Kumar, Binod, 2009, *"Study and analysis of knowledgebase of molecular systems and to develop model for prediction of molecular structure"*, thesis PhD, Saurashtra University

STUDY AND ANALYSIS OF KNOWLEDGEBASE OF MOLECULAR SYSTEMS AND TO DEVELOP MODEL FOR PREDICTION OF MOLECULAR STRUCTURE

A Thesis Submitted to

SAURASHTRA UNIVERSITY

For the award of the degree of

DOCTOR OF PHILOSOPHY
IN
COMPUTER SCIENCE

IN THE FACULTY OF SCIENCE

Submitted by

BINOD KUMAR
Assistant Professor in Computer Science – MCA
Institute of Science & Tech. For Advanced Studies & Research (ISTAR)
Vallabh Vidyanagar, Anand

Under the Guidance of

Dr. N. N. JANI
Ex. Prof. & Head, Computer Science Department,
Saurashtra University

DIRECTOR – MCA PROGRAMME, SKPIMCS
DEAN – FACULTY OF COMPUTER & IT
KADI SARVA VISHWAVIDYALAYA, GANDHINAGAR

AUGUST – 2009

Guide's Certificate

*I hereby certify that Mr. Binod Kumar has completed his thesis for the doctorate degree entitled "***Study and Analysis of Knowledgebase of Molecular Systems and to Develop Model for Prediction of Molecular Structure"***. I further certify that the research work done by him is of his own and original and carried out under my guidance and supervision. For the thesis that he is submitting, he has not been conferred any degree, diploma or distinction by either Saurashtra University or any other University according to best of my knowledge.*

Place:

Date:

Dr. N. N. Jani

Ex. Prof. & Head
Dept. of Computer Science, Saurashtra Univ.
Current Status:
Director- MCA Programme, SKPIMCS
Dean- Faculty of Computer & IT
Kadi Sarva Vishwavidyalaya, Gandhinagar

Researcher Certificate

I certify that the development model for Molecular Structure Prediction and strategies derived by analysis and described in the thesis has been based on the literature survey , bibliographical references, and research paper from International journals and National journals and various conferences proceedings and through study literature available on various websites in respect of related areas.

Apart from these, all analysis, hypothesis, inferences and interpretations of data and strategy have been my own and original creation. The model has been developed is my own and original creation. Moreover, I declare that for the work done in the thesis, either the Saurashtra University or any other University has not conferred any degree, diploma or distinction on me before.

Place:
Date:

Binod Kumar

Acknowledgement

Research is to see what everybody else has seen, and to think what nobody else has thought. This work is also no exception. It is my pleasure to convey my gratitude to all those who have directly or indirectly contributed to make this work successful.

First and foremost, this dissertation represents a great deal of time and effort not only on my part, but on the part of my supervisor, ***Dr. N. N. Jani****, Ex. Prof. & Head, Computer Science Dept., Saurashtra University, Rajkot. I expressed my profound gratitude to Dr. Jani sir for his endless encouragement throughout my research work. He has helped me shape my research from day one, pushed me to get through the inevitable research setbacks, and encouraged me to achieve to the best of my ability. A person with great concern for his students, he will remain an exemplar in my future.*

I take opportunity to express my deep sense of gratitude to **Dr. V. S. Patel***, Director, SICARD, Sardar Patel Centre for Science and Tech., Vallabh Vidyanagar, Anand, Gujarat for providing me visit at research centre. I got a chance to interact with sophisticated instruments like X-Ray Diffractometer (XRD) and Inductively Coupled Plasma Spectrometer (ICP).*

I express my respectful gratitude to **Dr. M. M. Patel***, Ex. Director, ISTAR, Vallabh Vidyanagar, Anand ,* **Dr. D. J. Desai***, Principal, V.P. & R.P.T.P Science College, Vallabh Vidyanagar and* **Dr. O. S. Srivastava***, HOD, MCA Dept., ISTAR for providing me all kinds of facility and moral support for completing my research work on time.*

Last but not least, I am thankful to my family members for their moral support and constant motivation for encouraging me in completing my work successfully.

Binod Kumar

TABLE OF CONTENTS

Page No.

Page No.

Page No.

LIST OF FIGURES

LIST OF TABLES

Chapter -1

INTRODUCTION

Chapter 1
INTRODUCTION

1.1 Introduction

This research work aims to analyze experimental data about biochemical properties and their corresponding kinetics. In this research the attempt has been made to analyze protein and DNA structure using tools such as DAMBE and Jemboss. Some Molecular Visualization or Analysis tools are already developed that reads, analyses, and cross-correlates experimental information which is useful for chemist, Organist Chemist, Biochemist and Druggist.

Under this research the analysis of different chemical and biochemical substances including drugs using tools like ACD/ChemSketch and NMR Prediction have been performed. The information obtained by the way of analysis that facilitates for in depth understanding of structures and that makes possible for a quantification of new chemical structure.

In this research using ACD/ChemSketch compounds are stored in databases and SMILE code (Simplified Molecular Input Line Specification) is generated. A SMILE defines the molecules in the form of alphanumeric chains. In this research work chemical shift of every carbon atom of the molecule have been displayed by using NMR Prediction.

Under this research CML codes of molecules have been developed and that codes have been used for molecular information like symmetry, and atom and bond attributes. Here multiple observations of the same molecule like conformational analysis and NMR prediction have been performed.

Using Pubchem/NCBI additional miscellaneous information such as bioactivity analysis by structure & activity similarity and revised compound selection after addition of similar compounds have been analyzed.

Under the research work geometric optimization of molecules, chemical structure visualization and calculation of electronic absorption spectra of chemical structure have been performed using ArgusLab tool. In this research Single Entry Point Calculation, Molecular Orbital calculation on grids for plotting HOMO and LUMO and ESP Mapped Density calculations have been also performed.

Under the research work of different types of analysis like prediction of protein secondary structure, isoelectric point calculation etc. have been performed on nucleotide and protein sequence using DAMBE and Jemboss tools.

The aim of this research work is to develop a model for the prediction of molecular structure. In research work bioinformatics and cheminformatics approaches on molecule has been covered. In this research an integrated bioinformatics and cheminformatics approach has been discussed that enables retrieval and visualization of biological relationships across heterogeneous data sources. So, now it is getting importance to integrate biological information on large molecules and their interaction networks with programs chemical information on small drug molecules.

Bioinformatists and Chemoinformatists have working independently in their respective fields. But now development of small molecule drugs and small drug molecules with known properties has been utilized to study the functions of large networks of biological molecules in the fields of chemical biology.

The objective of this research work is to assist the organic and biochemist in each step of the synthesis planning process for prediction of molecular structure. This research work provides a series of methods and tools for chemical or biochemical applications. Built-in catalogs of fine chemicals or biochemical provide suitable starting materials for a synthesis or molecular structure prediction target. Using *similarity searches* or *substructure* searches the connection between the target compound and available starting materials has been achieved.

This research work aims to search strategic bonds in target molecule for synthesis procedures. Structural criteria of each bond within the query molecule are also taken into account. In this research data mining tools has been used to predict physical properties of structures. In research work analysis on knowledgebase molecular system has been performed and a model has been developed that uses information to make decisions and suggest new strategies for chemistry and biochemistry problems.

The knowledgebase molecular system has three components:

A. Knowledgebase as Chemical Memory: An attempt has been made to concentrate on knowledge based data with an increasing number of chemical systems. Taking advantage of data sharing, each calculation increases the level of 'experience' of expert system extending the knowledge base upon which new hypothesis and chemical concept has been derived.

B. Data Mining: A component for increasing the chemical knowledge is extracting chemically meaningful data out of large scale chemical simulations with minimum human effort. The challenges lies in distinguishing data that is irrelevant for specific question under specific investigation for those that are important. To carry out this task an attempt has been made to concentrate on a knowledgebase system that process the molecular orbital and trace changes and similarities between molecules. Under this research visualization techniques have been used to enlarge scope of analysis.

C. Towards Artificial Chemical Intelligence: The final part of this research to formulate hypothesis based on data provided by molecules .Under this research work an attempt has been made for prediction of molecular structure. Finally, a research work result has been collected and then analyzed using analysis tools and then evaluated the result.

1.2 The Research Area, Problem Domain and Literature Survey

Bioinformatics and management of scientific data are critical to support life science discovery. As computational models of proteins, cells and organisms become increasingly realistic much biology research has migrated from the wetlab to the computer. Successfully accomplishing the translation of biology *in silico*, however, requires access to a huge amount of information from across the research community. Much of information is currently available from publicly accessible data sources and more is being added daily. Unfortunately, scientists are not currently able to identify easily and exploit this information because of the variety of semantics, interfaces and data formats used by the underlying data sources. Providing biochemist, medical researcher and computer scientist with integrated access to all information they need a consistent format requires overcoming a large number of technical, social and political challenges.

In the last decade, biologist have experienced a fundamental revolution from traditional research and development (R&D) consisting in discovering and understanding genes , metabolic pathways and cellular mechanisms to large scale computer-based R&D that simulates the disease , the physiology , the molecular mechanism and pharmacology. This represents a shift away from life science’s empirical roots in which it was an interactive process. Today it is systematic thematic and predictive with genomics, informatics and automation all playing a role. This fusion of biology and information science is expected to continue and expand for predictable futures. The first consequence of this revolution is the explosion of available data that bimolecular researchers have to exploit. For example, an average pharmaceutical company currently uses information from at least 40 databases [1] , each containing large amounts of data (e.g. as of June 2002, GenBank [2,3] provides access to 20,649,000,000 bases in 17,471,000 sequences) that can be analyzed using a variety of complex tools such as FASTA, BLAST etc.

Over past several years, bioinformatics has become both an all encompassing term for every thing relating to computer science and biology and an every trendy one. There are variety of reasons for this including : (1) As computational biology evolves and expands , the need for solutions to the data integration problems it faces increases; (2) the media are beginning to understand the implications of genomics revolution that has been going on the last 15 or more years ; (3) the recent headlines and debates surrounding the cloning of animals and humans ; and (4) to appear cutting edge , many companies have relabeled the work the work as they are doing as bioinformatics instead of geneticists , biologists or computer science.

The analysis of data sets is one of the most important tasks in investigation of properties of chemical or biochemical compounds. Especially in Drug Design, methods are used to characterize complete sets of chemical or biochemical compounds instead of describing individual molecule. Data Mining, i.e. the exploration of large amounts of data in search for consistent patterns, correlation or other systematic relationships, can be helpful tool to evaluate “hidden” information in a set of molecules. Finding the adequate information for representation of new chemical structures is one of the most important problems in chemical data mining.

With the progressive specialization in services and extensive use of computational methods the steady increase of data is barely manageable even by a team of scientist. Thereby the interest in specific information is pushed into backward while global information of complete sets of data is becoming more and more important. Thus, the recognition of superior information for complete data sets becomes one of the most important tasks for information management in science.

In Chemistry or Biochemistry the investigation of molecular structures and of their properties is one of the most important areas. In chemistry an own language and namespace for molecular exists, that is still in development stage. With increase of computational information processing several conventions and formats for chemical information have been developed.

But, in one of the most important communication media of modern times, the internet, the chemical language has been used only in a few applications. While a couple of databases were accessible via WWW, no service exists, that allows a data mining of chemical datasets by the use of this specific language.

The task of Data Mining in chemical or biochemical context is to evaluate “hidden” information in set of chemical data. One of the differences of Data Mining compared to conventional database queries is the production of new information that is used to characterize chemical data in a more general way. Generally, it is not be possible to hold all of the potentially required information in a data set of chemical structures. Thus, the extraction of relevant information and production of reliable secondary information are important.

The similarity of two compounds concerning their biological activity is one of central tasks in the development of pharmaceutical products. A typical application is retrieval of structures with defined biological activity from a database. Biological activity is of special interest the development of drugs. The diversity of structures in a data set of drugs has been the interest for the synthesis of new compounds. With increasing variety of data set, the chance to find a new way of synthesis for a compound with similar biological property is increasing.

Therefore, finding the adequate information for representation of chemical structures is one of the basic problems in chemical data mining. Several methods have been developed in the last decades for the description of molecules including their chemical or biochemical properties.

1.3 Relevance of the Research

Data Mining Service Chemistry (DMSC) [4] is a project for the development and exploration of chemical data sets. With this service it is possible to analyze chemical or biochemical data sets for molecular patterns and systematic relationships using the methods like Statistical analyses and neural networks of individual molecules.

System for Drug Discovery (QIS D2) [5] is a unique adaptive learning system designed to predict potential large-scale drug characteristics such as toxicity and efficacy. BioSpice is a set of software tools designed to represent and simulate cellular processes.

A new computer program is developed that describes, GRINSP (geometrically restrained inorganic structure prediction) [6], which allows the exploration of the possibilities of occurrence of 3-, 4-, 5- and 6-connected three-dimensional networks.

A global optimization method [7] is presented for predicting the minimum energy structure of small protein-like molecules. This method begins by collecting a large number of molecular conformations, each obtained by finding a local minimum of a potential energy function from a random starting point. The information from these conformers is then used to form a convex quadratic global underestimating function for the potential energy of all known conformers.

GenomeThreader [8] implements several data types in a reusable manner. Compared to its predecessor GeneSeqer, it is considerably faster, easier to maintain, and extensible. It is widely used for gene structure prediction.

The general approach [9] for the prediction of possible crystal structures consists of the global exploration of the energy landscape of the chemical system, with typical methods being simulated annealing or genetic algorithms. In the case of simulated annealing, combinations of model potentials and Ab initio calculations for the energy evaluation are state of the art.

The characteristics [10] of a free web-based spectral database for the chemical research community, containing ^{13}C NMR spectra data from more than 4000 natural compounds, and with a continuous increasing. This database allows flexible searching via chemical structure, substructure, name, and family of compounds, as well as spectral features as chemical shift, allowing the structural elucidation of known and unknown compounds by comparison of ^{13}C NMR data.

In this research work planning has been made to provide a centralized access to a wide variety of data mining methods, like statistical processing and prediction of molecular structure. With this service it is possible to submit data sets or to compile a data set by extracting structures from chemical databases via Internet. For submitting or compiled data sets descriptors have been calculated with an extensive set of options. On the basis of these descriptors, several methods of data analysis have been performed on the data set.

1.4 Details of Remaining Chapters

This thesis is meant to be a major step in my personal interest in prediction of molecular structure.

Second chapter of this thesis provides an overview of tools like ACD/ChemSketch, NMR Prediction, Argus Lab, DAMBE and Jemboss. **ACD/Labs** is used for developed molecular structures, reactions, and schematic diagrams and calculated chemical properties of different substances (chemical and biochemical). **NMR Prediction** tool is used to perform estimation of **^{1}H-NMR** and **^{13}C-NMR** of different substances. **ArgusLab** tool is used to build chemical structure and optimized its geometry. **DAMBE** tool is used to manipulate and analyze molecular sequence data. **Jemboss** can perform activities on sequences like predicting protein secondary structure etc. **CML** is designed to represent molecular information. **SMILES** (Simplified Molecular Input Line Entry System) is a line notation for entering and representing molecules.

Third chapter of this thesis provides an overview of pair wise sequence alignment and multiple sequence alignment. In this chapter alignment score and gap penalty between sequences has been calculated. Multiple sequence alignment is useful in finding patterns in nucleotide sequences and for identifying structural and functional domains in protein families. The method of converting MSA to a phylogenetic tree has been used to reduce the problem of a multiple alignment to an iterative process of pair-wise alignments.

Forth chapter of this thesis provides an overview of sequence alignment tools like BLAST and FASTA. Here their working methods and the syntax used by these tools has been discussed. FASTA uses algorithm to search for similarities between one sequence and any group of sequences of same type (nucleic acid or protein) as the query sequence. BLAST uses a heuristic algorithm that seeks local as opposed to global alignments and is therefore able to detect relationships among sequences that share only isolated regions of similarity.

Fifth chapter of this thesis provides an overview of protein structure and Cheminformatics. The subunits of a protein are amino acids. The primary structure is the sequence of residues in the polypeptide chain. Secondary structure is a local regularly occurring structure in proteins and is mainly formed through hydrogen bonds between backbone atoms. Tertiary structure describes the packing of alpha-helices, beta-sheets and random coils with respect to each other on the level of one whole polypeptide chain. Ab Initio method and Heuristic methods have been used for protein structure prediction.

Sixth chapter of this thesis shows the strong interaction between representation and the methods used for data analysis: molecular representation need to capture relevant information and be compatible with the statistical methods used to analyze the data. The chapters review molecular representations and put focus on model validation using statistics, visualization methods, and standardization approaches.

1.5 References

[1] M. Peitsch. “From Genome to Protein Space.” Presentation at the Fifth Annual Symposium in Bioinformatics, Singapore, October, 2000.

[2] D. Benson, I. Karsch –Mizarachi, D.Lipman . “Genbank.” Nucleic Acids research 31, no 1 (2003): 23-27 , http://www.ncbi.nlm.nih.gov/Genbank.

[3] “Growth of GenBank.” (2003): http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

[4] CRC NCRC Institute for Information technology Artificial Intelligence subject index: ftp.sas.com/pub/neural/FAQ&.html#A_app_chemistry

[5] Ying Zhao, Charles Zhou, Ian Oglesby, Cliff Zhou Quantum Intelligence, Inc. 3375 Scott Blvd Suite 100 ,Santa Clara CA 95054.

[6] Universite´ du Maine, Laboratoire des oxydes et Fluorures, CNRS UMR 6010, Avenue O. Messiaen, 72085 Le Mans Cedex 9, France.

[7] K.A. Dill, A.T. Phillips, and J.B. Rosen, Molecular Structure Prediction by Global Optimization,

[8] Gordon Gremme , Volker Brendel , Michael E. Sparks , Stefan Kurtz, Engineering a software tool for gene structure prediction in higher organisms, Information and Software Technology 47 (2005) 965–978

[9] K Doll, J C Sch¨on and M Jansen , Structure prediction based on ab initio simulated Annealing , Max-Planck-Institute for Solid State Research, Heisenbergstr. 1, D-70569 Stuttgart, Germany.

[10]Kochev, N., Monev, V., Bangov, I.: Searching Chemical Structures. In: Chemoinformatics: A textbook. Wiley-VCH (2003) 291–318

Chapter -2

Computational teChniques, Tools and Technologies To support Bioinformatics

Chapter 2

COMPUTATIONAL TECHNIQUES, TOOLS AND TECHNOLOGIES TO SUPPORT BIOINFORMATICS

2.1 Introduction

Under this research work tools like ACD/ChemSketch, NMR Prediction, Argus Lab, DAMBE and Jemboss have been discussed.

ACD/Labs [1] has been used for developed molecular structures, reactions, and schematic diagrams and calculated chemical properties of different substances (chemical and biochemical). **NMR Prediction** [2] tool has been used to perform estimation of **^{1}H-NMR** and **^{13}C-NMR** of different substances. The proton shift estimation program has been invoked by this tool and the result has been displayed written to the drawing. Sometimes the drawing is changed to allow the display of certain shifts. **ArgusLab [3]** tool has been used to build chemical structure and its geometry has been optimized. It is being used for visualization of frontier p molecular orbitals of chemical structure.

DAMBE [4] tool has been used to manipulate and analyze molecular sequence data. DAMBE is used for calculation of genetic distances or phylogenetic reconstruction. **Jemboss** [5] has been used for interactively editing sequence alignment. Different activities on sequences have been performed by this tool like Editing Functions, Locking Sequences, Trim Sequences, Colour Schemes, Scoring Matrix, Consensus Sequence, Identity Table and Consensus Plot etc.

2.2 ACD/ChemSketch

2.2.1 Introduction

ACD/ChemSketch is the powerful all-purpose chemical drawing and graphics package from ACD/Labs developed to help chemists quickly and easily draw molecular structures, reactions, and schematic diagrams, calculate chemical properties, and design professional reports and presentations. ACD/Labs has been fully dedicated to building integrated solutions that enable data transfer and connection with in chemical organizations.

ChemBasic is a simple, convenient, and functionally rich *programming language* for presentation and manipulation of molecular structure related objects and all the contents of ACD /Labs current and future programs. ChemBasic is founded on, and fully integrated with, ACD/Labs existing functionality. At the same time, ChemBasic has all of the things a programming language should have: numeric and string variables, arrays, flow control and conditional operators, input output procedures, etc.

ChemBasic inherits from *generic BASIC* and some of its extensions. Most evident is a product of Microsoft's Visual Basic for Applications (VBA). ChemBasic is designed as object oriented language. This means that all the chemistry related things are described as objects—that is, specific data structures which correspond to molecules, conformations, etc. I can design multi item input forms using ChemBasic programs using ACD/Forms Manager.

2.2.2 ACD/ChemSketch includes

- Structure mode for drawing chemical structures and calculating their properties.
- Draw mode or text and graphics processing.
- Additional modules that extend the ChemSketch possibilities (most of them should be purchased separately).

Structure mode. General information

In the Structure mode, following actions can be performed:

- Chemical structures can be drawn using the buttons located on the Structure toolbar, Atoms toolbar and References toolbar.

- For the selected structure the molar refractivity, molar volume, parachor, index of refraction, surface tension, density, and some other physicochemical properties can be calculated.
- Chemical structures can be finding according to their systematic or non-systematic names, therapeutic category or inhibited enzyme by using the integrated ACD/Dictionary .
- Most favorable tautomeric forms of the drawn structure can be checked and can be automatically corrected the structure by using the integrated Tautomeric Forms function on the Structure toolbar.
- An optimized 3D model of a 2D structure can be get.

In **Draw mode**, the following actions can be performed:

- Graphical objects such as lines, arrows, rectangles, ellipses, arcs, polylines, and polygons can be drawn by using the Drawing toolbar buttons.
- Objects can be manipulated.
- Location of objects on the page with a ruler and gridlines can be controlled.

2.2.3 Structure Representation

Antialiasing has been supported by ACD/ChemSketch that displays chemical structures drawn with smooth lines. Antialiasing is a computer rendering technique that blurs the hard edges and adds shaded pixels to create the appearance of smoothness. This addresses the common issue with printers and computer monitors, when, due to the relatively low resolution, the tilted lines appear “stairlike” instead of smooth straight lines or curves. For example, compare the two pictures below:

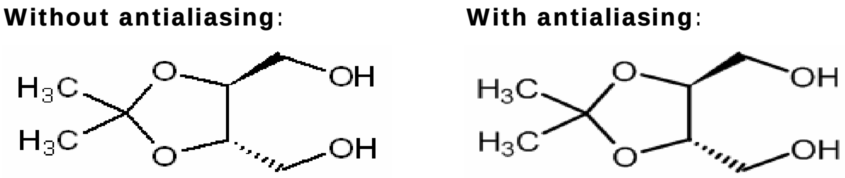


Figure 2.1: Stairlike curves

Treat some bonds to metal atoms as coordination bonds

ACD/Labs support the usage of a special *coordination bond* to represent a specific bonding between a ligand and a metal center in coordination structures. Such a bond indicates a connection but does not affect the valence of the corresponding atoms. However, often use of the regular *single bond* to represent a coordination that leads to formal violation of valence rules. Such a violation is marked in ACD/ChemSketch by “crossed atoms”.

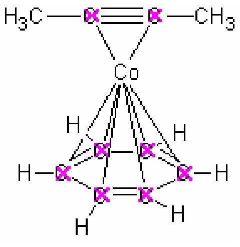
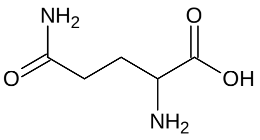


Figure 2.2 : Coordination Bond Representation

2.2.4 IUPAC International Chemical Identifier (InChI)

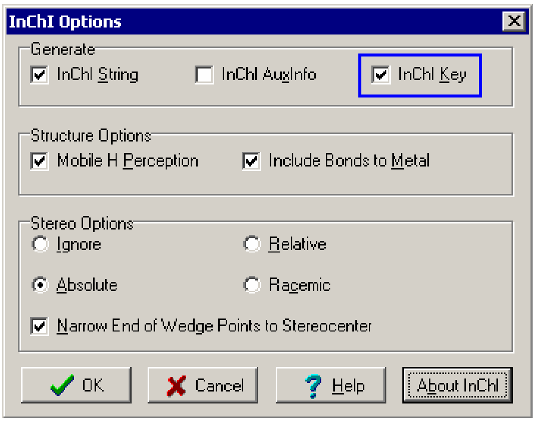
The IUPAC International Chemical Identifier (InChI™) is a non-proprietary identifier enabling unambiguous identification of chemical substances for electronic handling of chemical structural information. InChI codes significantly expand the use of InChI encoding for structure specification and searching over the Internet. For example:



InChI=1/C5H10N2O3/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H2,7,8)(H,9,10)

Figure 2.3:2,5-diamino-5-oxopentanoic acid

InChI generation options include an option for InChIKey generation:



InChI Options

Generate

☑ InChI String □ InChI AuxInfo ☑ InChI Key

Structure Options

☑ Mobile H Perception ☑ Include Bonds to Metal

Stereo Options

○ Ignore ○ Relative

● Absolute ○ Racemic

☑ Narrow End of Wedge Points to Stereocenter

Figure 2.4: InChIKey Option

For quick access of InChI generation, a special button "Generate InChI" has been added to the top toolbar:

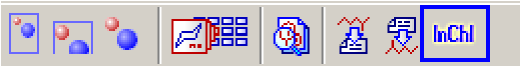


Figure 2.5: Generate InChIKey Button

2.3 NMRPrediction

2.3.1 Introduction

This software performs different estimation of a structure. It estimates ^{1}**H-NMR**. It invokes the proton shift estimation program and displays the results written to the drawing. Sometimes the drawing is changed to allow the display of certain shifts. It estimates ^{13}**C-NMR**. It invokes the carbon-13 shift estimation program and displays the results written to the drawing. **Show Protocol** command displays detailed information about the most recently invoked shift estimation. **Calculate 3D Coordinates** command displays the currently drawn structure as a 3D display in its own window. The molecule can be rotated by moving the mouse.

2.3.2 **Taking** example of **Glutamyl**

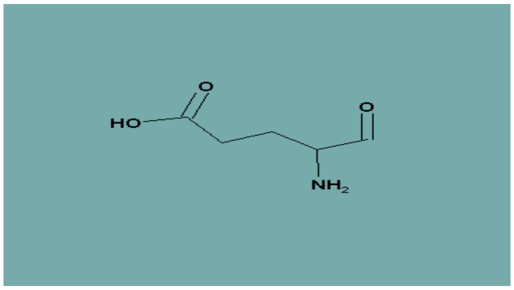


Figure 2.6: Example of Glutamyl

1H-NMR spectra of Glutamyl

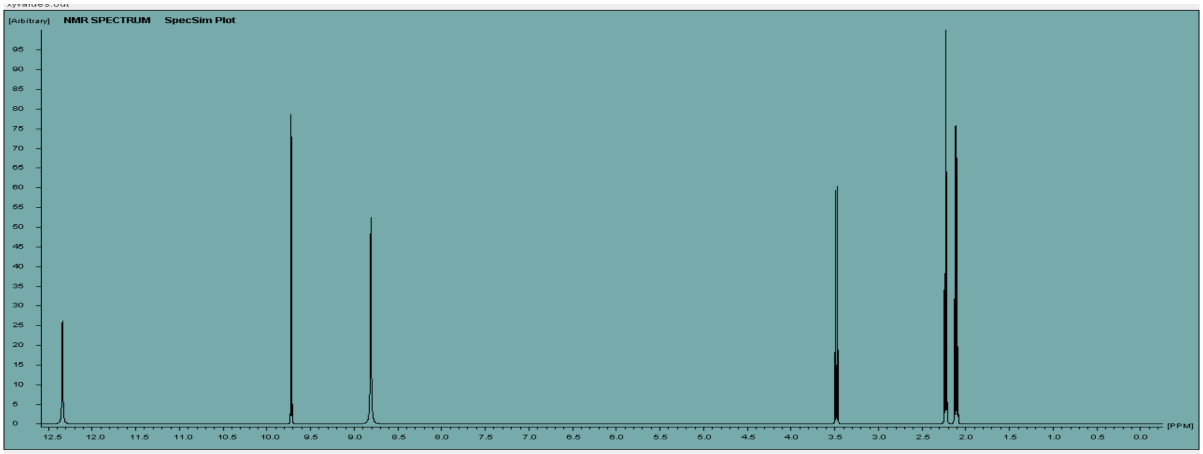


Figure 2.7: 1H-NMR spectra of Glutamyl

Table 2.1: Shift Prediction Protocol

Node	**Shift**	**Base + Inc.**	**Comment (ppm rel. to TMS)**
NH2	8.81	2.00	amine
		6.81	general corrections
CH	3.48	1.50	methine
		1.13	1 alpha -N
		0.86	1 alpha -C=O
		-0.01	1 beta -C
CH	9.72	9.60	CHO
		0.12	1 -C
CH2	2.11	1.37	methylene
		0.22	1 beta -N
		0.29	1 beta -C=O
		0.23	1 beta -C(=O)O
CH2	2.23	1.37	methylene
		0.90	1 alpha -C(=O)O
		-0.04	1 beta -C
OH	12.34	11.00	carboxylic acid

2.4 ArgusLab

2.4.1 Introduction

Argus Lab performing following capabilities:

- Build chemical structure and optimize its geometry.
- Visualize frontier p molecular orbital's of chemical structure.
- Calculate the electronic absorption spectra of chemical structure.
- Use a surface to visualize the spin-density in a molecule with unpaired spins.
- Make a surface that maps the electrostatic potential to the electron density.
- Using surfaces to see what happens to the electron density when a molecule absorbs light.

2.4.2 Building of Benzene

Benzene structure can be built from scratch and its geometry can be optimized. After addition of atoms from editor Benzene molecule can be generated and bonds can be made automatically. Following structure can be shown.

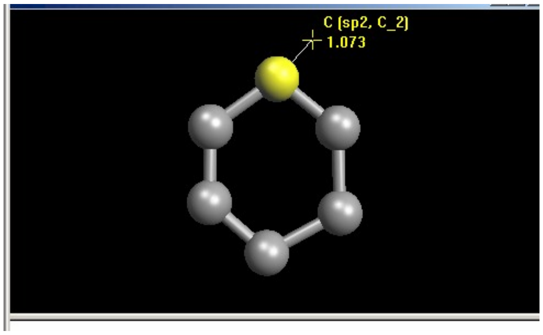


Figure 2.8: Building of Benzene

All bonds can be shown in that Benzene structure.

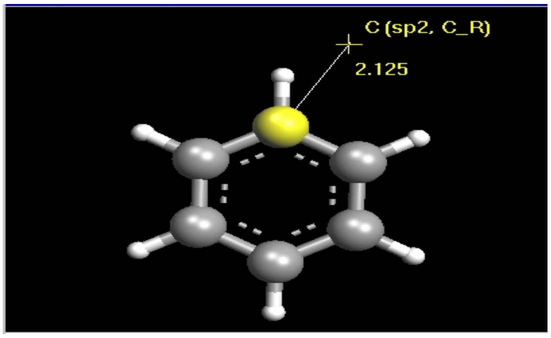
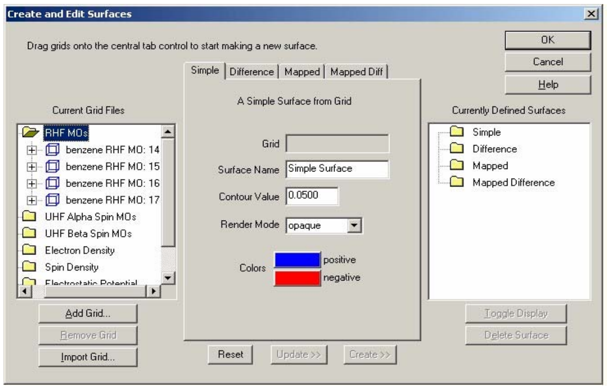


Figure 2.9: Bonds in Benzene Structure

Visualize the Building of Benzene

ArgusLab with generated MO grid files.


Create and Edit Surfaces

Drag grids onto the central tab control to start making a new surface.

A Simple Surface from Grid

Grid:

Surface Name: Simple Surface

Contour Value: 0.0500

Render Mode: opaque

Colors: 🟦 positive, 🟥 negative

Figure 2.10: Building of Benzene Visualization

Visualization of MOs of Benzene

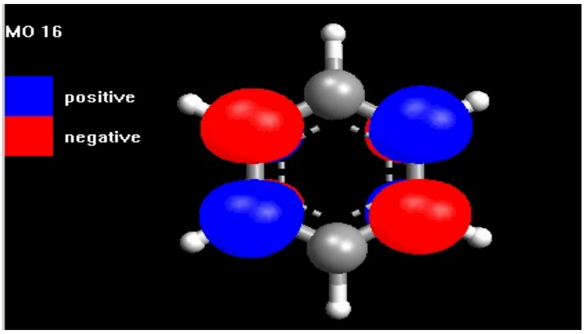


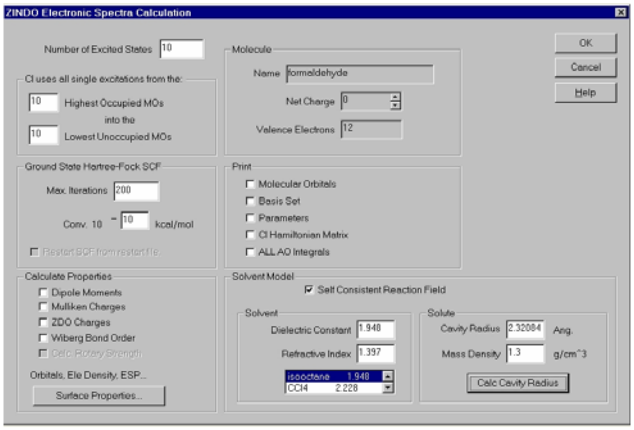
Figure 2.11: Visualization of MOs of Benzene

Calculating the electronic UV/Visible absorption spectrum of Benzene

The electronic excited states of benzene can be calculated using the semi-empirical ZINDO method which is parameterized for low-energy excited states of organic and organo-metallic molecules.

Calculating the ZINDO Electronic Spectra of a Molecule

The calculation consists of a ground-state closed shell SCF calculation followed by a configuration interaction calculation, using single-excited configurations, to solve for the excited states. Currently, only singlet excited states can be calculated.



ZINDO Electronic Spectra Calculation

Number of Excited States 10

CI uses all single excitations from the:
10 Highest Occupied MOs
into the
10 Lowest Unoccupied MOs

Molecule
Name formaldehyde
Net Charge 0
Valence Electrons 12

Ground State Hartree-Fock SCF
Max. Iterations 200
Conv. 10 - 10 kcal/mol
□ Restart SCF from restart file

Print
□ Molecular Orbitals
□ Basis Set
□ Parameters
□ CI Hamiltonian Matrix
□ ALL AO Integrals

Calculate Properties
□ Dipole Moments
□ Mulliken Charges
□ ZDO Charges
□ Wiberg Bond Order
□ Calc. Rotary Strength
Orbitals, Ele Density, ESP...

Solvent Model
☑ Self Consistent Reaction Field

Solvent
Dielectric Constant 1.948
Refractive Index 1.397
isooctane 1.948
CCl4 2.228

Solute
Cavity Radius 2.32084 Ang.
Mass Density 1.3 g/cm^3

Figure 2.12: Calculating the ZINDO Electronic Spectra of a Molecule

Making an electrostatic potential-mapped electron density surface

ArgusLab can generate Mapped surfaces. These are surfaces where one property is mapped onto a surface created by another property. The most popular example of this is to map the electrostatic potential (ESP) onto a surface of the electron density. In an ESP-mapped density surface, the electron density surface gives the shape of the surface while the value of the ESP on that surface gives the colors.

The electrostatic potential is the potential energy felt by a positive "test" charge at a particular point in space. If the ESP is negative, this is a region of stability for the positive test charge. Conversely, if the ESP is positive, this is a region of relative instability for the positive test charge. Thus, an ESP-mapped density surface can be used to show regions of a molecule that might be more favorable to nucleophilic or electrophilic attack, making these types of surfaces useful for qualitative interpretations of chemical reactivity.

Steps for calculating the following surface:

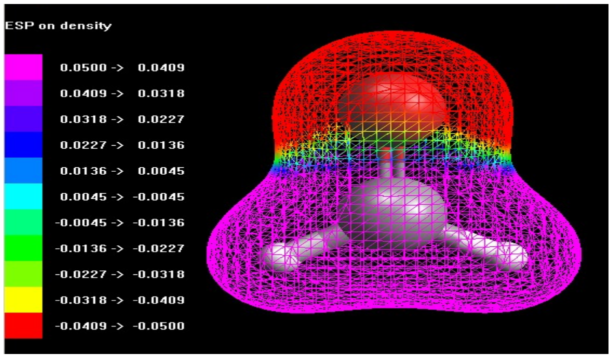
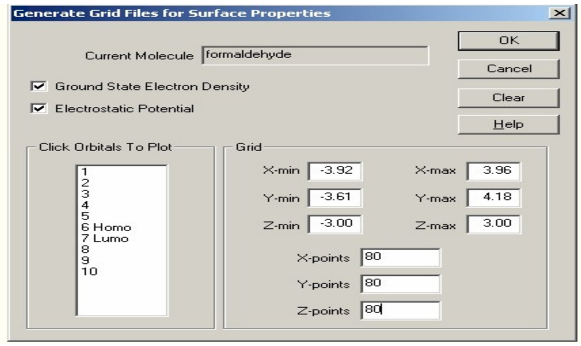


Figure 2.13: The surface in a mesh rendering to make it easier to see the underlying molecular structure

This is an ESP-mapped density surface of formaldehyde. The colors are the value of the ESP at the points on the electron density surface. The color map is given on the left. The large red region around the oxygen-end of the molecule. There is enhanced electron density here. The red color indicates the most negative regions of the electrostatic potential where a positive test charge would have favorable interaction energy. The hydrogen-end of the molecule, with the magenta color, shows regions of relatively unfavorable energy for the ESP.

Making the Surface: Generate the grid data

All surfaces are constructed from grid data that is generated from a calculation. To generate the grid data, a single-point energy calculation of formaldehyde can be run.



Generate Grid Files for Surface Properties

Current Molecule: formaldehyde

☑ Ground State Electron Density
☑ Electrostatic Potential

Click Orbitals To Plot: 1, 2, 3, 4, 5, 6 Homo, 7 Lumo, 8, 9, 10

Grid

X-min	-3.92	X-max	3.96
Y-min	-3.61	Y-max	4.18
Z-min	-3.00	Z-max	3.00

X-points: 80
Y-points: 80
Z-points: 80

Figure 2.14: Generate the Grid Data

Seeing the lone pairs on the oxygen

Some of the surface's settings can be altered to visualize the lone pair electron density on the oxygen.

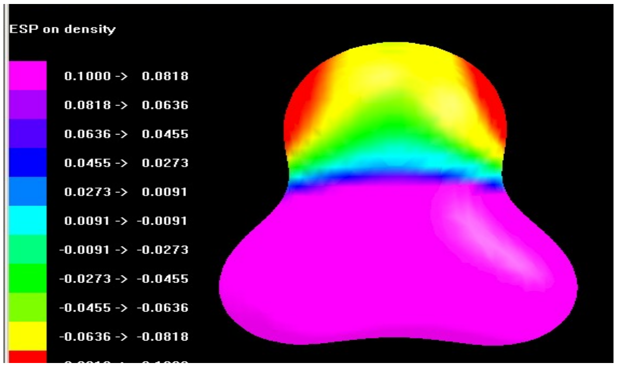
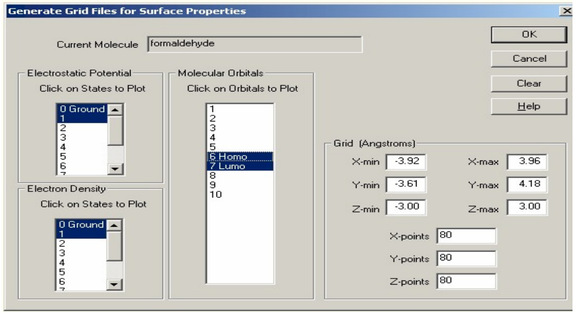


Figure 2.15: The lone pairs on the oxygen

Using surfaces to see change in the electron density when a molecule absorbs light

Here the first excited state of simple molecule formaldehyde (CH_2O) has been examined. The highest occupied molecular orbital (HOMO) of formaldehyde is a non-bonding type MO that is in the plane of the molecule. The lowest unoccupied molecular orbital (LUMO) is a p MO perpendicular to the plane of the molecule. The first excited state of formaldehyde is an n->p* transition that is composed almost exclusively of the HOMO->LUMO transition.

Calculate the electronic absorption spectra or formaldehyde



Generate Grid Files for Surface Properties

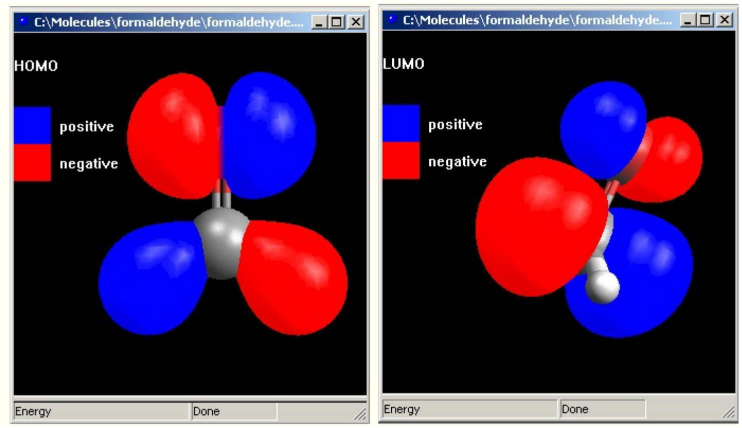
Current Molecule: formaldehyde

Grid (Angstroms)

X-min	-3.92	X-max	3.96
Y-min	-3.61	Y-max	4.18
Z-min	-3.00	Z-max	3.00

X-points: 80
Y-points: 80
Z-points: 80

Figure 2.16 : Visualizing the frontier MOs



C:\Molecules\formaldehyde\formaldehyde....
HOMO
🟦 positive
🟥 negative
Energy
Done
C:\Molecules\formaldehyde\formaldehyde....
LUMO
🟦 positive
🟥 negative
Energy
Done

Figure 2.17 : Visualizing the frontier MOs(Diagram)

Electron Density Difference

Different surface can be made to show the difference of the excited state minus the ground state electron density.

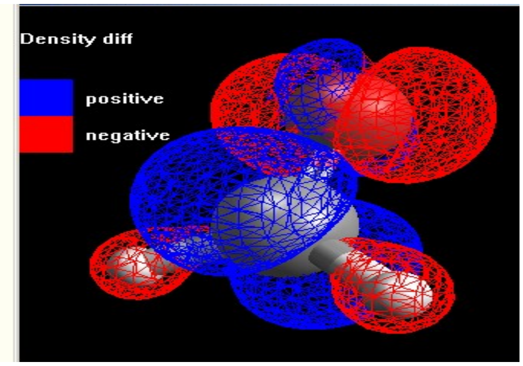
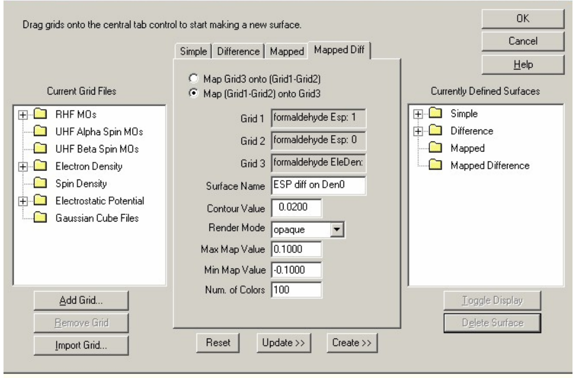


Figure 2.18: Electron Density Difference of benzene

Mapping the ESP difference onto the electron density



Drag grids onto the central tab control to start making a new surface.

○ Map Grid3 onto (Grid1-Grid2)
● Map (Grid1-Grid2) onto Grid3

Grid 1: formaldehyde Esp: 1
Grid 2: formaldehyde Esp: 0
Grid 3: formaldehyde EleDen:
Surface Name: ESP diff on Den0
Contour Value: 0.0200
Render Mode: opaque
Max Map Value: 0.1000
Min Map Value: -0.1000
Num. of Colors: 100

Figure 2.19: Mapping the ESP difference onto the electron density

2.5 DAMBE

DAMBE stands for Data Analysis in Molecular Biology and Evolution. It is an integrated software package for retrieving, organizing, manipulating aligning and analyzing molecular sequence data. Allele frequency data can also be used by DAMBE for calculating genetic distances or phylogenetic reconstruction.

2.5.1 Main Feature

DAMBE's main features can be classified into the following five categories:

1 Database and network functions:

(a) Molecular sequences can be directly read from GenBank or other networked computers;

(b) Specific sequences from GenBank sequences can be extracted by using information contained in the FEATURES table of GenBank sequence files;

2 Sequence conversion and manipulation utilities:

(a) It can be automatically detected and can be converted to 18 most commonly used molecular data formats;

(b) Complementary sequences can be getting.

(c) Protein-coding nucleotide sequences can be translated into amino acid sequences, with 12 different genetic codes implemented;

(d) Sequence can be aligned,

(e) Site-wise unresolved nucleotide, amino acid or codon sites can be deleted.

(f) Particular sites can be extracted, e.g., first, second or third codon positions, for particular analyses;

3 Sequence analysis can be focused on, factors affecting the frequency parameters in substitution models:

(a) Nucleotide and Dinucleotide frequencies

(b) Codon frequencies

(c) Amino acid frequencies

(d) Amino acid properties can be plotted along the sequence; with the following properties implemented:

- Polarity
- Polar requirement
- Chemical composition of the side chain
- Volume
- Hydropathy
- Isoelectric point
- Aromaticity

4 Basic comparative sequence analysis can be performed that focus on factors affecting the rate ratio parameters in substitution models:

(a) Nucleotide substitution pattern

(b) Codon substitution pattern

(c) Amino acid substitution pattern

(d) Substitution saturation

5 Advanced comparative sequence analysis can be performed

(a) Phylogenetic reconstruction based on the distance, maximum parsimony and maximum likelihood methods

(b) Reconstruction of ancestral sequences

(c) Testing the molecular clock hypothesis

(d) Evaluating relative statistical support of alternative phylogenetic hypotheses (e.g., alternative phylogenetic trees)

(e) Fitting probability distributions to substitution data over sites.

2.5.2 Sequence Analysis

This command computes the nucleotide and dinucleotide frequencies.

A part of a sample output (for one sequence) is shown below:

```
                  A        C        G       U      Other     Sum(ACGU)
===================================================================
FLAHAOHF
Freq             339     211     209     210       0           969
Prop.            .22      35     .22     .22      .22                       1
===================================================================
                A        C       G        U            Sum
===================================================================

Obs. A         117      75       72       75           339
Exp.           119      74       73       74

Obs. C          88      52       26       44           210
Exp.            74      46       45       46

Obs. G          79      32       53       45           209
Exp.            73      46       45       45

Obs. U          55       52      57       46           210
Exp.            74       46      45       46

Subtotal        339     211     208       210          968
```

The output is of two parts for each sequence, the first part lists the nucleotide frequencies, with "Other" stands for all characters that are not "acgtu", e.g., "-?.". The second part lists the di-nucleotide frequencies and the expected frequencies when there is no association or repulsion between nucleotides (i.e., the probability of two nucleotides sitting next to each other depends entirely on their frequencies). The di-nucleotides are counted from the beginning to the end of the sequence, with the nucleotides on the left column being the first, and those on the top row being the second, of the dinucleotide. From the first part of the output, it has been interpreted that A is being used more frequently than other nucleotides.

2.5.3 Codon Frequency

This opens a dialog box for computing codon frequencies and codon usage bias. A part of a default sample output, based on a segment of the Influenza A viruses, is shown below:

Output from sequences in file C:\MS\virus\virus.rst on

Sequence length = 969 (After excluding '?', '-' and 'n'.)

Number of codons = 323

From pooled sequences

```
AA      Codon    Mean Number(Sum)      RSCU
===================================================
A       GCA            8.0(32)              1.87
        GCC            3.3(13)              0.75
        GCG            2.3(9)               0.56
        GCU            3.5(14)              0.82
V       GUA            6.5(26)              1.29
        GUC            4.0(16)              0.80
       GUG              0.3(21)             1.04
       GUU             4.5(18)              0.88
```

The codon usage table is based on the following sequences:

```
  1   FLAHAOHF
  2   FLAHA1N
  3   IAU11858
  4   IVHATG391
```

```
             CodSite     A        C         G          U      Sum
====================================================================
   1           Freq.     433       221       357       281     1292
               Prop.     .34       .17       .28       .22        1

   2           Freq.     428       318       240       306     1292
               Prop.     .33       .25       .19       .24        1

   3           Freq.     461       319       228       284     1292
               Prop.     .36       .25       .18       .22        1
```

The output is in two parts. The first is a table of codon frequencies categorized into codon families, and the second lists nucleotide frequencies separately for each of the three codon positions designated as CodSite in the output.

2.5.4 Nonsynonymous codon substitution:

The sequence pairs available for selection on the left list depends on what input file format that is being used. If input format is NOT the RST format, then the number of possible sequence pairs is simply N*(N-1)/2. A partial sample output for a set of elongation factor 1-sequences (for only one pair-wise comparison between two chelicerate species) is shown below:

```
N     Cod1   Cod2   AA1  AA2    DG     DM
===========================================
node#8 vs. node#9

 037    AGG    AGU     R    S   109.00   2.73
 106    ACC    GCC     T    A    58.00   0.90
 150    UCA    CCA     S    P    73.00   0.55
 241    AAG    GAG     K    E    53.00   1.05
 244    GUU    CUU     V    L    32.00   0.91
 357    GAA    GAC     E    D    61.00   1.46
-------------------------------------------------------
Mean                          64.33   1.27
Num NS: 6

node#11 vs. Bra90058
 26   UAC    UUC     Y    F     22.00   0.48
 106  GCC    AAC     A    N    110.00   1.77
 117  ACU    UCU     T    S     58.00   0.89
 147  GCC    ACC     A    T     58.00   0.90
 148  AAG    AAC     K    N     94.00   1.83
 149  AUG    UUG     M    L     14.00   0.41
 156  AAC    GCC     N    A    110.00   1.77
 171  GAA    GAC     E    D     61.00   1.46
 213  AUG    AUC     M    I     10.00   0.29
 272  AAC    AGC     N    S     46.00   1.31
 282  UCU    UAC     S    Y    143.00   3.32
------------------------------------------------------
Mean                            66.00   1.31
Num NS: 11
```

Pair-wise comparisons along the tree are either between internal nodes, or between an internal node and a terminal node. This information is shown at the beginning of each pair-wise comparison. The first column shows the sequential numbering of codons along the DNA sequences (after deleting unresolved codons). The second and third columns show which codon pairs are involved in the substitution, and the fourth and fifth columns show the corresponding amino acids.

2.6 Jemboss

2.6.1 Introduction

Jemboss is developed by the EMBOSS team and is a graphical interface to the European Molecular Biology Open Software Suite, EMBOSS. Jemboss incorporates the 200+ applications of both the EMBOSS and EMBASSY packages.

The job manager is used to monitor the status of batch processes. These are those EMBOSS applications that are computationally intensive. Instead of waiting for the results these processes are submitted as batch, which frees the interface for other analyses to be carried out. This product includes code licensed from RSA Data Security.

2.6.2 Local and Remote File Manager

The users local and the remote file systems can be displayed. The local files are those stored on the computer that Jemboss is being run on. The remote files are the users files located on the server machine that runs the EMBOSS applications.

The activities performed by file manager are:

- Drag and Drop Files
- Transferring Files
- Refresh' File Manager
- Multiple File Selection

2.6.3 Jemboss Results Manager

Applications in Jemboss can be run **'interactively'** or in **'batch'** mode. Interactive applications wait for the process to finish and the results pop up on the screen. Batch process run in the background so that other tasks can be performed in Jemboss while the application is running. In both cases the results are stored on the server machine and can be retrieved at any time.

2.6.4 Sequence List

This window allows us to store their commonly used sequences.

An EMBOSS list file contains "references" to sequences, for example the file has been looked like: opsd_abyko.fasta, sw: opsd_xenla, sw: opsd_c* and @another_list etc.

The sequence length has been calculated by ‘Calculate sequence attributes' under the 'Tools' menu. The sequence start and end positions has been displayed.

2.6.5 Jemboss Alignment Editor

The Jemboss Alignment Editor has been used interactively to edit a sequence alignment (read in fasta or MSF format). It can also be used from the command line to produce image files of the alignment (e.g. within a script).

Following activities has been performed by alignment editor:

- Loading Sequences
- Editing Functions
- Locking Sequences
- Trim Sequences
- Colour Schemes
- Scoring Matrix
- Consensus Sequence

2.7 Chemical Markup Language (CML)

2.7.1 Introduction

This is a variety of XML [6] designed to represent molecular information. It has been used to store chemical formulas and to display the molecules in graphical formats.

CML [7] has been developed to carry molecules, crystallographic data and reactions using an XML language. A universal, platform and application independent format for storing and exchanging chemical information has been offered by CML. CML outlines a variety of general purpose 'data-holder' elements and a smaller number of more specifically chemical elements (e.g. <molecule>, <reaction>, <crystal>) used to indicate chemical 'objects'. For example, a <molecule> will contain a <list> of <atom>s, which in turn have three <float>s specifying Cartesian coordinates for each atom.

CML provides no default conventions for labeling data elements and puts few restrictions on element ordering. The design of CML and contains minimal preconceptions as to the type of chemical information that has been stored using it.

2.7.2 Reading XML Documents [8]

Here is an example from the CML Schema:

```
<cml>
  <molecule id="m1">
    <atomArray>
      <atom elementType="N"/>
      <atom elementType="O"/>
    </atomArray>
  </molecule>
</cml
```

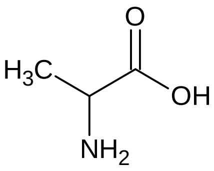
The first tag is `<cml>`. This is the top level tag. The next tag is `<molecule id="m1">`. The CML Schema reference says that the `<molecule>` tag is "a container for atoms, bonds and submolecules." The 'id' attribute is used as a unique identifier so that the molecule can be referred to from elsewhere. Similarly, the `<atomArray>` is "a container for a list of atoms." The tag `<atom elementType="N"/>` specifies a nitrogen atom and the tag `<atom elementType="O"/>` an oxygen atom.

The tag `</atomArray>` closes the `<atomArray>` element. Tags must always be closed with a `</...>` pattern in XML to create well formed documents. Also, tags must be fully enclosed within other tags and cannot overlap. For example, `<a><b></b></a>` is well formed XML but `<a><b></a></b>` is not. If a tag does not have anything inside it then the shorthand `<.../>` can be used to indicate both opening and closing an empty tag.

2.7.3 Examples of the molecules with CML

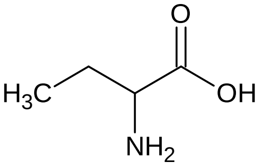
In this research substance like Alanine, Amino butyric Acid, Asparagine and Glutamine have been studied.

(A) Alanine



```
 <list xmlns:cml="http://www.xml-cml.org/schema/cml2/core" xmlns:stm="http://www.xml-
cml.org/schema/stmml" xmlns:ichi="http://www.iupac.org/foo/ichi" xmlns="http://www.xml-
cml.org/schema/cml2/core" title="/var/wwwtmp/mn_convert29053.xml">
 <molecule convention="CACTVS">
 <metadataList>
  <metadata name="dc:title">chemical structure data</metadata>
  <metadata name="dc:creator">wwwrun</metadata>
  <metadata name="dc:date">2009-06-03</metadata>
  </metadataList>
 <atomArray>
 <atom id="1">
  <string builtin="elementType">C</string>
  <float builtin="x2">16.9045</float>
  <float builtin="y2">-7.5353</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="2">
  <string builtin="elementType">O</string>
  <float builtin="x2">16.9045</float>
  <float builtin="y2">-6.2053</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="3">
  <string builtin="elementType">C</string>
  <float builtin="x2">15.7527</float>
  <float builtin="y2">-8.2003</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="4">
  <string builtin="elementType">N</string>
  <float builtin="x2">15.7527</float>
  <float builtin="y2">-9.5303</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="5">
  <string builtin="elementType">C</string>
  <float builtin="x2">14.6008</float>
  <float builtin="y2">-7.5353</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="6">
  <string builtin="elementType">O</string>
  <float builtin="x2">18.0563</float>
  <float builtin="y2">-8.2003</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
  </atomArray>
 <bondArray>
 <bond id="1">
  <string builtin="atomRef">1</string>
  <string builtin="atomRef">2</string>
  <string builtin="order">2</string>
  </bond>
 <bond id="2">
  <string builtin="atomRef">1</string>
  <string builtin="atomRef">3</string>
  <string builtin="order">1</string>
  </bond>
 <bond id="3">
  <string builtin="atomRef">3</string>
  <string builtin="atomRef">4</string>
  <string builtin="order">1</string>
  </bond>
 <bond id="4">
  <string builtin="atomRef">3</string>
  <string builtin="atomRef">5</string>
  <string builtin="order">1</string>
  </bond>
 <bond id="5">
  <string builtin="atomRef">6</string>
  <string builtin="atomRef">1</string>
  <string builtin="order">1</string>
  </bond>
  </bondArray>
  </molecule>
  </list>
```

(B) Amino Butyric Acid

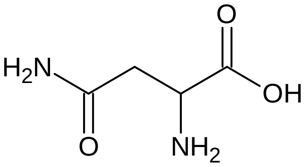


```
 <list xmlns:cml="http://www.xml-cml.org/schema/cml2/core" xmlns:stm="http://www.xml-
cml.org/schema/stmml" xmlns:ichi="http://www.iupac.org/foo/ichi" xmlns="http://www.xml-
cml.org/schema/cml2/core" title="/var/wwwtmp/mn_convert28995.xml">
 <molecule convention="CACTVS">
 <metadataList>
  <metadata name="dc:title">chemical structure data</metadata>
  <metadata name="dc:creator">wwwrun</metadata>
  <metadata name="dc:date">2009-06-03</metadata>
  </metadataList>
```

```
<atomArray>
<atom id="1">
 <string builtin="elementType">C</string>
 <float builtin="x2">16.7272</float>
 <float builtin="y2">-9.5383</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
<atom id="2">
 <string builtin="elementType">O</string>
 <float builtin="x2">16.7272</float>
 <float builtin="y2">-8.2083</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
<atom id="3">
 <string builtin="elementType">C</string>
 <float builtin="x2">15.5755</float>
 <float builtin="y2">-10.2033</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
<atom id="4">
 <string builtin="elementType">N</string>
 <float builtin="x2">15.5755</float>
 <float builtin="y2">-11.5333</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
<atom id="5">
 <string builtin="elementType">C</string>
 <float builtin="x2">14.4236</float>
 <float builtin="y2">-9.5383</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
<atom id="6">
 <string builtin="elementType">O</string>
 <float builtin="x2">17.8791</float>
 <float builtin="y2">-10.2033</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
<atom id="7">
 <string builtin="elementType">C</string>
 <float builtin="x2">13.2718</float>
 <float builtin="y2">-10.2034</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
 </atomArray>
<bondArray>
<bond id="1">
 <string builtin="atomRef">1</string>
 <string builtin="atomRef">2</string>
 <string builtin="order">2</string>
 </bond>
<bond id="2">
 <string builtin="atomRef">1</string>
 <string builtin="atomRef">3</string>
 <string builtin="order">1</string>
 </bond>
<bond id="3">
 <string builtin="atomRef">3</string>
 <string builtin="atomRef">4</string>
 <string builtin="order">1</string>
 </bond>
<bond id="4">
 <string builtin="atomRef">3</string>
 <string builtin="atomRef">5</string>
 <string builtin="order">1</string>
 </bond>
<bond id="5">
```

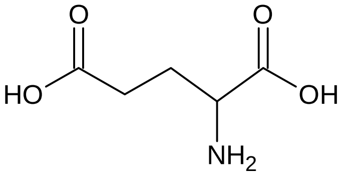
```
  <string builtin="atomRef">6</string>
  <string builtin="atomRef">1</string>
  <string builtin="order">1</string>
  </bond>
 <bond id="6">
  <string builtin="atomRef">7</string>
  <string builtin="atomRef">5</string>
  <string builtin="order">1</string>
  </bond>
  </bondArray>
  </molecule>
  </list>
```

(C)Asparagine



```
 <list xmlns:cml="http://www.xml-cml.org/schema/cml2/core" xmlns:stm="http://www.xml-
cml.org/schema/stmml" xmlns:ichi="http://www.iupac.org/foo/ichi" xmlns="http://www.xml-
cml.org/schema/cml2/core" title="/var/wwwtmp/mn_convert29126.xml">
 <molecule convention="CACTVS">
 <metadataList>
  <metadata name="dc:title">chemical structure data</metadata>
  <metadata name="dc:creator">wwwrun</metadata>
  <metadata name="dc:date">2009-06-03</metadata>
  </metadataList>
 <atomArray>
 <atom id="1">
  <string builtin="elementType">C</string>
  <float builtin="x2">21.5693</float>
  <float builtin="y2">-8.2244</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="2">
  <string builtin="elementType">C</string>
  <float builtin="x2">23.873</float>
  <float builtin="y2">-8.2244</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="3">
  <string builtin="elementType">C</string>
  <float builtin="x2">20.4176</float>
  <float builtin="y2">-8.8894</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="4">
  <string builtin="elementType">C</string>
  <float builtin="x2">22.7212</float>
  <float builtin="y2">-8.8894</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="5">
  <string builtin="elementType">O</string>
  <float builtin="x2">23.873</float>
  <float builtin="y2">-6.8944</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="6">
  <string builtin="elementType">O</string>
 <float builtin="x2">20.4176</float>
 <float builtin="y2">-10.2194</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
<atom id="7">
 <string builtin="elementType">N</string>
 <float builtin="x2">19.2658</float>
 <float builtin="y2">-8.2244</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
<atom id="8">
 <string builtin="elementType">N</string>
 <float builtin="x2">22.7212</float>
 <float builtin="y2">-10.2194</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
<atom id="9">
 <string builtin="elementType">O</string>
 <float builtin="x2">25.0248</float>
 <float builtin="y2">-8.8894</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
 </atomArray>
<bondArray>
<bond id="1">
 <string builtin="atomRef">1</string>
 <string builtin="atomRef">3</string>
 <string builtin="order">1</string>
 </bond>
<bond id="2">
 <string builtin="atomRef">1</string>
 <string builtin="atomRef">4</string>
 <string builtin="order">1</string>
 </bond>
<bond id="3">
 <string builtin="atomRef">2</string>
 <string builtin="atomRef">4</string>
 <string builtin="order">1</string>
 </bond>
<bond id="4">
 <string builtin="atomRef">2</string>
 <string builtin="atomRef">5</string>
 <string builtin="order">2</string>
 </bond>
<bond id="5">
 <string builtin="atomRef">3</string>
 <string builtin="atomRef">6</string>
 <string builtin="order">2</string>
 </bond>
<bond id="6">
 <string builtin="atomRef">3</string>
 <string builtin="atomRef">7</string>
 <string builtin="order">1</string>
 </bond>
<bond id="7">
 <string builtin="atomRef">4</string>
 <string builtin="atomRef">8</string>
 <string builtin="order">1</string>
 </bond>
<bond id="8">
 <string builtin="atomRef">9</string>
 <string builtin="atomRef">2</string>
 <string builtin="order">1</string>
 </bond>
 </bondArray>
 </molecule>
 </list>
```

(D) Glutamine



```
 <list xmlns:cml="http://www.xml-cml.org/schema/cml2/core" xmlns:stm="http://www.xml-
cml.org/schema/stmml" xmlns:ichi="http://www.iupac.org/foo/ichi" xmlns="http://www.xml-
cml.org/schema/cml2/core" title="/var/wwwtmp/mn_convert29157.xml">
 <molecule convention="CACTVS">
 <metadataList>
  <metadata name="dc:title">chemical structure data</metadata>
  <metadata name="dc:creator">wwwrun</metadata>
  <metadata name="dc:date">2009-06-03</metadata>
  </metadataList>
 <atomArray>
 <atom id="1">
  <string builtin="elementType">C</string>
  <float builtin="x2">14.2172</float>
  <float builtin="y2">-13.937</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="2">
  <string builtin="elementType">C</string>
  <float builtin="x2">18.8245</float>
  <float builtin="y2">-13.937</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="3">
  <string builtin="elementType">O</string>
  <float builtin="x2">14.2172</float>
  <float builtin="y2">-12.607</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="4">
  <string builtin="elementType">C</string>
  <float builtin="x2">15.3691</float>
  <float builtin="y2">-14.602</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="5">
  <string builtin="elementType">O</string>
  <float builtin="x2">18.8246</float>
  <float builtin="y2">-12.607</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="6">
  <string builtin="elementType">C</string>
  <float builtin="x2">16.5209</float>
  <float builtin="y2">-13.937</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="7">
  <string builtin="elementType">C</string>
  <float builtin="x2">17.6726</float>
  <float builtin="y2">-14.7782</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="8">
 <string builtin="elementType">O</string>
 <float builtin="x2">19.9763</float>
 <float builtin="y2">-14.602</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
<atom id="9">
 <string builtin="elementType">O</string>
 <float builtin="x2">13.0654</float>
 <float builtin="y2">-14.602</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
<atom id="10">
 <string builtin="elementType">N</string>
 <float builtin="x2">17.6726</float>
 <float builtin="y2">-16.1082</float>
 <float builtin="x3">0</float>
 <float builtin="y3">0</float>
 <float builtin="z3">0</float>
 </atom>
 </atomArray>
<bondArray>
<bond id="1">
 <string builtin="atomRef">1</string>
 <string builtin="atomRef">3</string>
 <string builtin="order">2</string>
 </bond>
<bond id="2">
 <string builtin="atomRef">1</string>
 <string builtin="atomRef">4</string>
 <string builtin="order">1</string>
 </bond>
<bond id="3">
 <string builtin="atomRef">2</string>
 <string builtin="atomRef">7</string>
 <string builtin="order">1</string>
 </bond>
<bond id="4">
 <string builtin="atomRef">2</string>
 <string builtin="atomRef">5</string>
 <string builtin="order">2</string>
 </bond>
<bond id="5">
 <string builtin="atomRef">2</string>
 <string builtin="atomRef">8</string>
 <string builtin="order">1</string>
 </bond>
<bond id="6">
 <string builtin="atomRef">4</string>
 <string builtin="atomRef">6</string>
 <string builtin="order">1</string>
 </bond>
<bond id="7">
 <string builtin="atomRef">6</string>
 <string builtin="atomRef">7</string>
 <string builtin="order">1</string>
 </bond>
<bond id="8">
 <string builtin="atomRef">9</string>
 <string builtin="atomRef">1</string>
 <string builtin="order">1</string>
 </bond>
<bond id="9">
 <string builtin="atomRef">10</string>
 <string builtin="atomRef">7</string>
 <string builtin="order">1</string>
 </bond>
 </bondArray>
 </molecule>
 </list>
```

2.8 SMILES - A Simplified Chemical Language

2.8.1 Introduction

SMILES [9] (Simplified Molecular Input Line Entry System) is a line notation for entering and representing molecules and reactions. Some examples are:

Table 2.2: SMILE examples

SMILES	Name	SMILES	Name
CC	ethane	[OH3+]	hydronium ion
O=C=O	carbon dioxide	[2H]O[2H]	deuterium oxide
C#N	hydrogen cyanide	[235U]	uranium-235
CCN(CC)CC	triethylamine	F/C=C/F	E-difluoroethene
CC(=O)O	acetic acid	F/C=C\F	Z-difluoroethene
C1CCCCC1	cyclohexane	N[C@@H](C)C(=O)O	L-alanine
c1ccccc1	benzene	N[C@H](C)C(=O)O	D-alanine

Reaction SMILES	Name
[I-].[Na+].C=CCBr>>[Na+].[Br-].C=CCI	displacement reaction
(C(=O)O).(OCC)>>(C(=O)OCC).(O)	intermolecular esterification

SMILES are a true language, although with a simple vocabulary (atom and bond symbols) and only a few grammar rules. SMILES representations of structure can in turn be used as "words" in the vocabulary of other languages designed for storage of chemical information (information about chemicals) and chemical intelligence (information about chemistry).

2.8.2 Canonicalization

A SMILE denotes a molecular structure as a graph with optional chiral indications. This is essentially the two-dimensional picture chemists draw to describe a molecule. SMILES describing only the labeled molecular graph (i.e. atoms and bonds, but no chiral or isotopic information) are known as generic SMILES.

It can be shown in the following examples.

Table 2.3: Canonicalization of SMILE

Input SMILES	Unique SMILES
OCC	CCO
[CH3][CH2][OH]	CCO
C-C-O	CCO
C(O)C	CCO
OC(=O)C(Br)(Cl)N	NC(Cl)(Br)C(=O)O
ClC(Br)(N)C(=O)O	NC(Cl)(Br)C(=O)O
O=C(O)C(N)(Br)Cl	NC(Cl)(Br)C(=O)O

2.8.3 SMILES Specification Rules

SMILES notation consists of a series of characters containing no spaces. Hydrogen atoms may be omitted (hydrogen-suppressed graphs) or included (hydrogen-complete graphs). Aromatic structures may be specified directly.

There are five generic SMILES encoding rules, corresponding to specification of atoms, bonds, branches, ring closures, and disconnections.

2.8.3.1 Atoms

Atoms are represented by their atomic symbols: this is the only required use of letters in SMILES. Each non-hydrogen atom is specified independently by its atomic symbol enclosed in square brackets, [ ]. The second letter of two-character symbols must be entered in lower case. Atoms in aromatic rings are specified by lower case letters, e.g., aliphatic carbon is represented by the capital letter C, aromatic carbon by lower case c.

Table 2.4: SMILE atoms

C	methane	(CH4)
P	phosphine	(PH3)
N	ammonia	(NH3)
S	hydrogen sulfide	(H2S)
O	water	(H2O)
Cl	hydrochloric acid	(HCl)

2.8.3.2 Bonds

Single, double, triple, and aromatic bonds are represented by the symbols -, =, #, and :, respectively. Adjacent atoms are assumed to be connected to each other by a single or aromatic bond (single and aromatic bonds may always be omitted). Examples are:

Table 2.5: SMILE bonds

CC	ethane	(CH3CH3)
C=O	formaldehyde	(CH2O)
C=C	ethene	(CH2=CH2)
O=C=O	carbon dioxide	(CO2)
COC	dimethyl ether	(CH3OCH3)
C#N	hydrogen cyanide	(HCN)
CCO	ethanol	(CH3CH2OH)
[H][H]	molecular hydrogen	(H2)

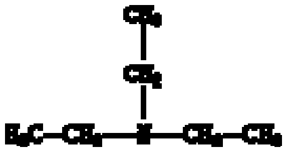
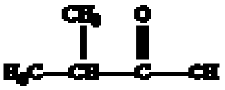
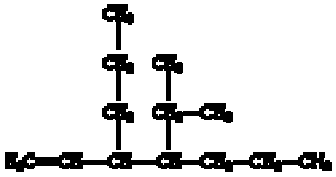
For linear structures, SMILES notation corresponds to conventional diagrammatic notation except that hydrogens and single bonds are generally omitted. For example, 6-hydroxy-1,4-hexadiene can be represented by many equally valid SMILES, including the following three:

Structure	Valid SMILES
	C=CCC=CCO
CH2=CH-CH2-CH=CH-CH2-OH	C=C-C-C=C-C-O
	OCC=CCC=C

2.8.3.3 Branches

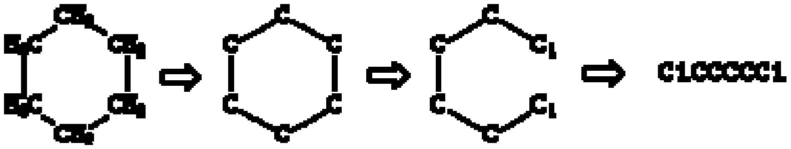
Branches are specified by enclosing them in parentheses, and can be nested or stacked. In all cases, the implicit connection to a parenthesized expression (a "branch") is to the left. Examples are:

Table 2.6: SMILE branches

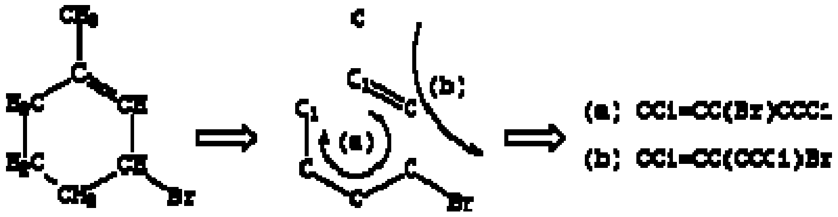
		
CCN(CC)CC	CC(C)C(=O)O	C=CC(CCC)C(C(C)C)CCC
Triethylamine	Isobutyric acid	3-propyl-4-isopropyl-1-heptene

2.8.3.4 Cyclic Structures

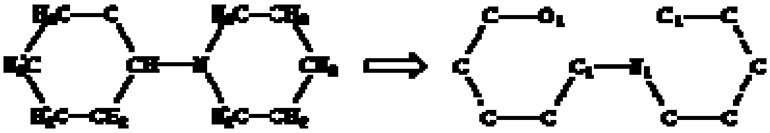
Cyclic structures are represented by breaking one bond in each ring. The bonds are numbered in any order, designating ring opening (or ring closure) bonds by a digit immediately following the atomic symbol at each ring closure. This leaves a connected non-cyclic graph which is written as a non-cyclic structure using the three rules described above. Cyclohexane is a typical example:



There are usually many different, but equally valid descriptions of the same structure, e.g., the following SMILES notations for 1-methyl-3-bromo-cyclohexene-1:



Digits denoting ring closures has been reused. As an example, the digit 1 used twice in the specification:

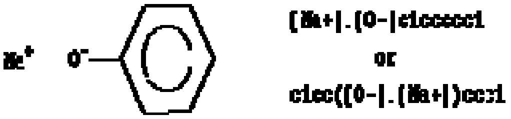


O1CCCCC1N1CCCCC1

The ability to re-use ring closure digits makes it possible to specify structures with 10 or more rings. Structures that require more than 10 ring closures to be open at once are exceedingly rare.

2.8.3.5 Disconnected Structures

Disconnected compounds are written as individual structures separated by a "." (period). If desired, the SMILES of one ion may be imbedded within another as shown in the example of sodium phenoxide.



Matching pairs of digits following atom specifications imply that the atoms are bonded to each other. The bond may be explicit (bond symbol and/or direction preceding the ring closure digit) or implicit (a nondirectional single or aromatic bond). This is true whether or not the bond ends up as part of a ring.

Adjacent atoms separated by dot (.) imply that the atoms are not bonded to each other. This is true whether or not the atoms are in the same connected component.
For example, C1.C1 specifies the same molecule as CC(ethane) .

2.10 References

[1] www.acdlabs.com

[2] www.upstream.ch

[3] www.arguslab.com

[4] http://dambe.bio/uottawa.ca

[5] http://mEMBOSS/jemboss/jar/resources/readme.html

[6] http://www.xml-cml.org/

[7] http://zvon.org/xxl/CML1.0/Output/index.html

[8] http://cml.sourceforge.net/

[9] http://www.daylight.com

Chapter -3

Alignment of PAirs and MulTiple SequenceS and Phylogenetic AnAlysis

Chapter 3

ALIGNMENT OF PAIRS AND MULTIPLE SEQUENCES AND PHYLOGENETIC ANALYSIS

3.1 Introduction

In this chapter Pair wise Sequence Alignment and Multiple Sequence Alignment has been discussed. In this chapter alignment score and Gap Penalty between sequences has been calculated. The gap penalty formula can be extended to include a penalty for alignments for the gaps at the end of a sequence of equal length. Multiple sequence alignment is useful in finding patterns in nucleotide sequences and for identifying structural and functional domains in protein families.

Multiple sequence alignment (MSA) is an extension of the similarity concepts to determine levels of homology (relatedness) between members of a series of globally related sequences are aligned together in column.

3.2 Sequence Description

Patterns provide appropriate representations of conserved regions in biosequences. In most cases one is given a set of sequences (DNA or proteins) and is looking for patterns that appear in some minimum number (or percentage) of these sequences. The exact definition of a pattern varies from algorithm to algorithm. In general, a pattern is a member of a well defined subset C of all the possible regular expressions over $\sum^{1}$ (the set C is called a pattern language). Being a regular expression, every pattern P defines a language L (P) in the natural way: a string belongs to L (P) if it is recognized by the automaton of P. A sequence $s \in \sum^{*}$ is said to “match" a given pattern P if s contains some substring that belongs to L (P).

For an illustration, consider the following set of strings over the English alphabet:

S = {LARGE, FINGER, AGE}

In this case the pattern “L..GE" has support 2 since it is matched by the first two strings of S (`.' is called the don't-care character and is used to indicate position that can be occupied by an arbitrary alphabet character). The term support denotes the number of input strings matching a given pattern. As another example, the pattern “A*GE” has also support 2 (it is matched by the first and the last strings). Here, the character `*' is used to match substrings of arbitrary length.

3.3 Pair wise Sequence Alignment

Pair wise sequence alignment [1] involves the matching of two sequences, one pair of elements at a time. The challenge in pair wise sequence alignment is to find the optimum alignment of two sequences with some degree of similarity. This optimum condition is based on a score that reflects the number of paired characters in two sequences and number and length of gaps required to adjust the sequences so the maximum number of characters are in alignment. For example, consider the ideal case of identical nucleotide sequences, (A) and (B)

A) **ATTCGGCATTCAGTGCTAGA**
B) **ATTCGGCATTCAGTGCTAGA**

Assuming that the alignment scoring algorithm counts one point per pair of aligned characters, then the score for each of the 20 pairs, or 20 points. Now, consider the case when several of character pairs aren’t aligned:

C) **ATTCGGCATT**CAGT**G**CTAGA
D) **ATTCGGCATT**GCTA**G**A

In this case, the score is 11, because only 11 pairs of characters in sequences (C) and (D) are aligned. By moving last six characters ahead in sequence (D) by adding four gaps, the sequences become:

E) **ATTCGGCATT**CAGT**GCTAGA**
F) **ATTCGGCATT - - - -GCTAGA**

Now, the score, based on the original algorithm of character pairing, is 16. However, because the score would have been 11 without the inserted gaps, a penalty should be extracted for each gap inserted into the sequence to favor alignments that can be made with as few gaps as possible. Assuming a gap penalty of -0.5 per gap, the alignment score becomes 10 + 6 + (4 × -0.5) or 14.

Another scenario is that in which the areas of similarity and difference are not obvious. Consider the sequences (G) and (H):

G) **ATTCG**G**CATT**CAG**AG**C**G**AG**A**
H) **ATTCG**A**CATT**GCT**AG**T**G**GT**A**

Unlike the previous cases, there are no relatively long runs of character pairings, and the matching pairs are separated by unaligned characters. The alignment score is 1 point per aligned pair or 13. One attempt at visual alignment by adding four gaps into sequence (H) results in:

G) **ATTCG**G**CATT**CAGA**GCTAG**A
I) **ATTCG**A**CATT - - - - GCTAG**TGGTA

This alignment results in a score of 12, or 14 alignments minus 2 points for the 4 gaps introduced into sequence (H), transforming it to sequence (I). In addition, a penalty of -0.5 per character pair is scored for an inexact match. In case of sequences (G) and (I), there are 6 inexact matches. In case of sequences (G) and (I), there are 6 inexact matches, for a penalty of (6 × -0.5 = -3). Using this new alignment scoring algorithm, and ignoring the length difference between the two sequences, the alignment score for the (G)-(I) alignment becomes:

$$\begin{aligned} \text{Alignment Score} &= 14 \text{ alignments} + 4 \text{ gaps} + 6 \text{ inexact matches} \\ &= 14 + (4 \times -0.5) + (6 \times -\ 0.5) \\ &= 14 - 2 - 3 \\ &= 9 \end{aligned}$$

In this example, adding gaps results in a lower alignment score, illustrating how the relative worth of exact matches, inexact matches and gaps determines the eventual alignment of two sequences.

Although a simple gap penalty of - 0.5 point per gap has been used to illustrate the role of alignment scores on sequence alignment, gap penalty is typically calculated as:

$$\text{Penalty}_{\text{gap}} = \text{Cost}_{\text{opening}} + \text{Cost}_{\text{extension}} \times \text{Length}_{\text{gap}}$$

In this formula, Penalty $_{\text{gap}}$ is the total gap penalty, Cost $_{\text{opening}}$ is the cost of opening is the cost of opening a gap in a sequence , Cost $_{\text{extension}}$ is the cost of extending an existing gap by one character, and Length $_{\text{gap}}$ is the length of the gap in characters. The minimum value of Length $_{\text{gap}}$ is one. Returning to sequence pair (E)-(F), assuming that Cost $_{\text{opening}}$ is (- 0.5) and Cost $_{\text{extension}}$ is (- 0.5), the gap penalty becomes:

$$\begin{aligned} \text{Penalty}_{\text{gap}} &= \text{Cost}_{\text{opening}} + \text{Cost}_{\text{extension}} \times \text{Length}_{\text{gap}} \\ &= -\ 0.5 + (-\ 0.5 \times 4) \\ &= -\ 2.5 \end{aligned}$$

With the expanded method of computing gap penalty, the score becomes 10 + 6 – 2.5 = 13.5 points. The gap penalty formula can be extended to include a penalty for alignments for the gaps at the end of a sequence of equal length.

3.3.1 Local versus Global Alignment

Sequence pair (E) – (F) is an example of global alignment- that is , an attempt to line up the two sequences matching as many characters as possible, for the entire length of each segment. Global alignment considers all characters in a sequence, and bases alignment on the total score, even at the expense of stretches in the sequence that share similarity as shown in figure. Global alignment is used to help determine whether two protein sequences are the same family.

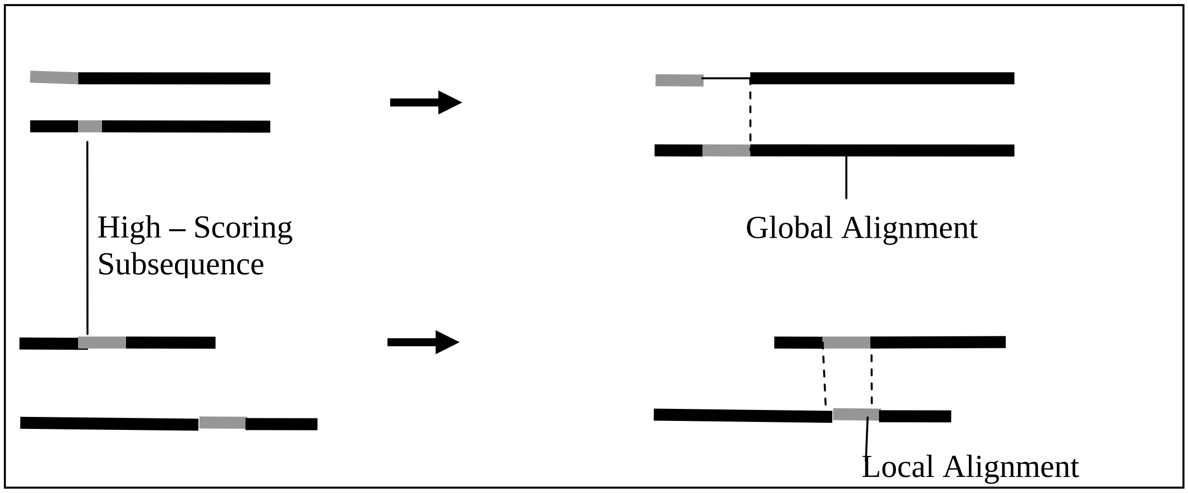


Figure 3.1: Local (top) versus Global (bottom) Alignment. In local alignment, the alignment of local, high – scoring sequences takes precedence over the overall alignment. In global alignment, the best overall alignment is sought, regardless of whether local, high-scoring subsequences are in alignment or not.

3.3.2 Methods of Sequence Alignment

There are several approaches for conducting sequence alignments. Many of these methods are heuristics methods. One of them is Dynamic Programming (DP) method for sequence alignment.

Dynamic Programming

In genetics, sequence alignment is an important application where dynamic programming is essential. Typically, the problem consists of transforming one sequence into another using edit operation that replace, insert, or remove an element. Each operation has an associated cost, and the goal is to find the sequence of edits with the lowest total cost.

The problem can be stated naturally as a recursion, a sequence A is optimally edited into a sequence B by either:

1. inserting the first character of B, and performing an optimal alignment of A and the tail of B
2. deleting the first character of A, and performing the optimal alignment of the tail of A and B
3. replacing the first character of A with the first character of B, and performing optimal alignments of the tails of A and B.

The partial alignments can be tabulated in a matrix, where cell (i, j) contains the cost of the optimal alignment of A [1...i] to B [1...j]. The cost in cell (i, j) can be calculated by adding the cost of the relevant operations to the cost of its neighboring cells, and selecting the optimum.

Dynamic programming is a form of recursion in which intermediate results are saved in a matrix. The comparison can solve complex mathematical equations, with the results of one equation feeding the input of another. With dynamic programming, the intermediate results can be recorded and next equation can be solved with regard to following equations.

The function has been used to illustrate the value of dynamic programming in sequence alignment:

$$MaxValue = f\,(A_i, B_j)$$

In this equation, *MaxValue* is function of variables A_i and B_j, where *i* and *j* are indices to the variables defined in tree structure illustrated in **Figure 3.3**. That is, the possible values of A_i are represented by A_1 through A_5, and possible values of B_j are represented by B_1 through B_{11}. The best solution to *MaxValue* depends on the equation that defines *MaxValue*. Following possible value of *MaxValue* have been considered as example:

$$MaxValue = (A_i \times B_j)$$

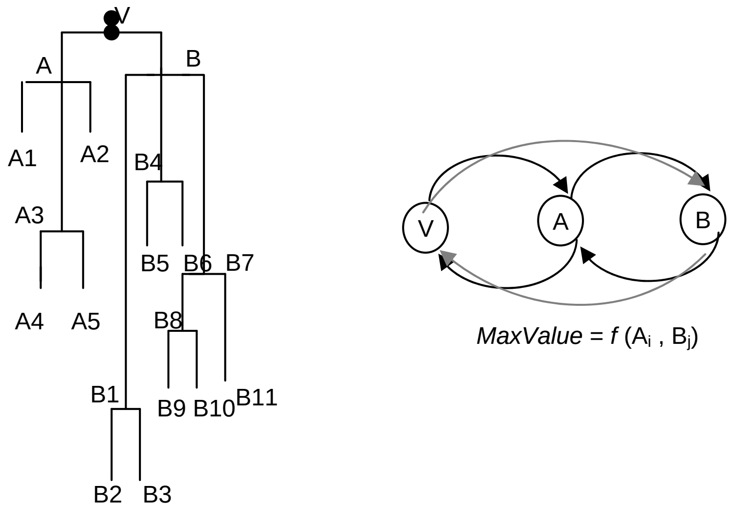


Figure 3.2: Dynamic Programming Problems. Values for A and B are defined in the tree structure. Maximizing *MaxValue* requires evaluating the equation for every combination of *i* and *j*.

In this example, the solution is simply the largest value of A and the largest values of B. However, following definition of *MaxValue* has been considered:

$$Ma\,\mathrm{x}Value = 3\sqrt{\frac{14\times A^2}{\log(A^2+B^2)}}$$

In Brute- Force Methods of solving for *MaxValue* every combinations of A and B has been required and each value has been found out recursively and defined in the tree structure.

For evaluating every value of B in the *MaxValue* equation needs evaluating every value of A that has been illustrated in **Figure 3.3**. For example, assume that the values of A_i and B_j are defined as:

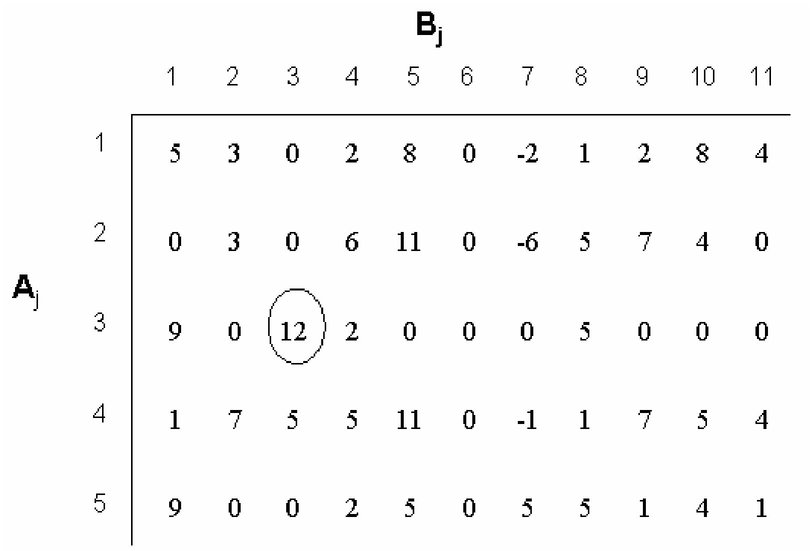
$$A=\begin{bmatrix}2\\3\\8\\4\\1\end{bmatrix} \qquad B=\begin{bmatrix}11\\1\\0\\3\\8\\1\\7\\5\\3\\2\end{bmatrix}$$

Solving for the first value of A_i ($A_1 = 2$) and ignoring the specific equation for MaxValue for Clarity:

MaxValue(1,1) =	$f(A_1, B_1)$ =	$f(2,9)$ =	5
MaxValue(1,2) =	$f(A_1, B_2)$ =	$f(2,11)$ =	3
MaxValue(1,3) =	$f(A_1, B_3)$ =	$f(2,1)$ =	0
MaxValue(1,4) =	$f(A_1, B_4)$ =	$f(2,0)$ =	2
MaxValue(1,5) =	$f(A_1, B_5)$ =	$f(2,3)$ =	8
MaxValue(1,6) =	$f(A_1, B_6)$ =	$f(2,8)$ =	0
MaxValue(1,7) =	$f(A_1, B_7)$ =	$f(2,1)$ =	-2
MaxValue(1,8) =	$f(A_1, B_8)$ =	$f(2,7)$ =	1
MaxValue(1,9) =	$f(A_1, B_9)$ =	$f(2,5)$ =	2
MaxValue(1,10) =	$f(A_1, B_{10})$ =	$f(2,3)$ =	8
MaxValue(1,11) =	$f(A_1, B_{11})$ =	$f(2,2)$ =	4

If the branches of A and B have hundreds of sub-branches, representing hundreds of values, then problem is likely computationally infeasible.

Dynamic programming can address this computational and time dilemma by creating a matrix to store the values for A_i, B_j and *MaxValue* for each combination of *i* and *j*. For example, consider the solution matrix for *MaxValue* in **Figure 3.3**. The solution set to *MaxValue* computed earlier for A_1 appears in first row of the matrix. Examining only this first row, it can be seen that there are two solutions to *MaxValue*, B_5 and B_{10}, each of which results in a value of 8.



		B_j										
		1	2	3	4	5	6	7	8	9	10	11
	1	5	3	0	2	8	0	-2	1	2	8	4
	2	0	3	0	6	11	0	-6	5	7	4	0
A_j	3	9	0	12	2	0	0	0	5	0	0	0
	4	1	7	5	5	11	0	-1	1	7	5	4
	5	9	0	0	2	5	0	5	5	1	4	1

Figure 3.3: Solution Matrix for *MaxValue* for A_i and B_j. The solution to *MaxValue* is A_3 and B_3 with *MaxValue* = 12

MaxValue has been considered as aligned score for pair wise alignment of two sequences. *MaxValue* takes into account gap penalties, correct alignments and imperfect alignments. After the matrix is filled in using the alignment score to determine *MaxValue*, the highest scoring path is followed back to the beginning of the alignment to define the best alignment of elements in sequence, including gaps.

Example of Global Sequence Alignment using Dynamic Programming

The following is an example of global sequence alignment using Needleman/Wunsch techniques. For this example, the two sequences to be globally aligned are

G A A T T C A G T T A (sequence #1)
G G A T C G A (sequence #2)

So M = 11 and N = 7 (the length of sequence #1 and sequence #2, respectively)

A simple scoring scheme is assumed where

- $S_{i,j}$ = 1 if the residue at position i of sequence #1 is the same as the residue at position j of sequence #2 (match score); otherwise
- $S_{i,j}$ = 0 (mismatch score)
- w = 0 (gap penalty)

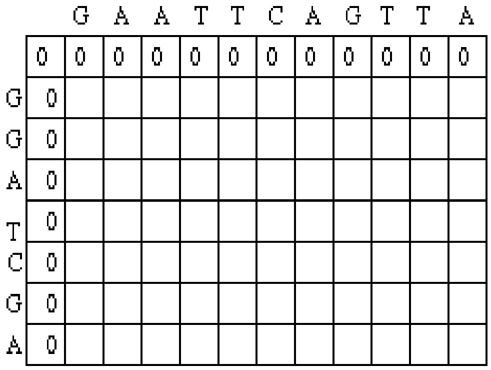
Three steps in dynamic programming

1. Initialization
2. Matrix fill (scoring)
3. Traceback (alignment)

Initialization Step

The first step in the global alignment dynamic programming approach is to create a matrix with M + 1 columns and N + 1 rows where M and N correspond to the size of the sequences to be aligned.

Since this example assumes there is no gap opening or gap extension penalty, the first row and first column of the matrix can be initially filled with 0.



		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0											
G	0											
A	0											
T	0											
C	0											
G	0											
A	0											

Matrix Fill Step

One possible (inefficient) solution of the matrix fill step finds the maximum global alignment score by starting in the upper left hand corner in the matrix and finding the maximal score $M_{i,\, j}$ for each position in the matrix. In order to find $M_{i,j}$ for any i,j it is minimal to know the score for the matrix positions to the left, above and diagonal to i, j. In terms of matrix positions, it is necessary to know $M_{i-1,j}$, $M_{i,j-1}$ and $M_{i-1,\, j-1}$.

For each position, $M_{i,j}$ is defined to be the maximum score at position i,j; i.e.

$M_{i,\, j}$ = MAXIMUM [

$M_{i-1,\, j-1}$ + $S_{i,\, j}$ (match/mismatch in the diagonal),

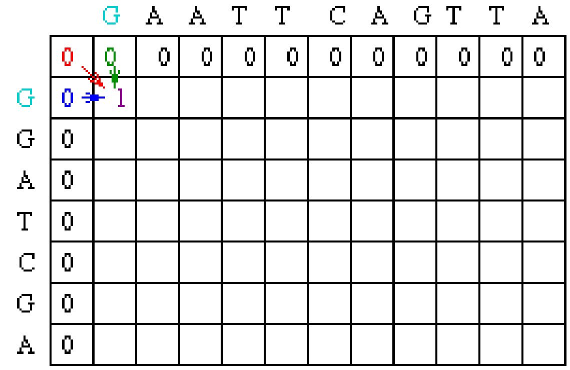
$M_{i,\, j-1}$ + w (gap in sequence #1),

$M_{i-1,\, j}$ + w (gap in sequence #2)**]**

In the example, $M_{i-1,j-1}$, $M_{i,j-1}$ and $M_{i-1,j}$ has been considered red, green and blue respectively.

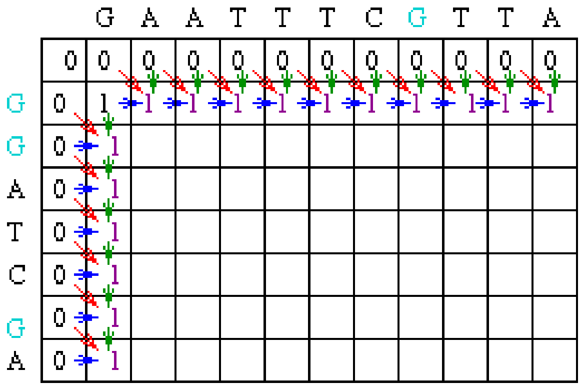
Using this information, the score at position 1, 1 in the matrix can be calculated. Since the first residue in both sequences is a G, $S_{1,1}$ = 1, and by the assumptions stated at the beginning, w = 0. Thus, $M_{1,\, 1}$ = MAX [$M_{0,\, 0}$ + 1, $M_{1,\, 0}$ + 0, $M_{0,\, 1}$ + 0] = MAX [1, 0, 0] = 1.

A value of 1 is then placed in position 1, 1 of the scoring matrix.



		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1										
G	0											
A	0											
T	0											
C	0											
G	0											
A	0											

Since the gap penalty (w) is 0, the rest of row 1 and column 1 can be filled in with the value 1. Take the example of row 1. At column 2, the value is the max of 0 (for a mismatch), 0 (for a vertical gap) or 1 (horizontal gap). The rest of row 1 can be filled out similarly until we get to column 8. At this point, there is a G in both sequences (light blue). Thus, the value for the cell at row 1 column 8 is the maximum of 1 (for a match), 0 (for a vertical gap) or 1 (horizontal gap). The value will again be 1. The rest of row 1 and column 1 can be filled with 1 using the above reasoning.

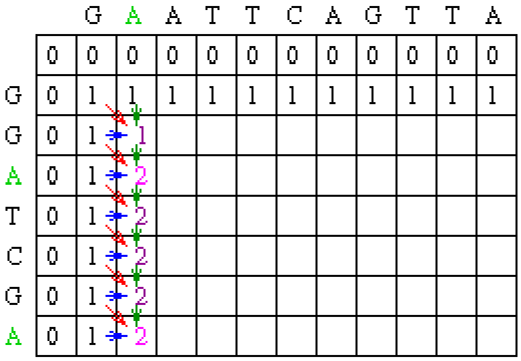


		G	A	A	T	T	T	C	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1										
A	0	1										
T	0	1										
C	0	1										
G	0	1										
A	0	1										

Then after column 2 will be considered. The location at row 2 will be assigned the value of the maximum of 1(mismatch), 1(horizontal gap) or 1 (vertical gap). So its value is 1.

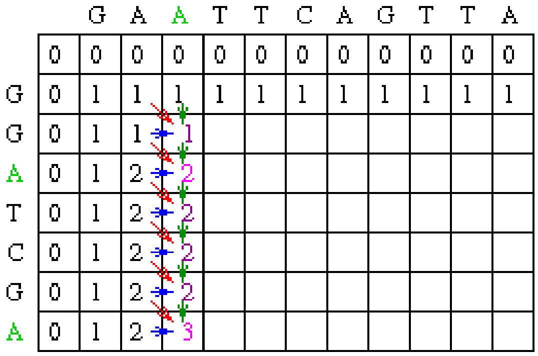
At the position column 2 row 3, there is an A in both sequences. Thus, its value will be the maximum of 2(match), 1 (horizontal gap), 1 (vertical gap) so its value is 2.

Moving along to position column 2 row 4, its value will be the maximum of 1 (mismatch), 1 (horizontal gap), 2 (vertical gap) so its value is 2. For all of the remaining positions except the last one in column 2, the choices for the value will be the exact same as in row 4 since there are no matches. The final row will contain the value 2 since it is the maximum of 2 (match), 1 (horizontal gap) and 2(vertical gap).



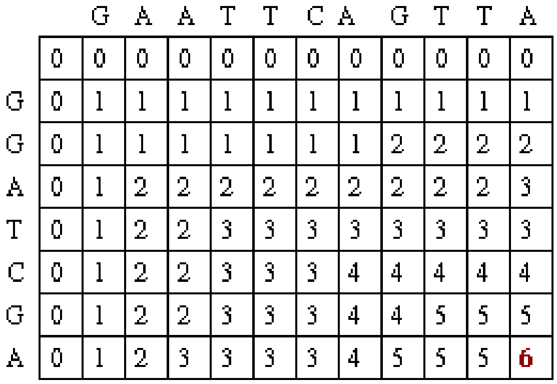
		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1									
A	0	1	2									
T	0	1	2									
C	0	1	2									
G	0	1	2									
A	0	1	2									

Using the same techniques as described for column 2, column 3 has to be filled.



		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1								
A	0	1	2	2								
T	0	1	2	2								
C	0	1	2	2								
G	0	1	2	2								
A	0	1	2	3								

After filling in all of the values the score matrix is as follows:



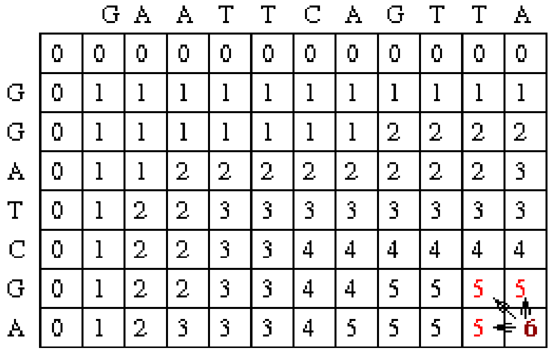
		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	3	4	4	4	4	4
G	0	1	2	2	3	3	3	4	4	5	5	5
A	0	1	2	3	3	3	3	4	5	5	5	6

Trace back Step

After the matrix fill step, the maximum alignment score for the two test sequences is 6. The traceback step determines the actual alignment(s) that result in the maximum score. With a simple scoring algorithm such as one that is used here, there are likely to be multiple maximal alignments.

The traceback step begins in the M, J position in the matrix, i.e. the position that leads to the maximal score. In this case, there is a 6 in that location.

Traceback takes the current cell and looks to the neighbor cells that could be direct predecessors. This means it looks to the neighbor to the left (gap in sequence #2), the diagonal neighbor (match/mismatch), and the neighbor above it (gap in sequence #1). The algorithm for traceback chooses as the next cell in the sequence one of the possible predecessors. In this case, the neighbors are marked in red. They are all also equal to 5.

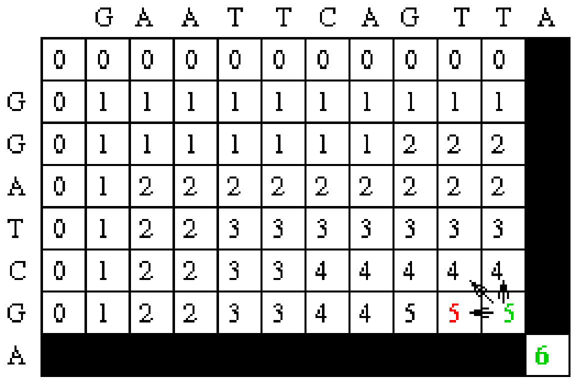


		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	1	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5	5
A	0	1	2	3	3	3	4	5	5	5	5	6

Since the current cell has a value of 6 and the scores are 1 for a match and 0 for anything else, the only possible predecessor is the diagonal match/mismatch neighbor. If more than one possible predecessor exists, any can be chosen. This gives us a current alignment of

```
(Seq #1)      A
              |
(Seq #2)      A
```

So now we look at the current cell and determine which cell is its direct predecessor. In this case, it is the cell with the red 5.

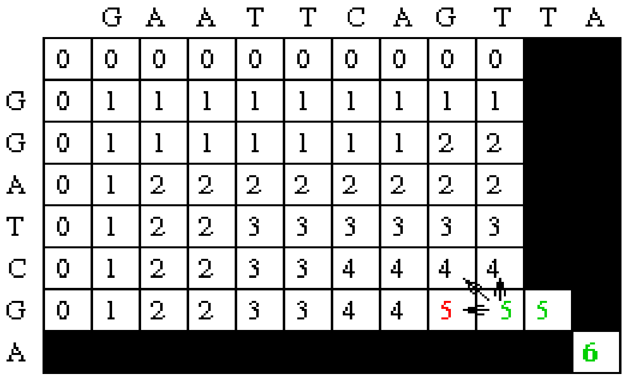


		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	
G	0	1	1	1	1	1	1	1	1	1	1	
G	0	1	1	1	1	1	1	1	2	2	2	
A	0	1	2	2	2	2	2	2	2	2	2	
T	0	1	2	2	3	3	3	3	3	3	3	
C	0	1	2	2	3	3	4	4	4	4	4	
G	0	1	2	2	3	3	4	4	5	5	5	
A												6

The alignment as described in the above step adds a gap to sequence #2, so the current alignment is

```
(Seq #1)      T A
                |
(Seq #2)      _ A
```

Once again, the direct predecessor produces a gap in sequence #2.



		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0		
G	0	1	1	1	1	1	1	1	1	1		
G	0	1	1	1	1	1	1	1	2	2		
A	0	1	2	2	2	2	2	2	2	2		
T	0	1	2	2	3	3	3	3	3	3		
C	0	1	2	2	3	3	4	4	4	4		
G	0	1	2	2	3	3	4	4	5	5	5	
A												6

After this step, the current alignment is

```
(Seq #1)      T T A
                  |
              _ _ A
```

Continuing on with the traceback step, a position in column 0 row 0 has been found which indicatives completion of traceback. One possible maximum alignment is:

		G	A	A	T	T	C	A	G	T	T	A
	0											
G		1										
G			1									
A				2	2							
T						3						
C							4	4				
G									5	5	5	
A												6

Giving an alignment of:

```
G A A T T C A G T T A
|   |   | |   |     |
G G A _ T C _ G _ _ A
```

An alternate solution is:

		G	A	A	T	T	C	A	G	T	T	A
	0											
G		1.										
G		1	1									
A				2	2							
T						3						
C							4	4				
G									5	5	5	
A												6

Giving an alignment of :

```
G _ A A T T C A G T T A
|     |   | |   |     |
G G _ A _ T C _ G _ _ A
```

There are more alternative solutions each resulting in a maximal global alignment score of 6. Since this is an exponential problem, only a single solution has been printed by most of dynamic programming algorithms.

3.4 Multiple Sequence Alignment

Multiple sequence alignment [2], in which three or more sequences must be aligned, is useful in finding patterns in nucleotide sequences and for identifying structural and functional domains in protein families.

Multiple sequence alignment is as an extension of the pair-wise alignment. The first step in multiple sequence alignment is pair-wise alignment of all the sequences. For example four sequences- S_1, S_2, S_3 and S_4 have been considered. The alignment of these sequences involves 5 pair-wise comparisons (S_1 and $S_{2,}$ S_1 and S_3 , S_1 and S_4 , S_2 and S_3 , S_2 and S_4 and S_3 and $S_{4.}$. This is shown in **Figure 3.4**. The result of this alignment has been represented as a tree or a dendogram.

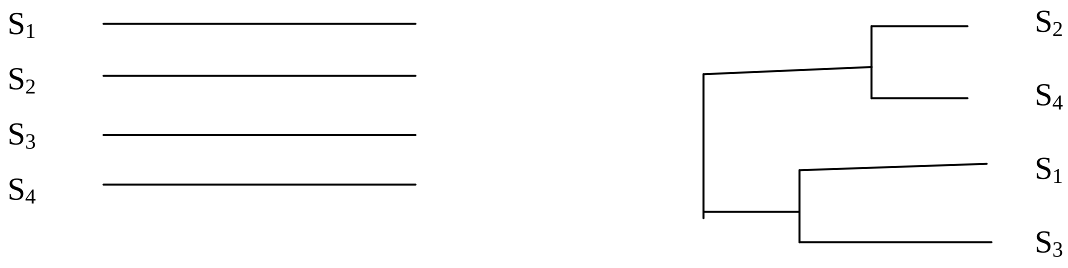


Figure 3.4: Pair-wise alignment of multiple sequences (6 pairwise comparisons then cluster analysis)

Then process of aligning the multiple sequences step-wise has been followed. The first step has been used to align the pair of most similar sequences, followed by less similar and so on. The gaps in the alignments have been used to optimize the alignment. This has been shown in **Figure 3.5**.

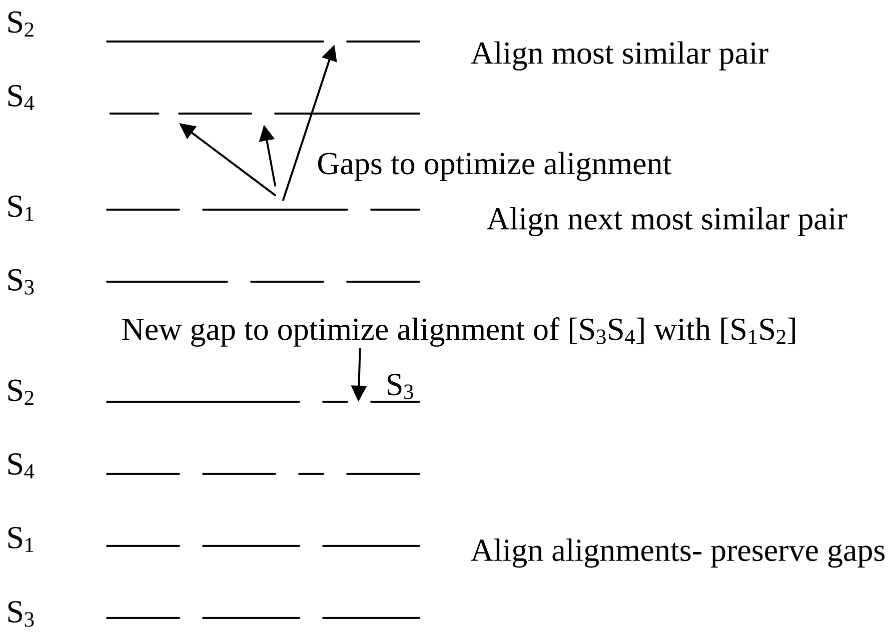

S2
S4
S1
S3
Align most similar pair
Gaps to optimize alignment
Align next most similar pair
New gap to optimize alignment of [S3S4] with [S1S2]
S2
S3
S4
S1
S3
Align alignments- preserve gaps

Figure 3.5: Step-wise alignment of sequences

3.4.1 Methods of Multiple Sequence Alignment

There are several approaches for conducting multiple sequence alignment. The most common approach to approach to multiple sequence alignment is Progressive Alignment.

Progressive Alignment Method

The approach of progressive alignment is to begin with an alignment of most alike sequences and then builds upon the alignment using other sequences. Progressive alignments work by first aligning the most alike sequences using dynamic programming and then progressively adding less related sequences to the initial alignment.

Cluster pair-wise alignment is a simple score between extensions of sequence alignment. In a pair-wise alignment, the comparison score between any two positions in those clusters is the arithmetic average of scores for all possible symbol comparisons at those positions. When gaps have inserted into a cluster to produce an alignment, they have been inserted at the same position in all of the sequences of the cluster. The full multiple alignments are obtained once all the sequences have been clustered into one. This hierarchical clustering can be plotted as a dendogram.

Since the alignment is calculated on a progressive basis, the order of the initial sequences can affect the final alignment. Different comparison matrices or gap weights affects the multiple alignments.

Consider the 5 sequences given in **Table 3.1** that need to be aligned.

Table 3.1: Set of 5 sequences

S_1	T	C	Y	G	I	F	V	L		
S_2	T	C	G	I	F	V	L			
S_3	S	C	Y	G	I	F	V	L	S	G
S_4	T	C	F	G	I	F	V	L		
S_5	A	C	G	I	F	V	L	S	G	

All pair-wise comparisons are performed, resulting in a matrix of scores. The scores have been represented in **Table 3.2**.

Table 3.2: Scores for the Pair-wise Comparisons

	S_1	S_2	S_3	S_4	S_5
S_1		26	38	38	26
S_2	26		26	26	32
S_3	38	26		36	36
S_4	38	26	36		26
S_5	26	32	36	26	

The most closely-related pair of sequences is aligned first. In this example S_1 and S_3 and S_1 and S_4 have the same score, so they can be used as the first pair. The interactions are shown in **Table 3.3**.

Table 3.3: Steps in Aligning 5 Sequences Given Above

There are four steps.

Step 1 – start with S_1 and S_3

S_1	T	C	Y	G	I	F	V	L	-	-
S_3	S	C	Y	G	I	F	V	L	S	G

Step 2 – add S_4

S_1	T	C	Y	G	I	F	V	L	-	-
S_3	S	C	Y	G	I	F	V	L	S	G
S_4	T	C	F	G	I	F	V	L	-	-

Step 3 – add S_2

S_1	T	C	Y	G	I	F	V	L	-	-
S_3	S	C	Y	G	I	F	V	L	S	G
S_4	T	C	F	G	I	F	V	L	-	-
S_2	T	C	-	G	I	F	V	L	-	-

Step 4 – add S_5

S_1	T	C	Y	G	I	F	V	L	-	-
S_3	S	C	Y	G	I	F	V	L	S	G
S_4	T	C	F	G	I	F	V	L	-	-
S_2	T	C	-	G	I	F	V	L	-	-
S_5	A	C	G	I	F	V	L	S	G	-

The alignment can be thought of as occurring in a "star" configuration, where the sequence with the greatest similarity to the others is at the centre and the rays of the star represent the pair-wise distances to the remaining sequences. Each time a sequence is added, gaps are inserted either in newly added sequence or in the entire alignment to optimize alignment.

The first problem with progressive methods is that they depend upon the initial pair-wise sequence alignments. If the sequences are closely related then the likelihood is good that the initial alignment contains relatively few errors. However, if initial sequences are distantly related, then there will be more errors in the alignment, which will propagate through the rest of the alignments. The second problem is that suitable scoring matrices and gap penalties must be chosen to apply to the sequences as a set.

3.4.2 Application of Multiple Alignments

The basic information from a multiple alignment of protein sequences is the position and nature of the conserved regions in each member of the group. Conserved sequence regions correspond to functionally and structurally important parts of the protein. Hypotheses about functional importance or specific roles can then be directly tested by mutagenesis and truncation experiments.

Multiple sequence alignments can be used to find regions of similar sequence in all of the sequences that defines a conserved consensus pattern. If the alignment is strong, MSA can also be used to derive the possible evolutionary relationships among the sequences.

Multiple alignments are powerful tools for identifying new members of the aligned group. It is possible to query databases of multiple alignments with single sequences and to query sequence databases with multiple alignments. It has been observed that such searches are more sensitive and selective than sequence-to-sequence searches.

Multiple alignments of many sequences and those with different sequence weights are difficult to visualize. Sequence logos are a graphical way for presenting multiple alignments. A different graphical view of multiply aligned sequences is by a tree relating their sequence similarity. This is very useful when the aligned sequences are of several functional subtypes and we wish to know to which one our sequence/s belongs. A way to estimate the significance of a tree is by bootstrap values. These values have been used to find number of times branching point have been observed with different models of the input data. The higher the fraction of the bootstrap value (number of observations/number of trials) indicates that the sequences emerging from that branch point cluster together.

Multiple alignments are powerful tools for identifying new members of the aligned group. It is possible to query databases of multiple alignments with single sequences and to query sequence databases with multiple alignments. It has been seen that such searches are more sensitive and selective than sequence-to-sequence searches.

The **Blimps** program has been used to query both protein and nucleotide sequence databases with protein blocks and vice versa. The queries are single sequences or blocks. The program is available on the WWW and by e-mail server for searching multiple alignment databases with single sequences.

The **MAST** program [3]has been used to query sequence databases with blocks. Protein or nucleotide databases are queried with protein blocks. The query can be a single block or all the blocks of a protein family.

The **LAMA** program has been used to search blocks databases with block queries. Queries can be obtained from the Blocks database, BlockMaker program or by reformatting multiple alignments.

3.5 Phylogenetic Analysis

The sequencing of DNA and proteins has become easy and fast with the use of automated tools. Similarity searches and multiple alignments of sequences have been used to find the relationship between the sequences.

Based on given a set of sequences, the evolutionary relationship among genes has been reconstructed. To reconstruct the evolutionary relationship, a branched structure termed a phylogeny or tree has been created. A phylogeny illustrates the relationship between the sequences. Analysis of phylogeny of a family has been done to know about molecular evolution. Molecular evolution is based on mutations imparting the DNA to derive during evolution. The number of types of changes in residues of a MSA can be used to start a phylogenetic analysis. Each column in a MSA denotes mutations that occur at one site during evolution of sequence family.

Phylogenetic Trees

The method [4] of converting MSA to a phylogenetic tree is used to reduce the problem of a multiple alignment to an iterative process of pair-wise alignments. The purpose works as follows:

- Compute all pair-wise distance between given sequence
- Compute a tree by single linkage clustering by using methods like UPGNA or Nearest Neighbor.
- Align the sequences in an orderly fashion.

An evolutionary relationship has been represented using phylogenetic trees. A tree is 2D graph showing evolutionary relationships among organisms. This separate source of sequences has been referred as taxa, defined as phylogenetically distinct units on the tree. The tree is composed of nodes representing the taxa and branches representing the relationships among the taxa. An example of a rooted tree of 4 taxa is shown in **Figure 3.4.**

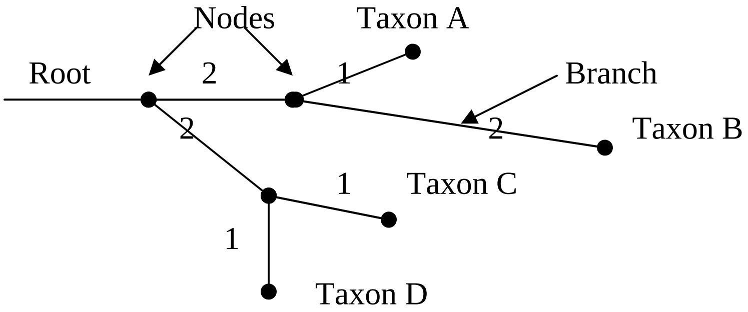


Figure 3.6: Example of a rooted tree of 4 taxa showing branch lengths proportional to the number of changes in branch

The most used terms in phylogenetic analysis are given in **Table 3.4**.

Table 3.4: Terms Used in Phylogenetic Analysis

Node	*A Node Represents a Taxonomic Units it can be a Taxon*
Branch	Defines the relationship between the taxa in terms of descent and ancestry.
Topology	Is the branching pattern
Branch length	Often represents the number of chances that have occurred in that branch.
Root	Is the common ancestor of all taxa.
Distance scale	Scale which represents the number of differences between sequences

3.5.1 Methods of Phylogenetic Analysis

There are two approaches to deriving phylogenetic trees. One approach makes no reference no reference to any historical model of relationships. Proceed by measuring a set of distances between species and generate the tree by a hierarchical *Clustering procedure*. This is called the **phonetic** approach. The alternative, the *cladistic approach*, is to consider possible pathways of evolution, infer the features of ancestor at each node and choose an optimal tree according to some model of evolutionary change. Phenetics is based on similarity; cladistics is based on genealogy.

Clustering Methods

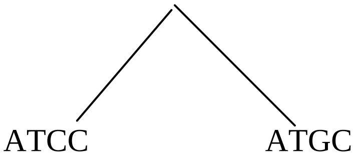
Clustering approaches to determination of phylogenetic relationships are explicitly non-historical. A simple clustering procedure works as follows: Given a set of species, determine for all pairs a measure of similarity or difference between them. To create a tree from the set of dissimilarities, first choose the two most closely related species and insert a node to represent their common ancestor. Then replace the two selected species by a set containing both and replace the distances from the pair to the others by the average of distances of the two selected species to the others. Now we have a set of pairwise dissimilarities not between individual species but between sets of species. (Regard each remaining individual species as a set containing only one element.) Then repeat the process, as the following example.

Example: Consider four species characterized by homologous sequences ATCC, ATGC, TTCG and TCGG. Taking the number of differences as the measure of dissimilarity between each pair of species, use a simple clustering procedure to derive a phylogenetic tree.

The distance matrix is:

	ATCC	ATGC	TTCG	TCGG
ATCC	0	1	2	4
ATGC		0	3	3
TTCG			0	2
TCGG				0

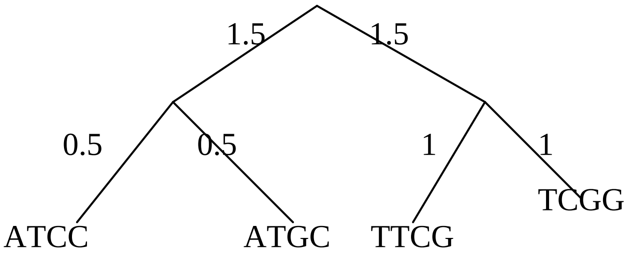
Because the matrix is symmetric, we need fill in only the upper half. The smallest nonzero distance is 1, between ATCC and ATGC. Therefore our first cluster is {ATCC, ATGC}. The tree will contain the fragment:



The reduced distance matrix is:

	(ATCC, ATGC)	TTCG	TCGG
(ATCC, ATGC)	0	½(2+3) =2.5	½(4+3) =3.5
TTCG		0	2
TCGG			0

The next cluster is {TTCG, TCGG}, distance 2. Finally, linking the clusters {ATCC, ATGC} and {TTCG, TCGG} gives the tree:



Branch lengths have been assigned according to the rule:

Branch length of edge between nodes X and Y = ½ distance between X and Y.

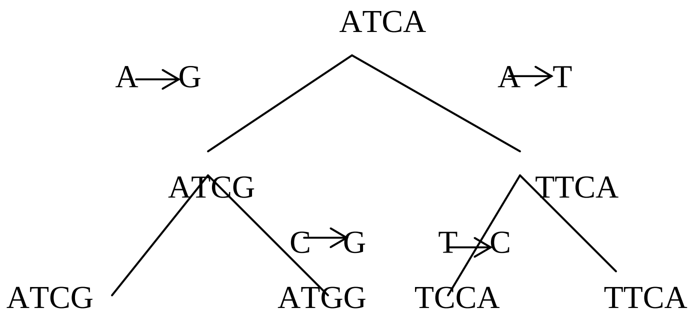
Whether the branch lengths are truly proportional to divergence times of the taxa represented by the nodes must be determined from external evidence.

This process of tree building is called the UPGMA method (Unweighted Pair Group Method with Arithmetic mean).

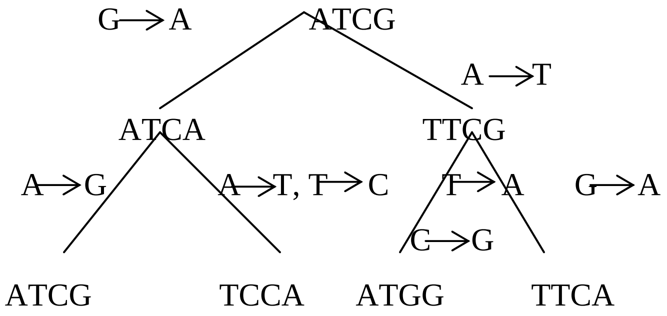
Cladistic Methods

Cladistic methods [5] deal explicitly with the patterns of ancestry implied by the possible trees relating a set of taxa. Their aim is to select the correct tree by utilizing an explicit model of the evolutionary process. The most popular cladistic methods in molecular phylogency are the *maximum parsimony* and *maximum likelihood* approaches. They are specialized to sequence to sequence data, starting from a multiple sequence alignment.

The *maximum parsimony* [6] method defines an optimal tree as the one that postulates the fewest mutations. For instance, given species characterized by homologous sequences ATGC, ATGG, TCCA and TTCA, the tree postulates four mutations:



An alternative tree postulates seven mutations.



Note that the second tree implies that the G → A mutation in the forth position occurred twice independently. The former tree is optimal according to the maximum parsimony method, because no other tree involves fewer mutations.

The *maximum likelihood* method assigns quantitative probabilities to mutational events, rather than merely counting them. Like maximum parsimony, maximum likelihood reconstructs ancestors at all nodes of each tree considered; but it also assigns branch lengths based on probabilities of mutational events. For each possible tree topology, the assumed substitution rates are varied to find the parameters that give the highest likelihood of producing the observed sequences. The optimal tree is the one with the highest likelihood of generating the observed data.

3.5.2 Computational Considerations

Cladistic methods- maximum parsimony and maximum likelihood requires large amounts of computer time. By considering Cladistic methods the total number of possible trees, increases very rapidly with the number of species.

Calculated phylogenies are often approximations, methods for testing them are:

1. Comparison of phylogenies obtained from different characters describing the same set of taxa. If trees produced from different characters share a subtree,

perhaps that portion of the phylogency has been determined reliably and other portions have not.

2. Analysis of subsets of taxa should give the same answer – respect to the subset- as appears within the full tree.
3. Formal statistical test, involving returning the calculation on subsets of the original data, are known as **jackknifing** and **bootstrapping**:
   - **Jackknifing** is calculation with data sets samples randomly from the original data. For phylogeny calculations from multiple sequence alignments. Select different subsets of the positions in the alignment and return the calculation. Finding that each subset gives the same phylogenetic tree lends it credibility. If such subset gives a different tree, none of them is trustworthy.
   - **Bootstrapping** is similar to Jackknifing except that the positions chosen at random may include multiple copies of the same position to form data sets of the same size as the original to preserve statistical properties of the sampling.
4. If there are very long edges, then this has been considered seriously because there is the possibility of unequal variation in evolutionary rate that may have disturbed the calculation. So outgroup taxa has been introduced to check this.

3.5 References

[1] Bryan Bergeron .“ Bioinformatics Computing” Eastern Economy Edition, 306-310,2003.

[2] S.C. Rastogi, N.Mendiratta, P. Rastogi, “Bioinformatics Methods and Applications”, PHI, 93-100, 2006.

[3] http://bioinformatics.weizmann.ac.il/blocks/process_blocks.html

[4] Arthur M. Lesk. “Introduction to Bioinformatics” Oxford University Press, 203-209, 2005.

[5] http://evolution.genetics.washington.edu/phylip/software.html

[6] Whelan , S. Lio , P. & Goldman, N. , Molecular Phylogenetics : State –of-the-art methods for looking into past, Trends in Genetics,17 , 262 -272 , 2001

Chapter -4

SimilaritieS Search and Sequence Alignment

Chapter 4
SIMILARITIES SEARCH AND SEQUENCE ALIGNMENT

4.1 FASTA Algorithm

The FASTA algorithm [1] is a heuristic method for string comparison. FASTA compares a query string against a single text string. When searching the whole database for matches to a given query, the query using the FASTA algorithm to every string in the database has been compared.

When looking for an alignment, a few segments have been found in which there has absolute identity between the two compared strings. The algorithm is using this property and focuses on these identical regions.

The stages in the FASTA algorithm are as follows:

1. An integer parameter called *ktup* (short for *k respective tuples*), has been specified and *ktup*-length matching substrings of the two strings has been looked. The standard recommended *ktup* values are six for DNA sequence matching and two for protein sequence matching. The matching *ktup*-length substrings are referred to as *hot spots*. Consecutive hot spots are located along the dynamic programming matrix diagonals. This stage can be done efficiently by using a lookup table or a hash to store all the *ktup*-length substrings from one string, and then search the table with the *ktup*-length substrings from the other string.

2. In this stage 10 best *diagonal runs* of hot spots in the matrix has been found out. A diagonal run is a sequence of nearby hot spots on the same diagonal.

In order to evaluate the diagonal runs, FASTA gives each hot spot a positive score, and the space between consecutive hot spots in a run is given a negative score that decreases with the increasing distance. The score of the diagonal run is the sum of the hot spots scores and the interspot scores. FASTA finds the 10 highest scoring diagonal runs under this evaluating scheme.

3. A diagonal run specifies a pair of aligned substrings. The alignment is composed of matches (the hot spots) and mismatches (from the interspot regions), but it does not contain any indels because it is derived from a single diagonal. The runs have been using an amino acid (or nucleotide) substitution matrix, and pick the best scoring run. The single best subalignment found in this stage is called $init_1$. Apart from computing $init_1$, a filtration has been performed and the diagonal runs achieving relatively low scores have been discarded.

4. Until now any indels in the subalignments have not been allowed. Now it has been tried to combine "good" diagonal runs from close diagonals, thus achieving a subalignment with indels allowed. Then after "good" subalignments have been taken from the previous stage (subalignments whose score is above some specified cutoff) and attempt to combine them into a single larger high-scoring alignment that allows some spaces.

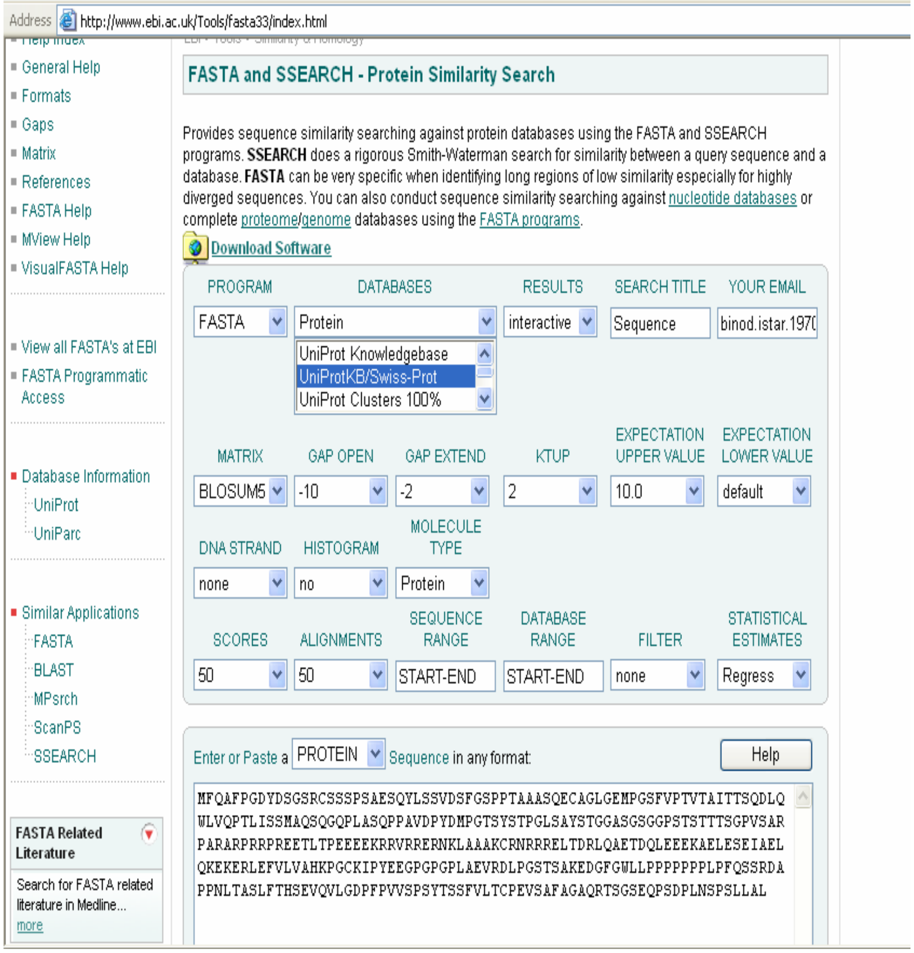
5. In this step FASTA computes an alternative local alignment score, in addition to $init_n$. A diagonal segment has been defined by $init_1$ in the dynamic programming matrix. A narrow diagonal band in the matrix has been considered, centered along this segment. The optimal local alignment in this band has been computed, using the ordinary dynamic programming algorithm. Assuming that the best local alignment is indeed within the defined band, the local alignment algorithm essentially merges diagonal runs found in the previous stages to achieve a local alignment which may contain indels. The band width is dependent on the *ktup* choice. The best local alignment computed in this stage is called *opt*.

6. In the last stage, the database sequences are ranked according to $init_n$ scores or *opt* scores, and the full dynamic programming algorithm is used to align the query sequence against each of the highest ranking result sequences.

Although FASTA is a heuristic, and as such it is possible to show instances in which the alignments found by the algorithm are not optimal, it is claimed that the resulting alignment scores well compare to the optimal alignment, while the FASTA algorithm is much faster than the ordinary dynamic programming alignment algorithm.

4.1.2 FASTA Implementation

FASTA at the EBI is one of the most popular FASTA implementation. The FASTA input form is given in **Figure 4.1**.



http://www.ebi.ac.uk/Tools/fasta33/index.html

FASTA and SSEARCH - Protein Similarity Search

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

PROGRAM	DATABASES	RESULTS	SEARCH TITLE	YOUR EMAIL
FASTA	Protein UniProt Knowledgebase UniProtKB/Swiss-Prot UniProt Clusters 100%	interactive	Sequence	binod.istar.197(

MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
BLOSUM5	-10	-2	2	10.0	default

DNA STRAND	HISTOGRAM	MOLECULE TYPE
none	no	Protein

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

Enter or Paste a PROTEIN Sequence in any format:

```
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTAITTSQDLQ
WLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGSGGPSTSTTTSGPVSAR
PARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELTDRLQAETDQLEEEKAELESEIAEL
QKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRDLPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDA
PPNLTASLFTHSEVQVLGDPFPVVSPSYTSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL
```

Figure 4.1: FASTA page at EBI

The Histogram

The histogram [2] compares the predicted extreme value distribution of local similarity scores, represented by asterisks with the actual number obtained, represented by equal signs. Each bar represents the number of local alignment having the z-opt score indicated in the first column. The second column is the actual number of sequences with that z-opt scores. The third column is the predicted number of alignments having z-scores in that interval.

For histogram for this particular sequence is given in **Figure 4.2**.

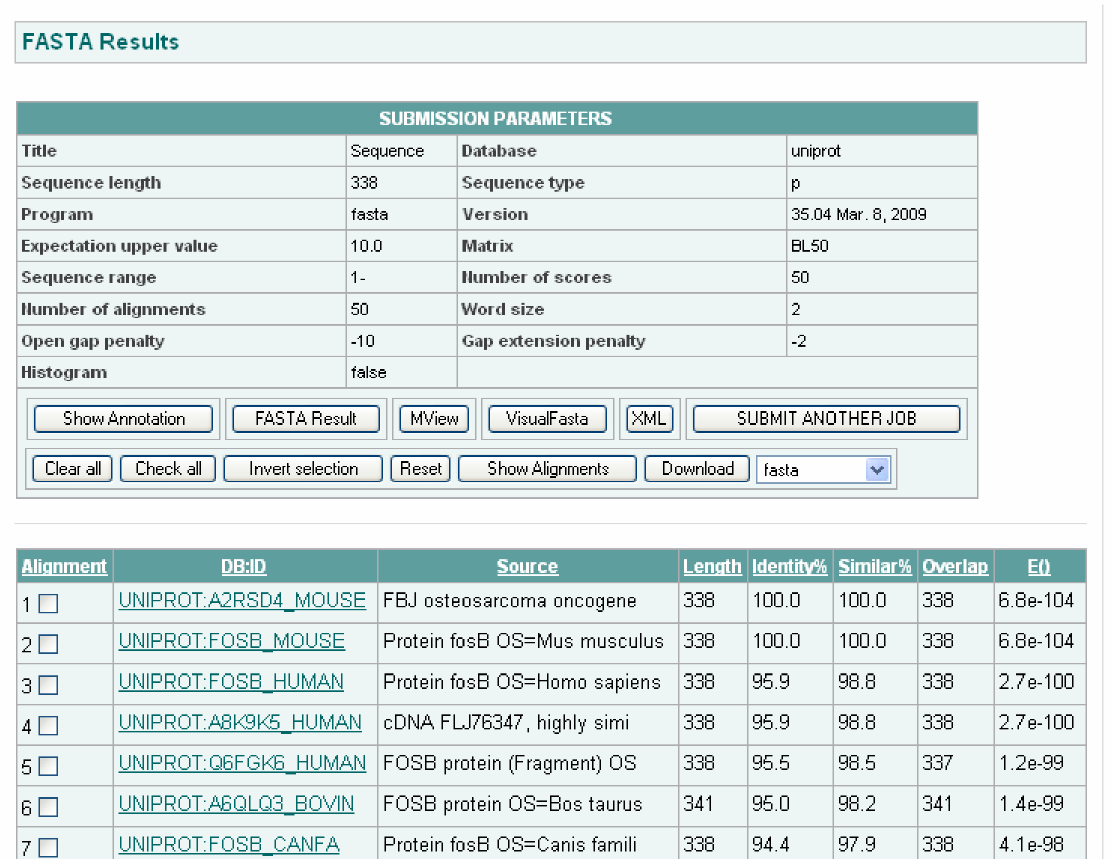
```
•         opt      E()
•   < 20  1040     0:=
•     22     0     0:           one = represents 1534 library sequences
•     24     4     1:*
•     26    14    19:*
•     28    53   201:*
•     30   227  1223:*
•     32  1161  4728:=  *
•     34  5065 12823:====    *
•     36 16019 26336:===========      *
•     38 33416 43523:======================      *
•     40 58656 60711:=======================================*
•     42 81562 74211:================================================*=====
•     44 87125
  81862:=====================================================*===
•     46 92000
  83378:======================================================*===
•     48 83023
  79825:====================================================*==
•     50 83389
  72841:===============================================*=======
•     52 68683 64039:=========================================*===
•     54 58489 54701:===================================*===
•     56 48841 45692:=============================*==
•     58 36330 37512:========================*
•     60 26755 30387:================== *
•     62 21306 24361:============== *
•     64 17453 19374:============*
•     66 13701 15313:=========*
•     68 10436 12045:=======*
•     70  8222  9439:======*
•     72  6047  7376:====*
•     74  5090  5751:===*
```

Figure 4.2: Histogram from FASTA output

The Sequence Listing

The next part of the output is a listing of the best scoring sequences in the database. The best alignments are reported first and the worst hits last.

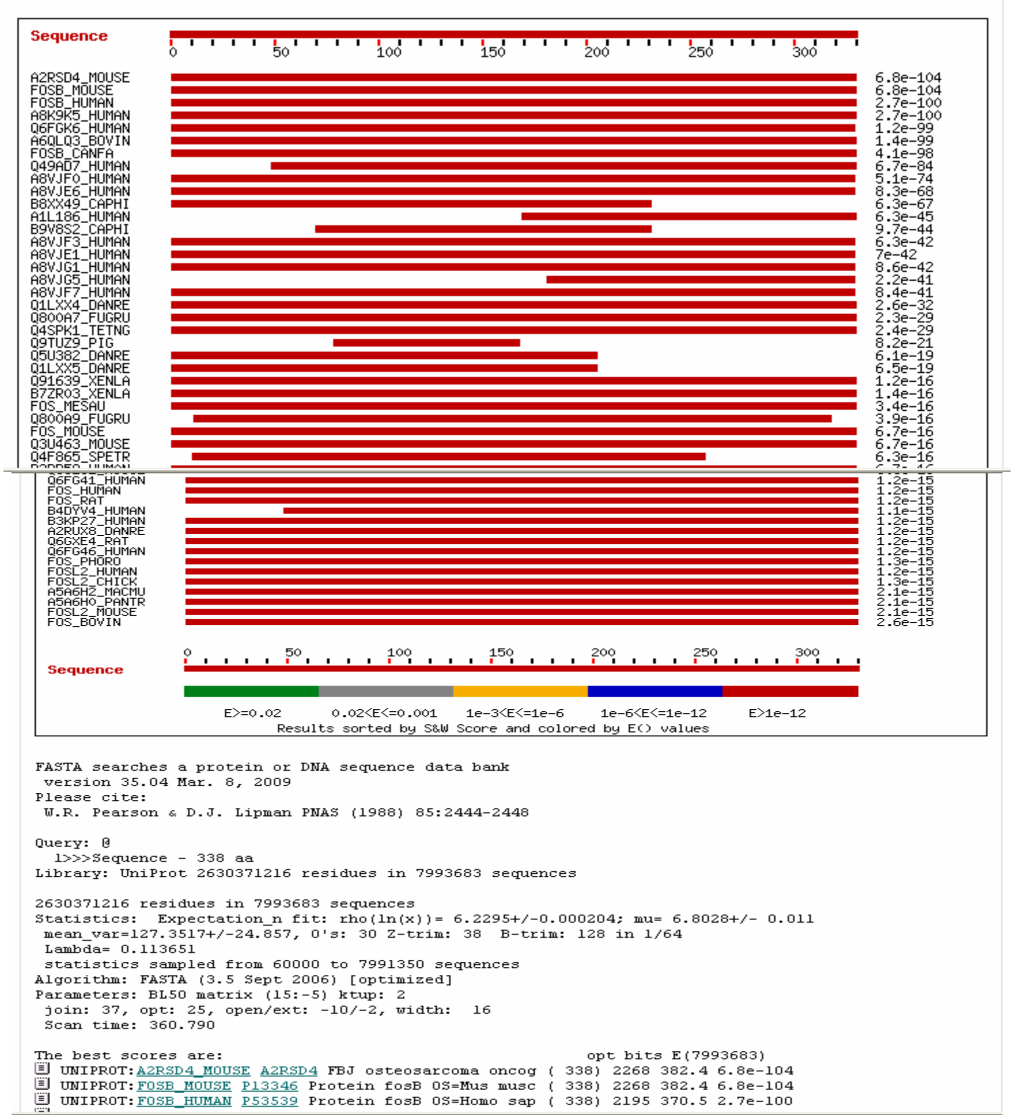
The first column identifies the database sequence reported by database, database accession number and database identifier. The next column reports the total length of database sequence. The final scores are reported in the opt column and the E ( ) value for that particular database sequence is reported in the last column.



SUBMISSION PARAMETERS			
Title	Sequence	Database	uniprot
Sequence length	338	Sequence type	p
Program	fasta	Version	35.04 Mar. 8, 2009
Expectation upper value	10.0	Matrix	BL50
Sequence range	1-	Number of scores	50
Number of alignments	50	Word size	2
Open gap penalty	-10	Gap extension penalty	-2
Histogram	false		

Alignment	DB:ID	Source	Length	Identity%	Similar%	Overlap	E()
1 □	UNIPROT:A2RSD4_MOUSE	FBJ osteosarcoma oncogene	338	100.0	100.0	338	6.8e-104
2 □	UNIPROT:FOSB_MOUSE	Protein fosB OS=Mus musculus	338	100.0	100.0	338	6.8e-104
3 □	UNIPROT:FOSB_HUMAN	Protein fosB OS=Homo sapiens	338	95.9	98.8	338	2.7e-100
4 □	UNIPROT:A8K9K5_HUMAN	cDNA FLJ76347, highly simi	338	95.9	98.8	338	2.7e-100
5 □	UNIPROT:Q6FGK6_HUMAN	FOSB protein (Fragment) OS	338	95.5	98.5	337	1.2e-99
6 □	UNIPROT:A6QLQ3_BOVIN	FOSB protein OS=Bos taurus	341	95.0	98.2	341	1.4e-99
7 □	UNIPROT:FOSB_CANFA	Protein fosB OS=Canis famili	338	94.4	97.9	338	4.1e-98

Figure 4.3: Best Scoring Sequences

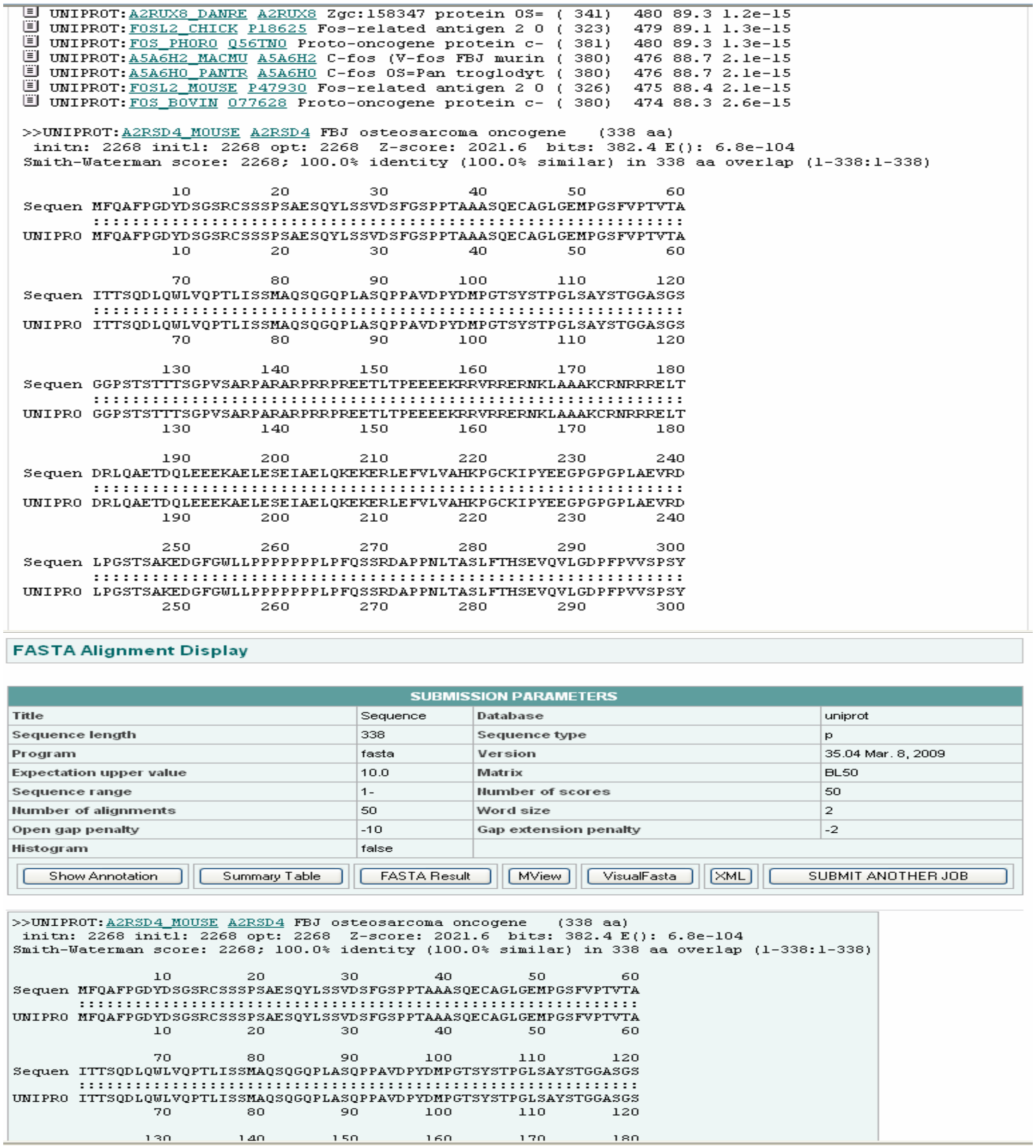


```
FASTA searches a protein or DNA sequence data bank
 version 35.04 Mar. 8, 2009
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

Query: @
  1>>>Sequence - 338 aa
Library: UniProt 2630371216 residues in 7993683 sequences

2630371216 residues in 7993683 sequences
Statistics:  Expectation_n fit: rho(ln(x))= 6.2295+/-0.000204; mu= 6.8028+/- 0.011
 mean_var=127.3517+/-24.857, 0's: 30 Z-trim: 38  B-trim: 128 in 1/64
 Lambda= 0.113651
 statistics sampled from 60000 to 7991350 sequences
Algorithm: FASTA (3.5 Sept 2006) [optimized]
Parameters: BL50 matrix (15:-5) ktup: 2
 join: 37, opt: 25, open/ext: -10/-2, width:  16
 Scan time: 360.790

The best scores are:                                      opt bits E(7993683)
 UNIPROT:A2RSD4_MOUSE A2RSD4 FBJ osteosarcoma oncog ( 338) 2268 382.4 6.8e-104
 UNIPROT:FOSB_MOUSE P13346 Protein fosB OS=Mus musc ( 338) 2268 382.4 6.8e-104
 UNIPROT:FOSB_HUMAN P53539 Protein fosB OS=Homo sap ( 338) 2195 370.5 2.7e-100
```

Figure 4.4: Visual FASTA Result

The Local Alignments

After the list of hits, the actual local alignments identified are displayed. The initn and init1 columns report the local similarity scores calculated at different stages of the FASTA procedure. For each alignment, the various scores and E ( ) values are reported again along with some new information. The Smith-Waterman score reported is the same as the opt score. The percent identity and length of the alignments are displayed.



```
UNIPROT:A2RUX8_DANRE A2RUX8 Zgc:158347 protein OS= ( 341)  480 89.3 1.2e-15
UNIPROT:FOSL2_CHICK P18625 Fos-related antigen 2 O ( 323)  479 89.1 1.3e-15
UNIPROT:FOS_PHORO Q56TN0 Proto-oncogene protein c- ( 381)  480 89.3 1.3e-15
UNIPROT:A5A6H2_MACMU A5A6H2 C-fos (V-fos FBJ murin ( 380)  476 88.7 2.1e-15
UNIPROT:A5A6H0_PANTR A5A6H0 C-fos OS=Pan troglodyt ( 380)  476 88.7 2.1e-15
UNIPROT:FOSL2_MOUSE P47930 Fos-related antigen 2 O ( 326)  475 88.4 2.1e-15
UNIPROT:FOS_BOVIN O77628 Proto-oncogene protein c- ( 380)  474 88.3 2.6e-15

>>UNIPROT:A2RSD4_MOUSE A2RSD4 FBJ osteosarcoma oncogene    (338 aa)
 initn: 2268 init1: 2268 opt: 2268  Z-score: 2021.6  bits: 382.4 E(): 6.8e-104
Smith-Waterman score: 2268; 100.0% identity (100.0% similar) in 338 aa overlap (1-338:1-338)

               10        20        30        40        50        60
Sequen MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
               10        20        30        40        50        60

               70        80        90       100       110       120
Sequen ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
               70        80        90       100       110       120

              130       140       150       160       170       180
Sequen GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
              130       140       150       160       170       180

              190       200       210       220       230       240
Sequen DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
              190       200       210       220       230       240

              250       260       270       280       290       300
Sequen LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
              250       260       270       280       290       300
```

FASTA Alignment Display

SUBMISSION PARAMETERS			
Title	Sequence	Database	uniprot
Sequence length	338	Sequence type	p
Program	fasta	Version	35.04 Mar. 8, 2009
Expectation upper value	10.0	Matrix	BL50
Sequence range	1-	Number of scores	50
Number of alignments	50	Word size	2
Open gap penalty	-10	Gap extension penalty	-2
Histogram	false		

```
>>UNIPROT:A2RSD4_MOUSE A2RSD4 FBJ osteosarcoma oncogene    (338 aa)
 initn: 2268 init1: 2268 opt: 2268  Z-score: 2021.6  bits: 382.4 E(): 6.8e-104
Smith-Waterman score: 2268; 100.0% identity (100.0% similar) in 338 aa overlap (1-338:1-338)

               10        20        30        40        50        60
Sequen MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
               10        20        30        40        50        60

               70        80        90       100       110       120
Sequen ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
               70        80        90       100       110       120

              130       140       150       160       170       180
```

Figure 4.5: Local Alignment Score

Significance of the E-values

FASTA calculates an E-value (expectation of significance). E ( ) values represent the number of sequences having a given alignment occurred by chance. In the above example of FASTA search, the best hits mostly have E ( ) values of zero (0). This can be interpreted as: “we can expect zero sequences to have the score this alignment has, strictly by chance”.

E ( ) values are calculated from the probability derived from the extreme value distribution for each z-opt score interval, and the number of sequences in the database. Therefore, as more sequences are added to a database, the E ( ) values can change and sequences identified in one search may not be found in a later one.

4.2. BLAST Algorithm

BLAST is one of the most widely used bioinformatics programs, because it addresses a fundamental problem and the algorithm emphasizes speed over sensitivity. This emphasis on speed is vital to making the algorithm practical on the huge genome databases currently available, although subsequent algorithms can be even faster.

BLAST is about 50 times faster than dynamic programming; however, it cannot guarantee the optimal alignments of the query and database sequences as in the dynamic programming, but just works to find the related sequences in a database search. BLAST is more time efficient than FASTA by searching only for the more significant patterns in the sequences, but with comparative sensitivity.

Algorithm

To run, BLAST requires a query sequence to search for, and a sequence to search against (also called the target sequence) or a sequence database containing multiple such sequences. BLAST finds subsequences in the database which are similar to subsequences in the query. In typical usage, the query sequence is much smaller than the database, e.g., the query may be one thousand nucleotides while the database is several billion nucleotides.

The main idea of BLAST is that there are often high-scoring segment pairs (HSP) contained in a statistically significant alignment. BLAST searches for high scoring sequence alignments between the query sequence and sequences in the database using a heuristic approach that approximates the Smith-Waterman algorithm. The exhaustive Smith-Waterman approach is too slow for searching large genomic databases such as

GenBank. Therefore, the BLAST algorithm uses a heuristic approach that is less accurate than the Smith-Waterman but over 50 times faster.

An overview of the BLASTP algorithm (a protein to protein search) is as follows:

1. **Remove low-complexity region or sequence repeats in the query sequence.** Low-complexity region means a region of a sequence is composed of few kinds of elements. These regions might give high scores that confuse the program to find the actual significant sequences in the database, so they should be filtered out. The regions is to be marked with an X (protein sequences) or N (nucleic acid sequences) and then be ignored by the BLAST program. To filter out the low-complexity regions, the SEG program is used for protein sequences and the program DUST is used for DNA sequences.
2. **Make a k-letter word list of the query sequence.**

   The words is listed of length 3 in the query protein sequence by taking k=3(k is usually 11 for a DNA sequence) "sequentially", until the last letter of the query sequence is included. The method can be illustrated in **Figure 4.6**.

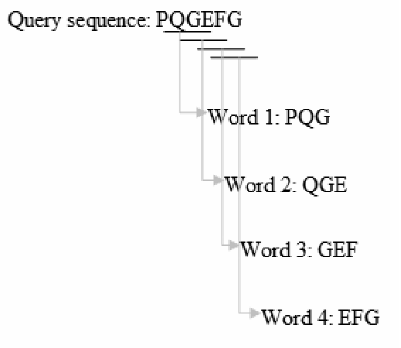


Figure 4.6: The method to establish the k-letter query word list

3. **List the possible matching words.**

   This step is one of the main differences between BLAST and FASTA. FASTA cares about all of the common words in the database and query sequences that are listed in step 2; however, BLAST cares about only the high-scoring words. The scores are created by comparing the word in the list in step 2 with all the 3-letter words. By using the scoring matrix (substitution matrix) to score the comparison of each residue pair, there are 20^3 possible match scores for a 3-letter word.

4. **Organize the remaining high-scoring words into an efficient search tree.** This is for the purpose that the program can rapidly compare the high-scoring words to the database sequences.
5. **Repeat step 1 to 4 for each k-letter word in the query sequence.**
6. **Scan the database sequences for exact match with the remaining high-scoring words.**
   The BLAST program scans the database sequences for the remaining high-scoring word, such as PEG, of each position. If an exact match is found, this match is used to seed a possible ungapped alignment between the query and database sequences.
7. **Extend the exact matches to high-scoring segment pair (HSP).**

   - The original version of BLAST stretches a longer alignment between the query and the database sequence in left and right direction, from the position where exact match is scanned. The extension does not stop until the accumulated total score of the HSP begins to decrease. A simplified example is presented in **Figure 4.7.**

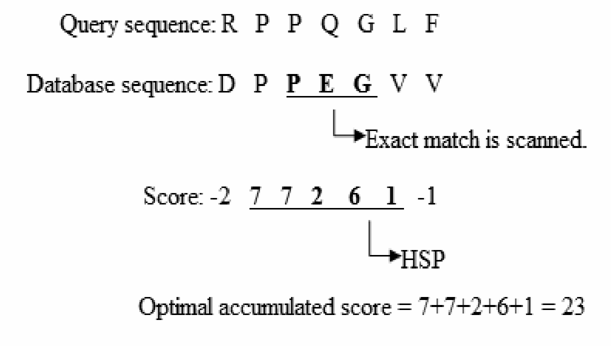

Query sequence: R P P Q G L F
Database sequence: D P P E G V V
Exact match is scanned.
Score: -2 7 7 2 6 1 -1
HSP
Optimal accumulated score = 7+7+2+6+1 = 23

Figure 4.7: The process to extension the exact match.

- BLAST2 adopts a lower neighborhood word score threshold to maintain the same level of sensitivity for detecting sequence similarity. Therefore, the possible matching words list in step 3 becomes longer. Next, the exact matched region, within distance A from each other on the same diagonal in **Figure 4.8**, has been joined as a longer new region. Finally, the new regions are then extended as the same method in the original version of BLAST, and the HSPs' (High-scoring segment pair) scores of the extended regions are then created by using a substitution matrix as before.

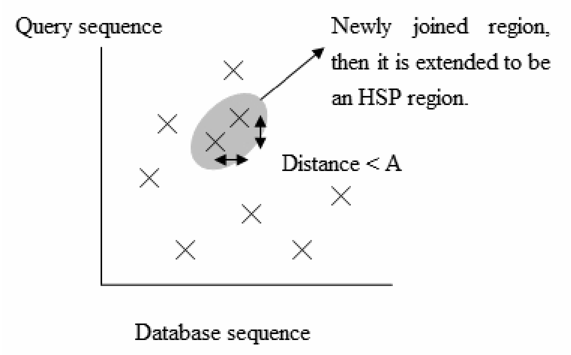

Query sequence
Newly joined region, then it is extended to be an HSP region.
Distance < A
Database sequence

Figure 4.8: The positions of the exact matches.

8. **List all of the HSPs in the database whose score is high enough to be considered.**
   The HSPs have been listed whose scores are greater than the empirically determined cutoff score S. By examining the distribution of the alignment scores modeled by comparing random sequences, a cutoff score S can be determined such that its value is large enough to guarantee the significance of the remained HSPs.
9. **Evaluate the significance of the HSP score.**
   BLAST next assesses the statistical significance of each HSP score by exploiting the Gumbel extreme value distribution (EVD). In accordance with the Gumbel EVD, the probability p of observing a score S equal to or greater than x is given by the equation

$$p(S \geq x) = 1 - \exp(-e^{-\lambda(x-\mu)})$$

Where $\mu = \frac{[\log Km'n']}{\lambda}$

The statistical parameters λ and K are estimated by fitting the distribution of the ungapped local alignment scores, of the query sequence and a lot of shuffled versions (Global or local shuffling) of a database sequence, to the Gumbel extreme value distribution. λ and K depend upon the substitution matrix, gap penalties, and sequence composition (the letter frequencies).The m' and n' is the effective length of the query and database sequence, respectively. The original sequence length is shortened to the effective length to compensate for the edge effect. They can be calculated as:

$$m' \approx m - \frac{(\ln Kmn)}{H}$$

$$n' \approx n - \frac{(\ln Kmn)}{H}$$

Where H is the average expected score per aligned pair of residues in an alignment of two random sequences. Altschul and Gish gave the typical values, λ = 0.318, K = 0.13, and H = 0.40, for ungapped local alignment. The expect score E of a database match is the number of times that an unrelated database sequence would obtain a score S higher than x by chance. The expectation E obtained in a search for a database of D sequences is given by

$$E \approx 1 - e^{-p(x>s)D}$$

Furthermore, when $p < 0.1$, E could be approximated by the Poisson distribution as: $E \approx pD$

10. Make two or more HSP regions into a longer alignment.

Sometimes, it has been found that two or more HSP regions in one database sequence that can be made into a longer alignment. This provides additional evidence of the relation between the query and database sequence. There are two methods, the Poisson method and the sum-of scores method, to compare the significance of the newly combined HSP regions. Suppose that there are two combined HSP regions with the sets of score (65, 40) and (52, 45), respectively. The Poisson method gives more significance to the set with the lower score of each set is higher (45>40). However, the sum-of-scores method prefers the first set, because 65+40 (105) is greater than 52+45(97). The original BLAST uses the Poisson method; gapped BLAST and the WU-BLAST use the sum-of scores method.

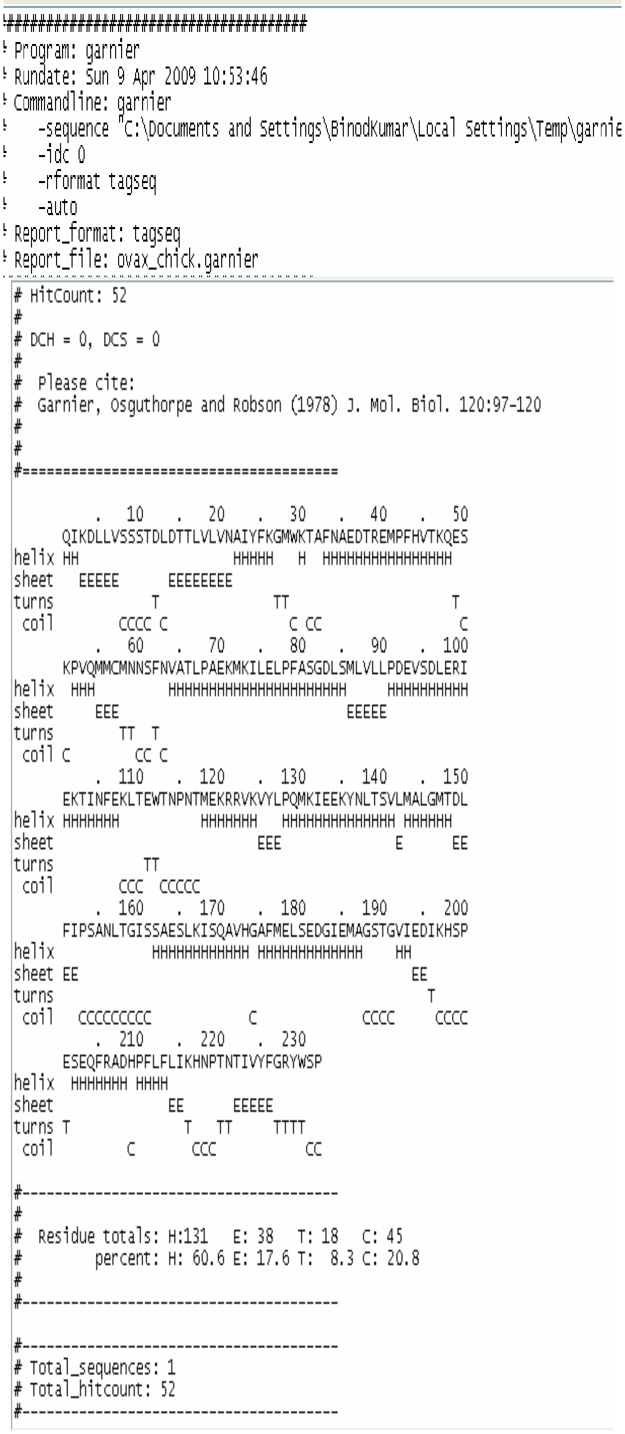
11 **Show the gapped Smith-Waterman local alignments of the query and each of the matched database sequences.**

- The original BLAST only generates ungapped alignments including the initially found HSPs individually, even when there is more than one HSP found in one database sequence.
- BLAST2 versions produce a single alignment with gaps that can include all of the initially found HSP regions. Note that the computation of the score and its corresponding E score is involved with the adequate gap penalties.

12 **Report matches whose expect score is lower than a threshold parameter E.**

4.2.1 BLAST: Output

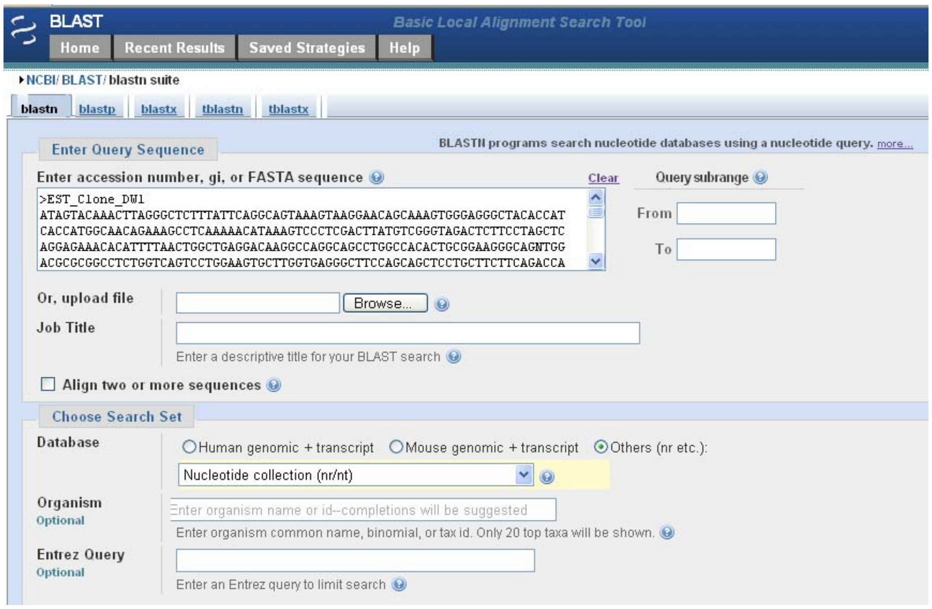
The protein sequence has been used for BLAST output [4]. First there is a short description of the program options chosen. Then there is a list of all of database sequences that match our query sequence. Several numbers are assigned to each of these sequences that represent the quality of the match.



```
########################################
# Program: garnier
# Rundate: Sun 9 Apr 2009 10:53:46
# Commandline: garnier
#    -sequence "C:\Documents and Settings\BinodKumar\Local Settings\Temp\garnie
#    -idc 0
#    -rformat tagseq
#    -auto
# Report_format: tagseq
# Report_file: ovax_chick.garnier
```

```
# HitCount: 52
#
# DCH = 0, DCS = 0
#
#  Please cite:
#  Garnier, Osguthorpe and Robson (1978) J. Mol. Biol. 120:97-120
#
#
#=======================================

          .   10    .   20    .   30    .   40    .   50
      QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQES
helix HH                   HHHHH   H  HHHHHHHHHHHHHHHH
sheet   EEEEE      EEEEEEEE
turns            T              TT                    T
 coil        CCCC C               C CC                 C
          .   60    .   70    .   80    .   90    .  100
      KPVQMMCMNNSFNVATLPAEKMKILELPFASGDLSMLVLLPDEVSDLERI
helix  HHH         HHHHHHHHHHHHHHHHHHHHHH     HHHHHHHHHH
sheet     EEE                            EEEEE
turns        TT  T
 coil C        CC C
          .  110    .  120    .  130    .  140    .  150
      EKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTSVLMALGMTDL
helix HHHHHHH          HHHHHHH   HHHHHHHHHHHHHH HHHHHH
sheet                        EEE              E      EE
turns           TT
 coil        CCC  CCCCC
          .  160    .  170    .  180    .  190    .  200
      FIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSP
helix            HHHHHHHHHHHH HHHHHHHHHHHHH    HH
sheet EE                                         EE
turns                                              T
 coil   CCCCCCCCC            C             CCCC     CCCC
          .  210    .  220    .  230
      ESEQFRADHPFLFLIKHNPTNTIVYFGRYWSP
helix  HHHHHHH HHHH
sheet              EE      EEEEE
turns T              T   TT     TTTT
 coil         C       CCC           CC

#---------------------------------------
#
#  Residue totals: H:131   E: 38   T: 18   C: 45
#         percent: H: 60.6 E: 17.6 T:  8.3 C: 20.8
#
#---------------------------------------

#---------------------------------------
# Total_sequences: 1
# Total_hitcount: 52
#---------------------------------------
```

Figure 4.9: Output from BLAST Query



Enter accession number, gi, or FASTA sequence

```
>EST_Clone_DW1
ATAGTACAAACTTAGGGCTCTTTATTCAGGCAGTAAAGTAAGGAACAGCAAAGTGGGAGGGCTACACCAT
CACCATGGCAACAGAAAGCCTCAAAAACATAAAGTCCCTCGACTTATGTCGGGTAGACTCTTCCTAGCTC
AGGAGAAACACATTTTAACTGGCTGAGGACAAGGCCAGGCAGCCTGGCCACACTGCGGAAGGGCAGNTGG
ACGCGCGGCCTCTGGTCAGTCCTGGAAGTGCTTGGTGAGGGCTTCCAGCAGCTCCTGCTTCTTCAGACCA
```

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Database: ○Human genomic + transcript ○Mouse genomic + transcript ●Others (nr etc.):

Nucleotide collection (nr/nt)

Organism Optional: Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query Optional

Enter an Entrez query to limit search

Figure 4.10: Query sequence section of Nucleotide Blast Form.

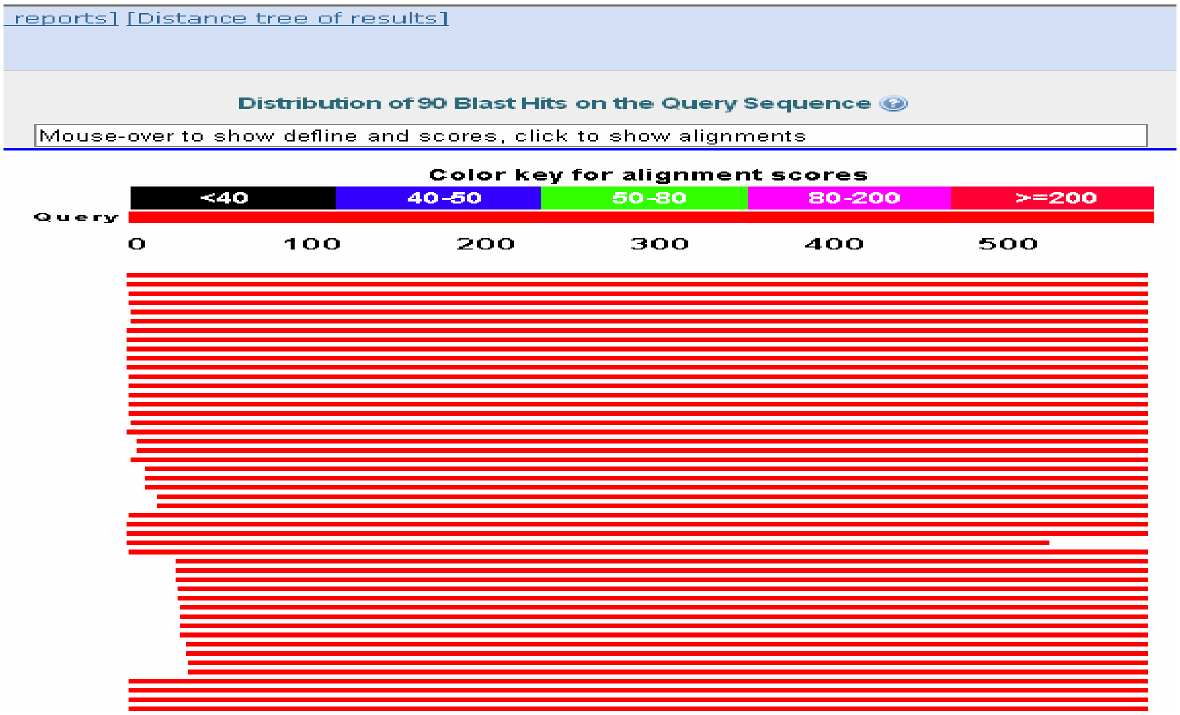

reports] [Distance tree of results]
Distribution of 90 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
100
200
300
400
500

Figure 4.11: Graphic Summary of Nucleotide Blast Form

4.2.2 BLAST Services

The BLAST web server, hosted by the NCBI, allows with a web browser to perform similarity searches against constantly updated databases of proteins and DNA that include most of the newly sequenced organisms.

BLAST is actually a family of programs (all included in the blastall executable). These include:

1. **Nucleotide-nucleotide BLAST (blastn)**
   This program, given a DNA query, returns the most similar DNA sequences from the DNA database that the user specifies.
2. **Protein-protein BLAST (blastp)**
   This program, given a protein query, returns the most similar protein sequences from the protein database that the user specifies.
3. **Position-Specific Iterative BLAST (PSI-BLAST)**
   This program is used to find distant relatives of a protein. First, a list of all closely related proteins is created. These proteins are combined into a general "profile" sequence, which summarizes significant features present in these sequences. A query against the protein database is then run using this profile, and a larger group of proteins is found. This larger group is used to construct another profile, and the process is repeated By including related proteins in the search, PSI-BLAST is much more sensitive in picking up distant evolutionary relationships than a standard protein-protein BLAST.
4. **Nucleotide 6-frame translation-protein (blastx)**
   This program compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
5. **Nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx)**
   This program is the slowest of the BLAST family. It translates the query nucleotide sequence in all six possible frames and compares it against the six-frame translations of a nucleotide sequence database. The purpose of tblastx is to find very distant relationships between nucleotide sequences.

6. **Protein-nucleotide 6-frame translation (tblastn)**

   This program compares a protein query against the all six reading frames of a nucleotide sequence database.

7. **Large numbers of query sequences (megablast)**

   When comparing large numbers of input sequences via the command-line BLAST, "megablast" is much faster than running BLAST multiple times. It concatenates many input sequences together to form a large sequence before searching the BLAST database, and then post-analyze the search results to glean individual alignments and statistical values.

Table 4.1: BLAST Program Options

Program	**Query Sequence**	**Database**	**Alignment Type**
Blastp	Protein	Protein	Gapped
Blastn	Nucleic acid	Nucleic acid	Gapped
Blastx	Translated nucleic Acid	Protein	Each frame gapped
Tblastn	Protein	Translated nucleic Acid	Each frame gapped
Tblastx	Translated nucleic Acid2	Translated nucleic Acid	Ungapped

4.2.3 FILTERING and GAPPED BLAST

Filtering is the process of removing the undesired sequences from the query sequence prior to the search. BLAST filters regions of low-complexity. If my sequence contains large regions of “low complexity” it may not significant hits to the database.

GAPPED- BLAST

Gapped –BLAST is BLAST 2.0[6]. It represents BLAST plus a new heuristic for gapped alignments. It allows the introduction of gaps (deletions and insertions) into alignments. With a gapped alignment tool, homologous domains do not have to be broken into several segments. The programs, blastn and blastp offer fully gapped alignments. blastx and tblastn have ‘in-frame’ gapped alignments and use sum statistics to link alignments from different frames.

Output of BLAST 2.0 is as follows:

- Information on Query sequence and Databases used
- Histogram, like the FASTA histogram
- Scores in bits-E value: number of hits expected to be reported by chance
- Alignments found(default =50)
- Parameters used in BLAST search.

4.2.4 FASTA and BLAST Algorithms Comparison

Table 4.2: Comparison of BLAST and FASTA

1.FASTA offers many of the same functionalities as BLAST. Although BLAST tools are faster, FASTA provides more accurate sequence alignments. BLAST uses a different algorithm, but its results are similar to those found by FASTA. BLAST uses a general set of rules to compare specific regions of similarity for a given search string. BLAST is more than a precision matching tool. It provides a method for comparing the structures and functions of samples to genetic sequences and proteins
2.FASTA is superior to BLAST for translated DNA–protein comparison and DNA database searches because it calculates a single alignment that allows frame shifts. In contrast, BLAST performs forward-frame searches separately. By treating forward-reading frames as a single sequence, FASTA makes it much easier to produce high-quality alignments that extend the length of the protein sequence, resulting in improved sensitivity.
3.FASTA is a little more flexible than BLAST for DNA sequence searches. It provides small word sizes to accommodate polymerase chain reaction primers having short sequences. And FASTA uses several different scoring matrixes to help identify sequences of varying lengths.

4.3 References

[1] httep://www.ebi.ac.uk/fasta33/genomics.html

[2] http://www.ebi.ac.uk/snpfasta3/index.html

[3] http://www.ncbi.nlm.nih.gov/BLAST

[4] www.ch.ebnet.org/software/bBLAST.html

[5] http://www.ncbi.nlm.gov/BLAST/

[6] S.C. Rastogi, N.Mendiratta, P. Rastogi, "Bioinformatics Methods and Applications", PHI, 144-145, 2006.

Chapter -5

Protein Structure and ChemiformatiCs

Chapter 5

PROTEIN STRUCTURE AND CHEMINFORMATICS

5.1 Introduction

The subunits of a protein are amino acids or to be precise **amino acid residues**. An amino acid consists of a central carbon atom (the alpha Carbon C_{alpha}) and an amino group (NH_2), a hydrogen atom (H), a carboxy group (COOH) and a side chain (R) which is bound to the C_{alpha}. Different **side chains** (R_i) make up different amino acids with different physico-chemical properties. Proteins are made out of 20 amino acids (there is a list with corresponding three- and one-lettercodes in the section on Biological Preliminaries). A **peptide bond** is formed via covalent binding of the Carbon atom of the Carboxy group of one amino acid to the nitrogen atom of the amino group of another amino acid by dehydration:

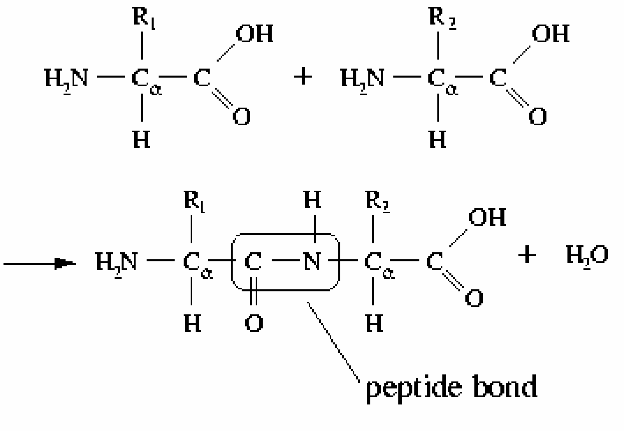


Figure 5.1: Peptide bond linking two amino acids

A **polypeptide chain** is a chain of amino acid residues linked together by peptide bonds. The **backbone** of the polypeptide is given by the repeated sequence of three atoms of each residue in the chain: the amide N, the alpha Carbon C_{alpha} and the Carbonyl C. Rotations in the chain take place about the bonds in the backbone, where as the peptide bond usually is rigid **(Figure 2)**. The existence of an amino group (**N-Terminal**) at one end of the chain and a carboxy group (**C-Terminal**) at the other end designs a direction to the chain. Conventionally the beginning of a polypeptide is its N-Terminal.

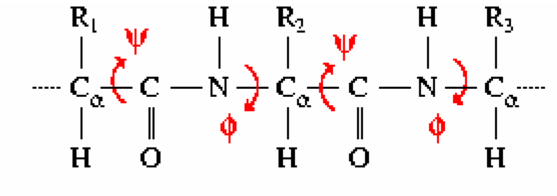


Figure 5.2: Torsion (or dihedral) angles of the backbone

5.2 Different Levels of Protein Structure

The wide variety of 3-dimensional protein structures corresponds to the diversity of functions proteins fulfill.

Proteins fold in three dimensions. Protein structure is organized hierarchically from so-called *primary structure* to *quaternary structure*. Higher-level structures are *motifs* and *domains*.

Above all the wide variety of conformations is due to the huge amount of different sequences of amino acid residues. The **primary structure** is the sequence of residues in the polypeptide chain.

Secondary structure is a local regularly occurring structure in proteins and is mainly formed through hydrogen bonds between backbone atoms. So-called random coils, loops or turns don't have a stable secondary structure. There are two types of stable secondary structures: **Alpha helices and beta-sheets** (**Figure 3 and Figure 4**). Alpha-helices and beta-sheets are preferably located at the core of the protein, where as loops prefer to reside in outer regions.

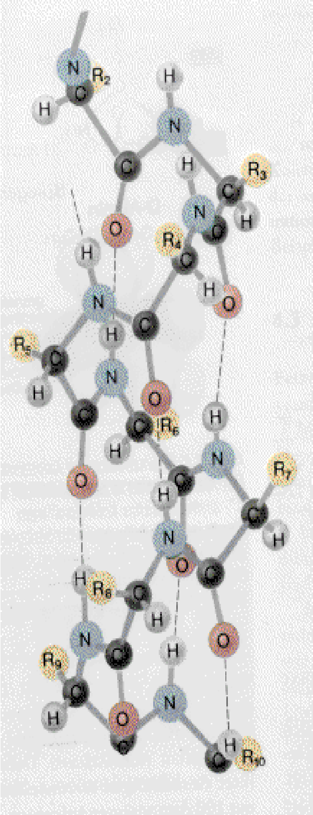


Figure 5.3: An alpha helix: The backbone is formed as a helix. An ideal alpha helix consists of 3.6 residues per complete turn. There are hydrogen bonds between the carboxy group of amino acid n and the amino group of another amino acid n+4[1][2]. The mean phi angle is -62 degrees and the mean psi angle is -41 degrees [3].

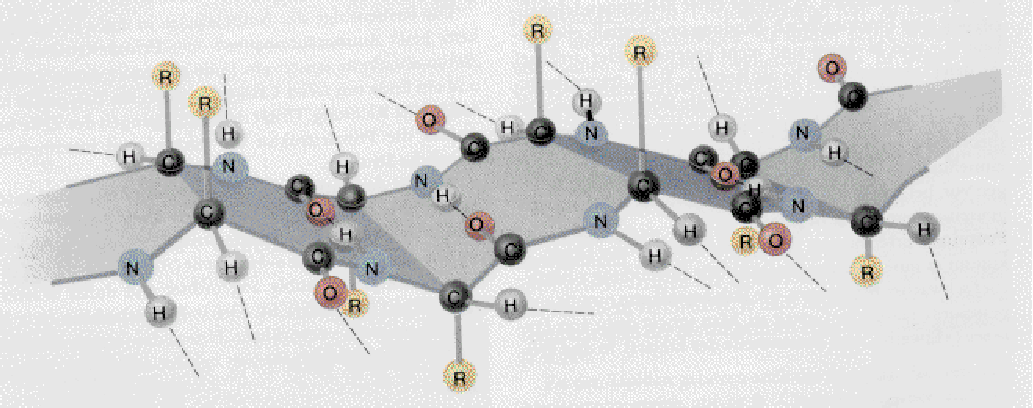


Figure 5.4: An antiparallel beta sheet. Beta sheets are created, when atoms of beta strands are hydrogen bond. Beta sheets may consist of parallel strands, antiparallel strands or out of a mixture of parallel and antiparallel strands [4].

Tertiary structure describes the packing of alpha-helices, beta-sheets and random coils with respect to each other on the level of one whole polypeptide chain. **Figure 5** shows the tertiary structure of Chain B of Protein Kinase C Interacting Protein.

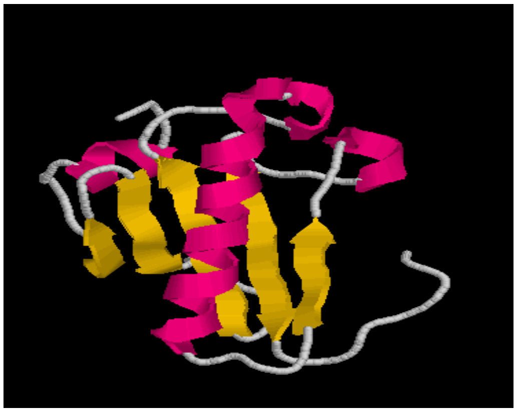


Figure 5.5: Secondary structure of Protein. Helices are visualized as ribbons and extended strands of betasheets by broad arrows

Quaternary structure only exists, if there is more than one polypeptide chain present in a complex protein. Then quaternary structure describes the spatial organization of the chains. **Figure 5.6** shows both, Chain A and Chain B of Protein Kinase C Interacting Protein forming the quaternary structure.

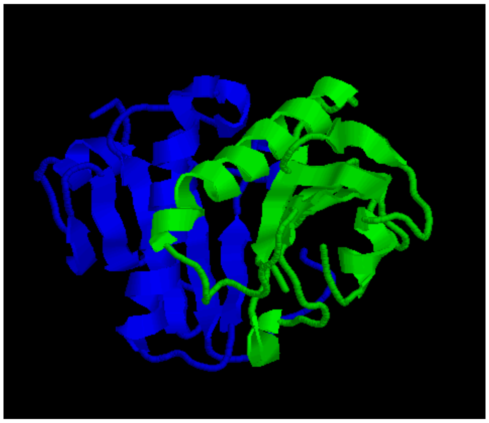


Figure 5.6: Quaternary structure of Protein.

5.3 Prediction Methods

Protein's 3D structure helps us to understand its functionality and provides means for planning experiments and drug design. Experimental methods given by X-ray crystallography and NMR spectroscopy to determine protein structure. The Brookhaven Protein Data Bank (PDB) is the repository for those structures. Files including atom coordinates which are suited for visualization by graphical molecule viewers like rasmol can be obtained at this site. PDB is also searchable with a sequence as a query, e.g. with the BLAST service located at NCBI with a polypeptide as a query.

The various prediction methods are based on the assumption, that the three-dimensional protein structure is determined by its primary structure.

Structure prediction methods are divided two categories:

1. Ab Initio Methods
2. Heuristic Methods

1. Ab Initio Methods

Ab Initio methods of determining protein structure are based on sequence data and molecular dynamics. One assumption is that a protein’s secondary structure can be completely defined as a function of bond lengths, bond angles and torsion angles. The overall process of predicting tertiary protein structure from known sequence is illustrated in **Figure 5.7**. Given a sequence of amino acids, the first step is to generate a secondary structure by using bond lengths, angles and torsion angles. The next phase of process, generating the tertiary structure, involves methods such as molecular dynamics to create a library of tertiary protein structure.

Molecular dynamics calculates the force on each atom and move that atom a distance in small unit of time. The process is repeated until a pre-determined time limit is reached.

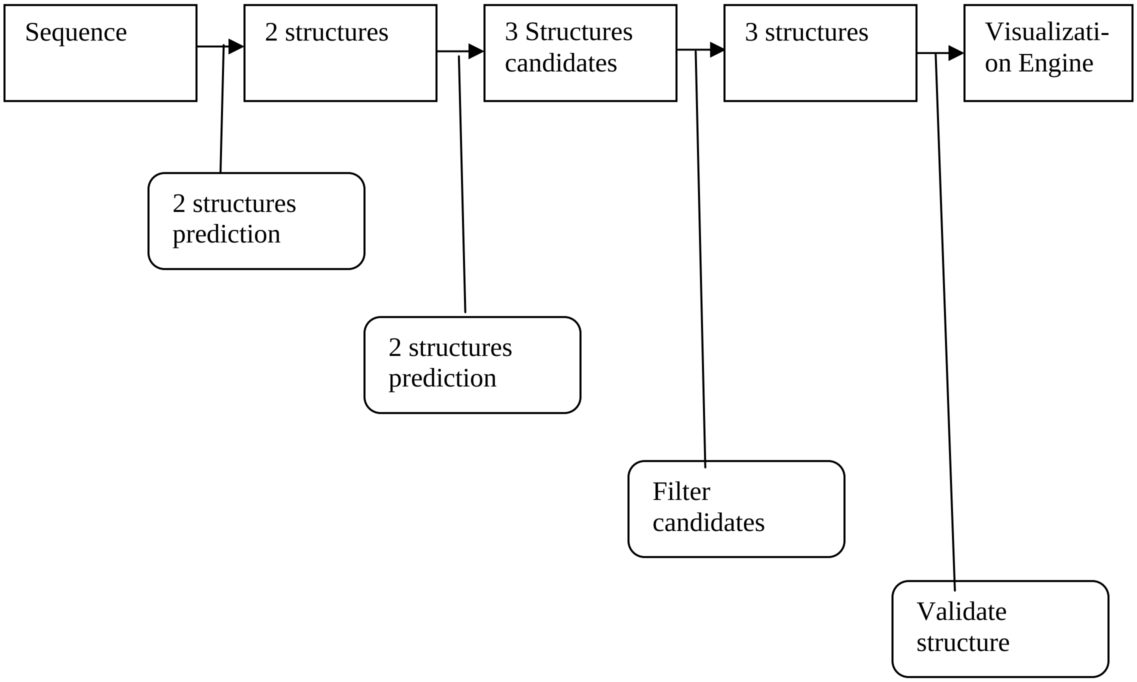


Figure 5.7: General Ab Initio Protein Structure Process

From a library of 3D structure candidates, the most promising structures are filtered from less capable structures. A common method of filtering to identify the most stable molecular conformations is based on the assumption that the native conformation of a protein is the conformation with the lowest energy. Once the top protein structure candidate is identified, it is validated and visualized. In validation process there is comparison of the predicted protein structure with a structure derived from NMR experiments. Validation process involves assigning a figure of merit to the predicted structure, based on comparison to the gold standard. The most often-used figure of merit in protein structure comparison is the root mean squared deviation (RMSD). The calculation for RMSD, expressed in Angstroms (**Figure 5.8**).

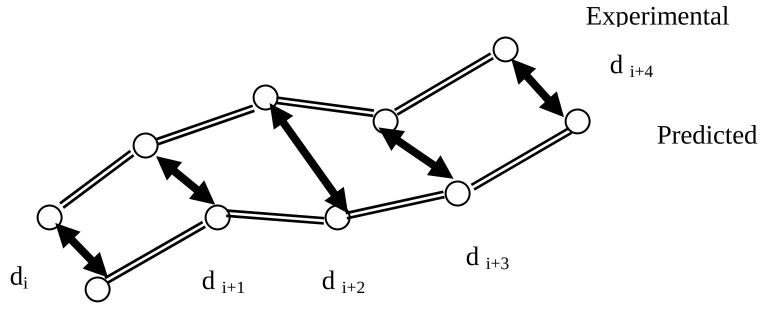


Figure 5.8: RMSD Calculation.

Where,

$$RMSD = \sqrt{\frac{\sum_i d_i^2}{N}}$$

N = Number of atoms

d = the distance in Angstroms between corresponding atoms in experimental and predicted protein structures.

Identical structures with perfect match would have an RMSD of 0; matching short to moderate –length protein structures have RMSD in the 1-3 Angstrom range. An RMSD of 5 or 6 Angstroms may be intolerable in a molecule with only 50 residues, but perfectly acceptable in large protein molecules for applications such as searching structure databases for known protein structures. However, even a relative measure, RMSD is valuable when working within a single family of proteins because the size of structures will be about same.

Visualization of the protein structure is performed through protein engines on the web, such as Rasmol or SWISS-PDBViewer.

2. Heuristic Methods

Heuristic methods use a database of protein structures to make prediction about the structure of newly sequenced proteins. In heuristic methods most newly sequenced proteins share structural similarities with proteins whose structures and sequences are known, and that these structures can serve as templates for new sequences. It is also assumed that relatively substantial changes in amino acid sequence may not alter the protein structure.

The main heuristic method of predicting protein structure from amino acid sequence data is comparative modeling i.e. to find similarities in amino acid sequence. Comparative modeling assumes that protein with similar amino acid sequences share the same basic 3D structure.

The basis for comparative modeling is typically the PDB, which contains description of 3D structures of proteins and other molecules as determined by NMR and X-ray crystallography experiments. Protein structures defined within Protein Database Modeling (PDM) and virtually every other protein structure database are based on assumptions that may not be completely valid. For example, the common assumption that amino acid sequences result in similar protein structures is known to have exceptions.

Comparative modeling is an iterative, multi-phase process. In **Figure 5.9**, given protein sequence data, the main phases of process are template selection, alignment, model building and evaluation. 3D visualization is often performed as part of the evaluation phase. The key activities in each phase of the comparative modeling process are outlined here.

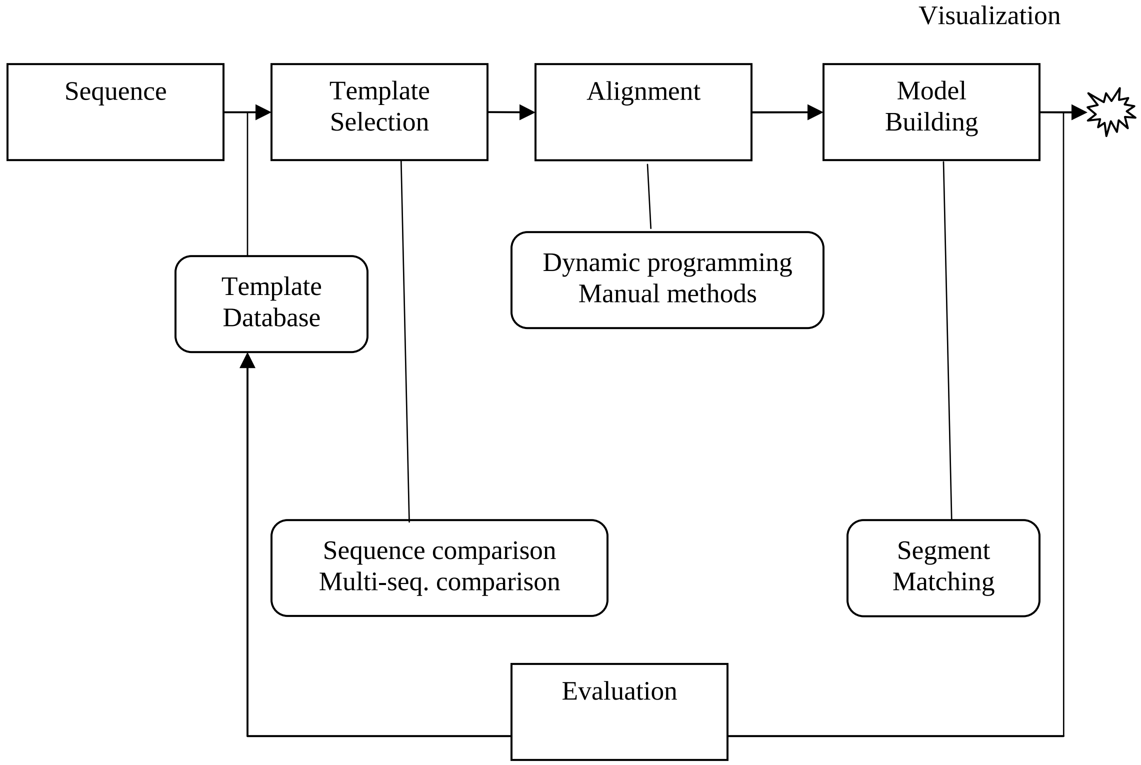


Figure 5.9: Comparative Modeling Process

Template Selection

Template selection involves searching a template database for the closest match or matches to the new (target) molecules, based on the target's amino acid sequence. The goal of template selection is to discover a link between the target protein and a known protein structure. For this usually PDB are used. Selecting an appropriate group of database entries from the database to serve as structure templates is based on sequences comparisons or threading.

Pairwise sequence comparison involves searching selections of the template candidate for amino acid sequences that are similar to sequences in the target protein. Multiple sequence comparison relies on an iterative algorithm that expands the template search to include all candidate templates from the template database.

Threading involves aligning the sequence of the target protein with the 3D structure of a template to determine whether the amino acid sequences is spatially and chemically similar to the template.

Alignment

The main aim of the alignment phase of comparative modeling is to align the sequence of polypeptides in the target sequence with that of the template structure in order to position the target and template in the same 3D orientation. Many of the alignment procedures are based on dynamic programming techniques.

Model Building

Actual model building begins only after identification of the libraries of templates that match the target protein. The structure of one of the template exactly fits the target protein, signifying that the structure of that target is identified to that of the template.

Rigid body assembly approach uses large segments of the template that are dissected at natural folds and reassembled over the superimposed structure of the target molecule. The accuracy of model building through rigid assembly is increased because there is an increased chance of availability of sub-assembly of molecule that matches the sequence in a corresponding area in the target protein as shown in **Figure 5.10**.

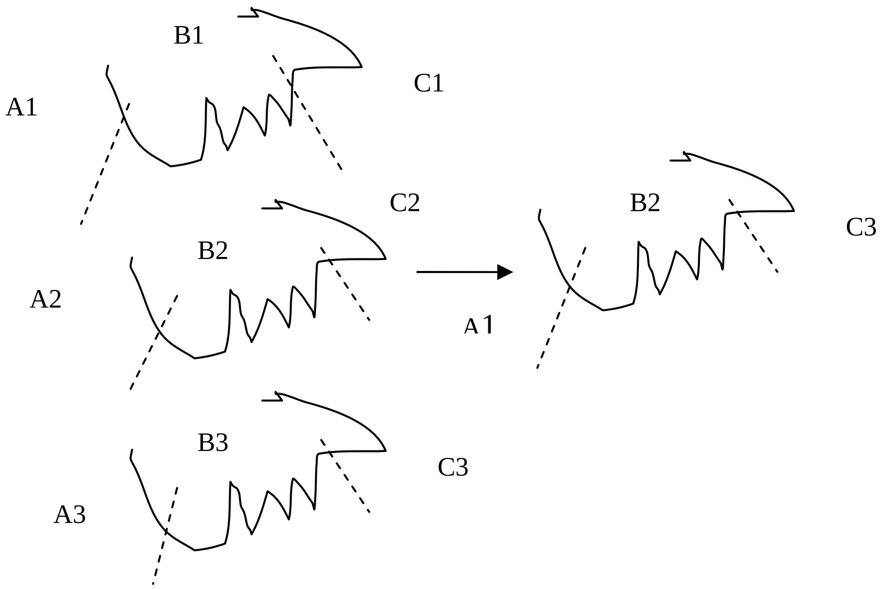


Figure 5.10: Rigid Body Assembly of Protein Structure

The aim of segment matching is to identify areas on structure templates that match areas in the target protein with similar sequences. These short matching segments in template are used as guiding positions in the target molecule, as shown in **Figure 5.11**.

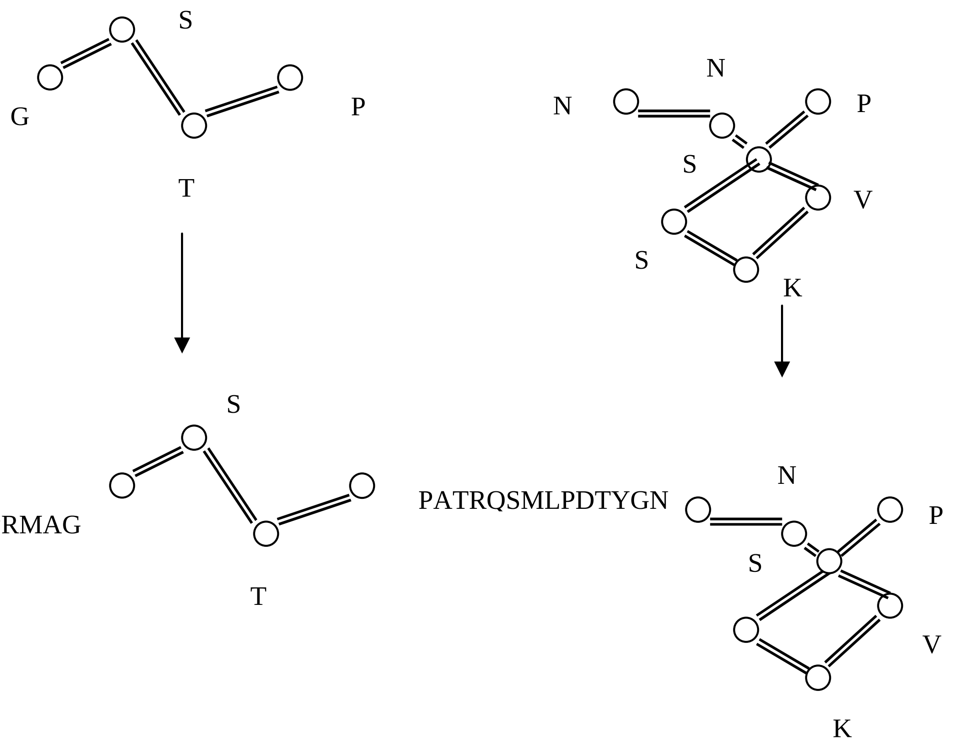


Figure 5.11: Short Segment Assembly of Protein Structure

Evaluation

In evaluating comparative modeling, even best methods like ab initio methods rarely achieve accuracies approaches 70 percent. The modeling process is repeated dozen of times before a reasonable target structured is constructed. In this evaluation process visualization tool is used to validate gross measures.

For quantitative evaluation, a measure of target-template similarity can be used. The greater the similarity of the model with the closest template, as measured by RMSD, the more likely the model is an accurate prediction of the actual structure.

5.4 Secondary Structure Prediction

Linus Pauling [5] suggested that amino acid chains could assume regular local structures, namely alpha helices and beta strands. In between these secondary structure elements there are turns or loops. There is a long tradition of attempts to predict local secondary structure based on sequence. State-of-the-art secondary structure prediction generally observes the frequencies of occurrences of k-tuples in particular secondary structures. Based on this statistic prediction can be made for a new sequence.

Chou and Fasman [6] apply a basic log-odds approach for the occurences of single amino acid residues in the sequence, while the GOR method [7] which is based on information theory uses all possible pair frequencies within a sliding window.

As long as one restricts to the problem to the prediction for a single sequence there seems to be an inherent limit in prediction accuracy of around 65%. Multiply aligned sequences offer a means to surpass this limit. The PHD-method [8] uses evolutionary information from multiple sequence alignments in a multi level system of neural networks. So, the average accuracy of PHD-method is greater than 72%.

The GOR method is used to estimate for the prediction of secondary structure of protein.

5.5 The Protein Folding Problem

It has long been known that the structure of a protein is determined purely by the amino acid sequence [9], and the structure of the protein determines the function. The function of a protein depends entirely on the ability of the protein to fold rapidly and reliably to its native structure.

This folding process satisfy two conditions - one thermodynamic, and one kinetic. The thermodynamic consideration is that the protein adopts a single, stable, folded conformation. The kinetic requirement is that the protein must fold to the native state on an appropriate timescale. It has been suggested that for a protein of 100 amino acids, a purely random conformational search would require around 10- 36 s or around 10- 29 years,[10] and yet proteins are able to fold on a timescale of milliseconds to seconds. This suggests that only a small amount of conformational space is sampled during the folding process and this in turn implies the existence of kinetic folding pathways, [11]. This paradox of how proteins fold rapidly and reliably to their native conformation is known as the protein folding problem.

The computational difficulty of protein folding is classified as an NP-complete problem. If a problem is NP-complete, it means that a particular solution can be checked in a polynomial time but to solve the whole problem requires an exponential time algorithm. A problem is in NP if it has a nondeterministic polynomial time solution. This means that the solution can be checked within polynomial time. As the exponential function in an NP-complete problem increases as much more rapid rate than a polynomial, these problems are untraceable.

5.6 Cheminformatics

5.6.1 Introduction

Cheminformatics (also known as **chemoinformatics** and **chemical informatics**) is the use of computer and informational techniques, applied to a range of problems in the field of chemistry. These in silico techniques are used in pharmaceutical companies in the process of drug discovery. These methods can also be used in chemical and allied industries in various other forms.

Chemoinformatics [12] is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization. Cheminformatics combines the scientific working fields of chemistry and computer science for example in the area of chemical graph theory and mining the chemical space. It is to be expected that the chemical space contains at least 10^{60} molecules. Cheminformatics can also be applied to data analysis for various industries like paper and pulp, dyes and such allied industries. Enzymes are a subset of receptor-like proteins that are directly responsible for catalyzing the biochemical reactions. DNA polymerase and related enzymes are crucial for cell division and replication. Enzymes are genetically programmed to be absolutely specific for their appropriate molecular targets.

The most important concept in drug design is to understand the methods by which the active site of a receptor selectively restricts the binding of inappropriate structures. Any potential molecule that can bind to a receptor is called a ligand. In order for a ligand to bind, it must contain a specific combination of atoms that presents the correct size, shape, and charge composition in order to bind and interact with the receptor **Figure 5.12** schematically shows a typical ligand-receptor binding interaction.

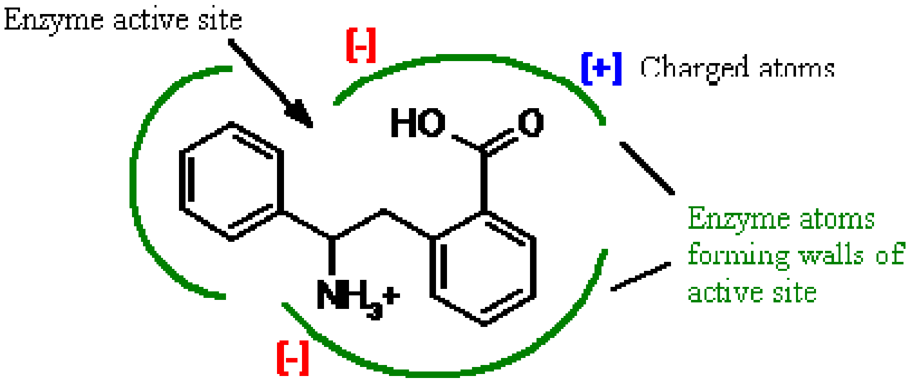


Figure 5.12: Enzyme substrate complementary interactions.

Ligand-receptor interaction must have complementary size and shape. This is termed steric complementarity. As is the case with an actual key, if a different molecule varies by even a single atom in the wrong place, it may not fit properly, and will most likely not interact with the receptor. However, the more closely the fit between the ligand and receptor, the more tightly the interaction becomes.

The main driving force for ligand and receptor binding is hydrophobic interaction. In order for ligand and receptor to interact, there must be a driving force that compels the ligand to leave the water and bind to the receptor. The hydrophobicity of a ligand is what causes this. Hydrophobicity stands for 'water fearing' and is a measure of how 'greasy' a compound is. It can be roughly approximated by the percentage of hydrogen and carbon in the molecule. As shown in **Figure 5.13**, the active site may contain a mixture of hydrophobic pockets and regions that are more polar.

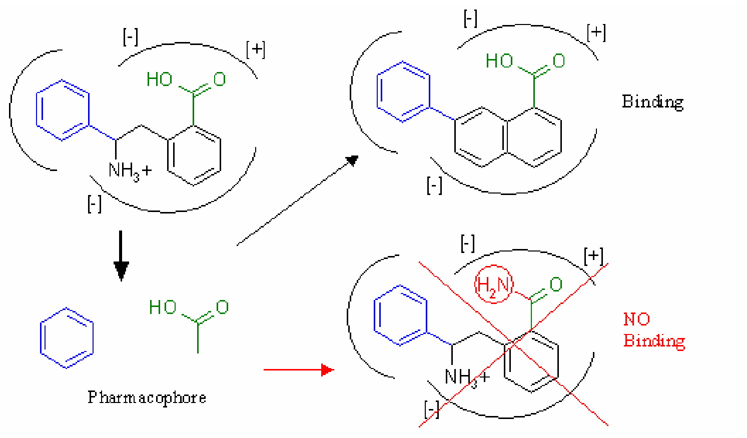


Figure 5.13: Pharmacophore and Receptor Binding

There are numerous potential interactions between ligand and receptor. Depending upon the size of the active site, there may be a numerous steric, electrostatic, and hydrophobic contact. The specific interactions that are crucial for ligand recognition and binding by the receptor are termed the pharmacophore. Usually, these are the interactions that directly factor into the structural integrity of a receptor or are involved in the mechanism of its action.

This is shown in **Figure 5.13** above. In the upper left frame of this figure, there is native ligand bound within the active site. Through biochemical investigation, phenyl ring (blue) and the carboxylic acid group (green) are vital to receptor interaction has been determined. Thus, it has been interpreted that these two groups must be the pharmacophore that a ligand must present to the receptor for binding.

5.6.2 The Challenge of Drug Design

There are difficulties in designing drugs towards specific target receptors.

Table 5.1. Major Tasks and concerns in Drug development [13]**.**

1. Characterize medical condition and determine receptor targets.
2. Achieve active site complementarity: steric, electrostatic, and hydrophobic.
3. Consider biochemical mechanism of receptor.
4. Adhere to laws of chemistry.
5. Synthetic feasibility.
6. Biological considerations.
7. Patent considerations.

When a medical condition exists where a drug could be beneficial, extensive scientific study must first be done in order to determine the biological and biochemical problems that underlie the disease process.

Once a receptor target has been established and well characterized, the process of ligand design begins. The first consideration is that the designed ligand must complement the active site of the receptor target. Steric, electrostatic, and hydrophobic complementarity must be established. The pharmacophore must be presented to the receptor in order for recognition and binding to occur. Otherwise, the designed ligands have no chance of interacting with the receptor.

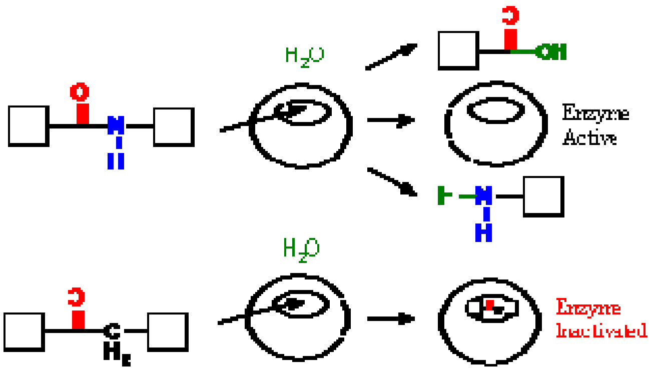


Figure 5.14: Designing ligands to offset enzyme mechanism.

In addition to adequately binding the receptor, the biochemical mechanism of the receptor target must be taken into consideration. This is shown in **Figure 5.14**. In this the biochemical mechanism of a protease has been schematically represented. A protease is an enzyme that cleaves proteins and peptides. In the top part of the figure, a specific group of atoms colored in red and blue, called a peptide bond has been recognized by protease. If the peptide bond is present at a specific position in the active site when the ligand binds, it is cleaved by the protease with the addition of water (H_2O) to form two separate fragments.

Having characterized the active site region and the mechanism of action of the target receptor, the challenge then becomes one of designing a suitable ligand. The optimal combination of atoms and functional groups to complement the receptor is often the natural ligand of the receptor.

There are biological considerations to the development of new drugs. The liver is the major organ of detoxification in the human body. Any drug that is taken undergoes a number of chemical reactions in the liver as the body attempts to neutralize foreign substances. Various chemical structures are highly toxic to biological systems, and these have been also well characterized.

5.6.3 The Drug Discovery Pipeline

The development [19] of any potential drug begins with years of scientific study to determine the biochemistry behind a medical problem for which pharmaceutical intervention is possible. The result is the determination of specific receptor targets that must be modulated to alter their activity in some way. Once these targets have been identified, the goal is then to find compounds that have to interact with the receptors.

The modern day drug discovery pipeline is outlined in **Figure 5.14**. The first step is to determine an estimate for the receptor. An estimation is a chemical or biological test that turns positive when a suitable binding agent interacts with the receptor. Usually, this test is some form of colorimetric estimation, in which an indicator turns a specific color when complementary ligands are present. This estimate is then used in mass screening, which is a technique whereby hundreds of thousands of compounds can be tested in a matter of days to weeks. Entire corporate database of known compounds has been first screened by pharmaceutical company. The reason is that if a successful match is found, the database compound is usually very well characterized. Then after, a synthetic method has been known for this compound. This enables the company to rapidly prototype a candidate ligand.

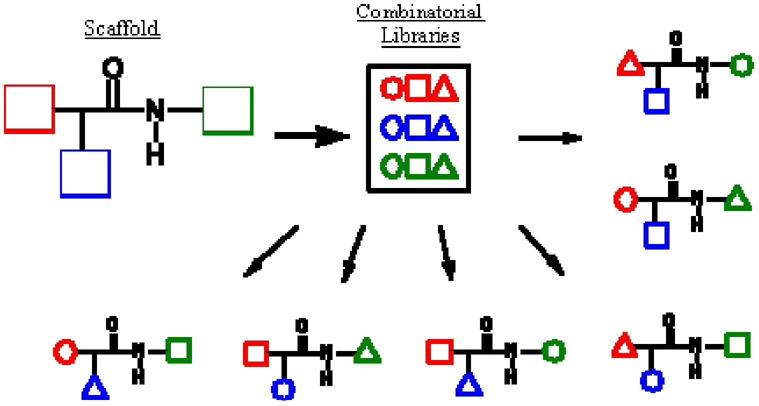


Figure 5.15: Combinatorial chemistry schematic.

Combinatorial chemistry is a very powerful technique that has been applied in the refinement of the lead compound. Combinatorial chemistry is a synthetic tool to rapidly generate thousands of lead compound derivatives for testing. As shown above in **Figure 5.15**, subsite groups (shown in red, green, and blue) are potential sites for derivatization. These subsites are then reacted with combinatorial libraries to generate a multitude of derivative structures, each with different substituent groups. By carefully selecting libraries based upon the study of the active site, the derivatization process towards optimizing ligand receptor interaction has been targeted.

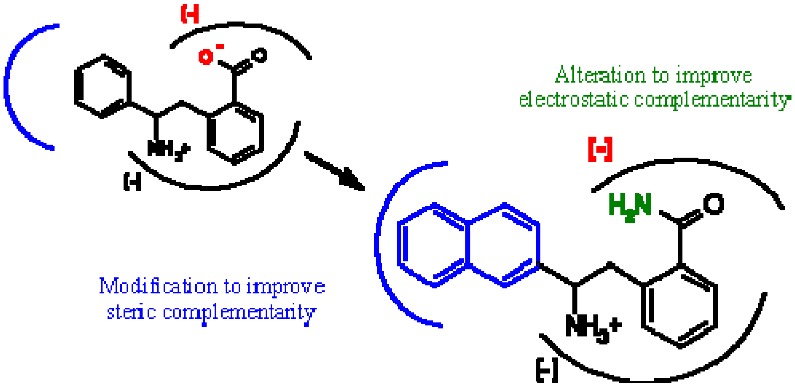


Figure 5.16: Structured Based Drug Design

Structure based design, often called rational drug design, and is much more focused than combinatorial chemistry. As shown above in **Figure 5.16**, it involves using the biochemical laws of ligand-receptor association discussed above to postulate ligand refinements to improve binding. Functional groups on the ligand can have been changed in order to expand electrostatic complementarity with the receptor. However, the danger in altering any portion of the ligand is the effect on the remaining ligand structures. Modifying even a single atom in the middle of the ligand can drastically change the shape of the overall structure. Even though complementarity in one portion of the ligand might be improved by the chemical revision, the overall binding might be severely compromised. This is the difficulty in any ligand refinement procedure.

5.6.4 Computer-Aided Drug Design (CADD)

Computer graphics technology has achieved the ability to generate vector models of chemical structures and manipulate them in real-time. The ability to study computer models of ligand structures and their binding interactions with a receptor has been offered by Computer graphics technology.

The time and effort required for drug synthesis and testing has been avoided by simply generating novel compounds using the computer. Testing has been replaced by calculating the ligand-receptor binding affinity using the physical laws of chemistry. The concept of generating **virtual lead compounds** entirely through computer simulation has been termed as Denovo Design.

Computer-Aided Drug Design (CADD) [14] is a specialized discipline that uses computational methods to simulate drug-receptor interactions. CADD methods are heavily dependent on bioinformatics tools, applications and databases.

5.6.5 Difficulties Implementing Denovo Design

Although computers have become exponentially faster, the complete number of calculations needed to accurately predict the binding of a denovo generated ligand to its receptor in a useful timeframe still requires significant approximations. In denovo design, a whole ligand from scratch has been generated and it has been docked within the receptor. A ligand is flexible structure, and can guess an excess of different conformations and orientations. The predicted binding structure has to be similar with the calculated one. Failure in this attempt has damaged the utility of denovo structure generating software.

The second most significant problem in computer aided denovo design [15] is the generation of undesired chemical structures. There are a nearly infinite number of potential combinations of atoms. However, the vast majority of these structures are of no use. As discussed above, undesired structures are rejected due to toxicity, chemical instability, or synthetic difficulty. Nearly all denovo design software packages are overwhelmed by this problem, especially with respect to synthetic feasibility.

5.6.7 Benefits of CADD

CADD methods and bioinformatics tools offer significant benefits for drug discovery programs.

- **Cost Savings**: Many biopharmaceutical companies now use computational methods and bioinformatics tools to reduce this cost burden. Virtual screening, lead optimization and predictions of bioavailability and bioactivity can help guide experimental research.

- **Time-to-Market:** The predictive power of CADD can help drug research programs choose only the most promising drug candidates. By focusing drug research on specific lead candidate's biopharmaceutical companies can get drugs to market more quickly.
- **Insight:** Molecular models of drug compounds can disclose atomic scale binding properties that are difficult to envision in any other way. Researcher's shows new molecular models to find out protein targets for new binding for new improved compound.

5.7 References

[1] Pauling, L., Corey, R.B. (1951) The structure of proteins: Two hydrogen-bonded helical cofigurations of the polypeptide chain. Proc.Natl.Acad.Sci. U.S.A. 37 p.205-211.

[2] Richmond, T.J. and Richards, F.M. (1978) Packing of alpha-helices: geometrical constraints and contact areas. J. Mol. Biol. 119 p537-555.

[3] Richardson, J. S. (1981) The anatomy and taxonomy of protein structure.Adv. Prot. Chem. 34, p.167-339.

[4] Richardson, J. S. (1977). Beta-Sheet topology and the relatedness of proteins. Nature 268. p.495-500

[5] Koehl, P. and Levitt M. (1999). A brighter future for protein structure prediction. Nature structural biology 6,2 p. 108-112.

[6] Chou, P.Y. and Fasman, G.D. (1978) Empirical predictions of protein conformations. Ann. Rev. Biochem. 47:251-276.

[7] Garnier, J., Gibrat J.-F., and Robson, B. (1996) GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence. Meth. Enz. 266:540-553.

[8] B Rost, and C Sander (1993) Prediction of protein secondary structure at better than 70% accuracy. J Molecular Biol, 232, 584-599.

[9] C. Tanford. ,Protein denaturation. ,*Adv. Prot. Chem.*, 1970, **24**, 1-95.

[10] D. Shortle. ,The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J.*, 1996, **10**, 1, 27-34.

[11] L. J. Smith, K. M. Fiebig, H. Schwalbe, and C. M. Dobson. The concept of the random coil - Residual structure in peptides and denatured proteins. *Fold. Des.*, 1996, **1**, 5, R95-R106.

[12] Binod Kumar and Dr. N. N. Jani**:** Recent Advances in Cheminformatics. International Journal of Intelligent Information Processing, Serials Publications, Vol.3 No.1 (2009) ISSN: 0973-3892.

[13] Binod Kumar and Dr. N. N. Jani: A new Dimension in Molecular Structure Analysis: Better Information for Biochemoinformatist. International Journal of Intelligent Information Processing, Serials Publications,Vol.3 No.2 (2009) ISSN: 0973-3892.

[14] Binod Kumar and Dr. N. N. Jani: Computational Approaches to Drug Design: Bioinformatics Approach. National Conference ETCT 2008, Surat, 2008.

[15] G.R. Marshall, C.D. Barry, H.E. Bosshard, R.A.. Dammkoehler and D.A. Dunn,The Conformational Parameter in Drug Design: The Active Analog Approach, in Computer Assisted Drug Design, ACS Symposia, **112**, E.C. Olson and R.E. Christofferson (Eds.), American Chemical Society, Washington D.C., 1979

Chapter -6

Conformational Study of Molecules using Tools

Chapter 6
CONFORMATIONAL STUDY OF MOLECULE USING TOOLS

6.1 Introduction

Under this research work various chemical and biochemical compounds have been analyzed including drugs using tools like ACD/ChemSketch, NMR Prediction and ArgusLab.

Under the research work of **Activity No.-1** molecules have been analyzed using tool like ACD/ChemSketch and NMR Prediction. In this research using ACD/ChemSketch compounds are stored in databases and SMILE code (Simplified Molecular Input Line Specification) is generated. A SMILE defines the molecules in the form of alphanumeric chains. In this research work chemical shift of every carbon atom of the molecule have been displayed by using NMR Prediction. Using Pubchem/NCBI additional miscellaneous information such as bioactivity analysis by structure & activity similarity and revised compound selection after addition of similar compounds have been found out.

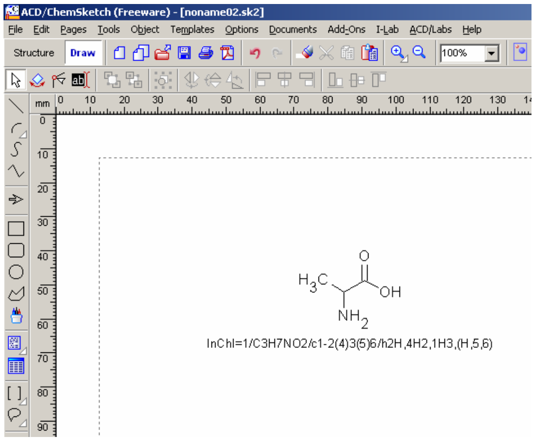
Under the research work of **Activity No.-2** geometry of molecules have been optimized, chemical structure has been visualized and electronic absorption spectra of chemical structure has been calculated by using ArgusLab tool.

Under the research work of **Activity No.-3** different types of analysis like prediction of protein secondary structure, isoelectric point calculation etc. have been performed on nucleotide Sequence and protein sequence using DAMBE and Jemboss tools.

6.2 Experimental Work

6.2.1 Activity No -1

In this research work Alanine (Amino acid) has been used on ACD/ChemSketch editor and its SMILE code has been generated. After that this structured has been transferred to NMRPrediction editor. Similar activity has been performed with Amino butyric acid, Asparagine and Glutamine.


ACD/ChemSketch (Freeware) - [noname02.sk2]
File Edit Pages Tools Object Templates Options Documents Add-Ons I-Lab ACD/Labs Help
Structure Draw
100%
mm
H_3C
O
OH
NH_2

InChI=1/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)

Figure 6.1: Alanine on ACD/ChemSketch editor

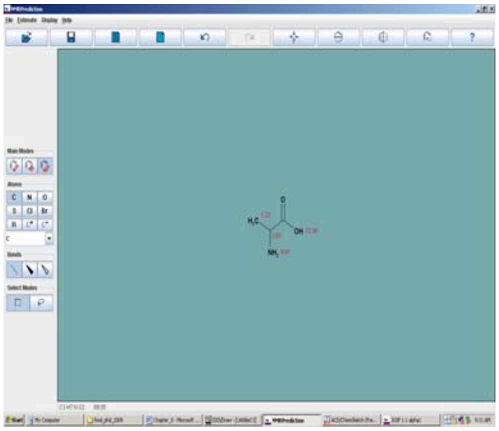


Figure 6.2: ^{13}C NMR of of Alanine

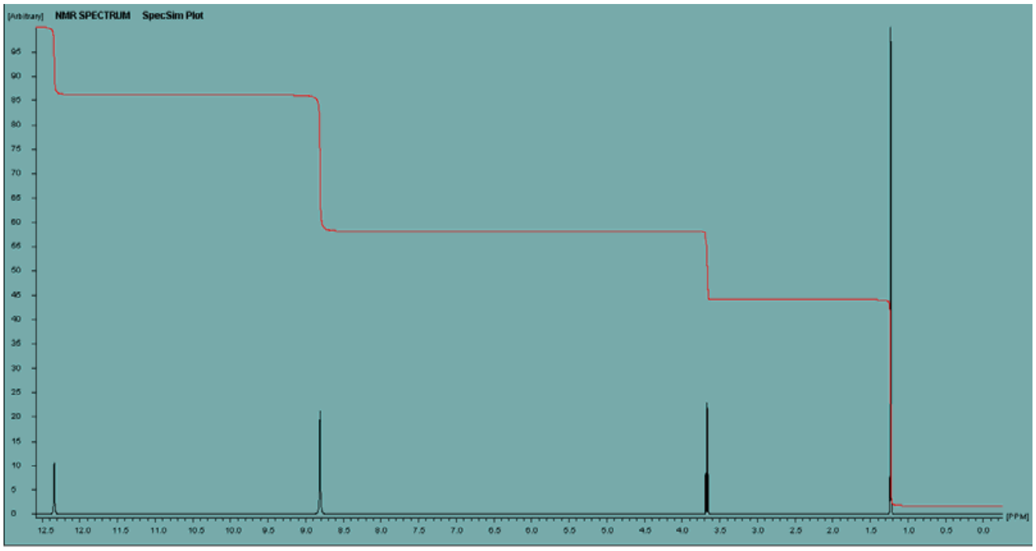


Figure 6.3: Estimation of ^{1}H NMR of Alanine

Table 6.1: Shift Prediction Protocol of Alanine using NMRPrediction

```
Node        Shift       Base + Inc. Comment (ppm rel. to TMS)
 C           174.7        166.0       1-carboxyl
                           11.0       1 -C-C
                           -2.2       general corrections
 CH           51.5         -2.3       aliphatic
                           21.8       1 alpha -C(=O)-O
                            9.1       1 alpha -C
                           28.3       1 alpha -N
                           -5.4       general corrections
 CH3          19.6         -2.3       aliphatic
                            9.1       1 alpha -C
                            2.0       1 beta -C(=O)-O
                           11.3       1 beta -N
                           -0.5       general corrections
```

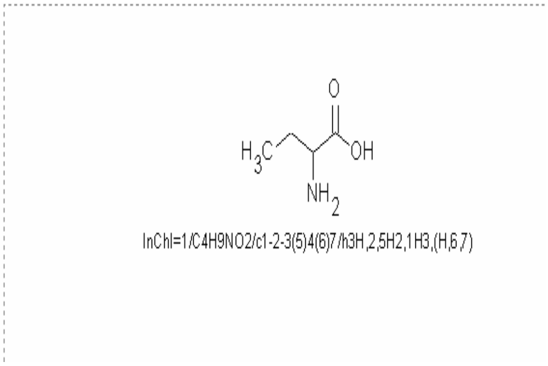


Figure 6.4:2-aminobutanoic acid on ACDChemSketch editor

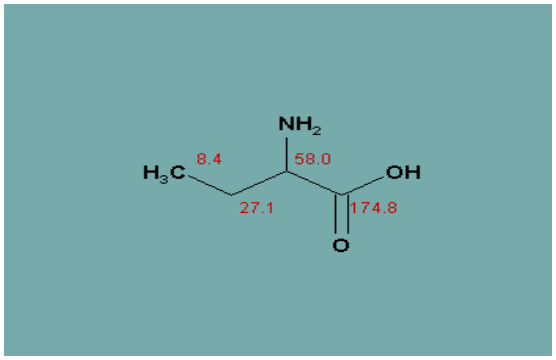


Figure 6.5: ^{13}C NMR of Aminobutanoic acid

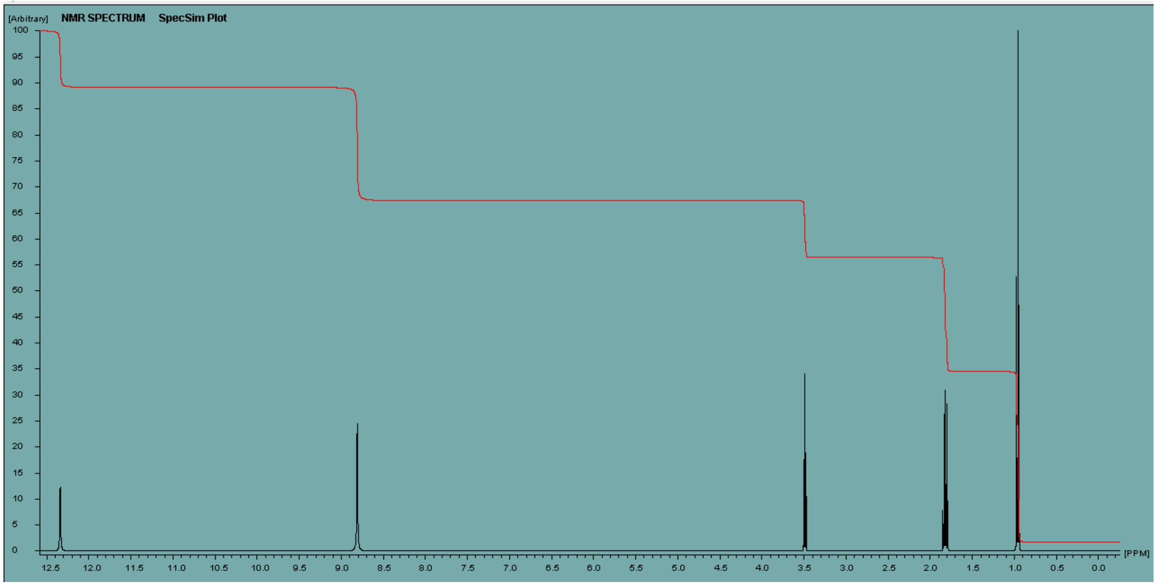
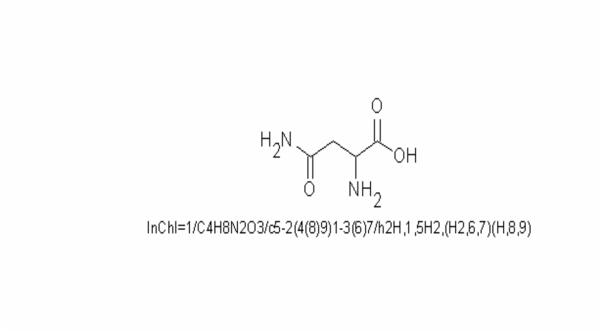


Figure 6.6: Estimation of ^{1}H NMR of Aminobutanoic acid

Table 6.2: Shift Prediction Protocol of 2-aminobutanoic acid using NMRPrediction

Node	Shift	Base + Inc.	Comment (ppm rel. to TMS)
CH	3.49	1.50	methine
		1.13	1 alpha -N
		0.87	1 alpha -C(=O)O
		-0.01	1 beta -C
NH2	8.81	2.00	amine
		6.81	general corrections
CH2	1.82	1.37	methylene
		0.00	1 alpha -C
		0.22	1 beta -N
		0.23	1 beta -C(=O)O
OH	12.34	11.00	carboxylic acid
		1.34	general corrections
CH3	0.96	0.86	methyl
		0.10	1 beta -C-R


H2N
O
OH
O
NH2
InChI=1/C4H8N2O3/c5-2(4(8)9)1-3(6)7/h2H,1,5H2,(H2,6,7)(H,8,9)

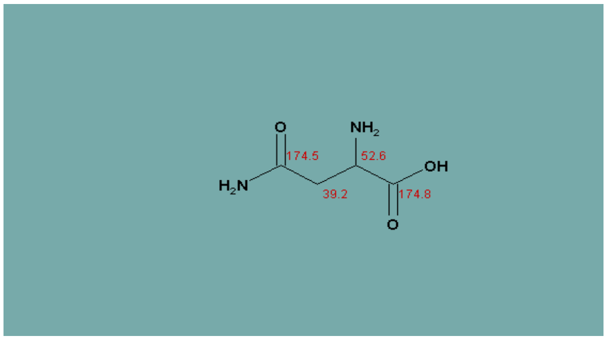


Figure 6.7: Asparagine on ACD/ChemSketch editor

Figure 6.8: ^{13}C NMR of Asparagine

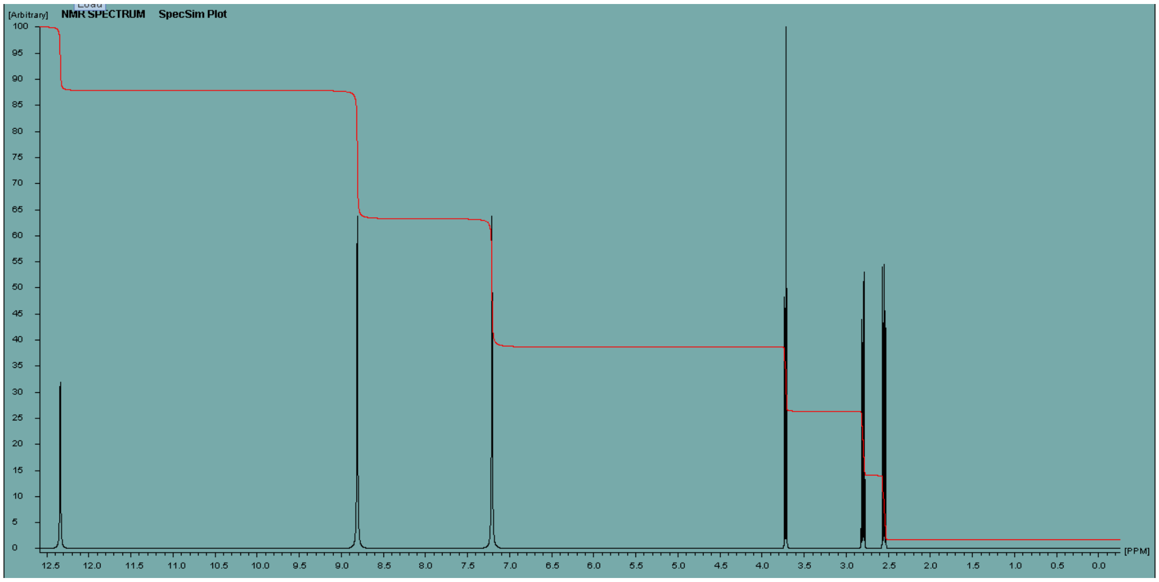


Figure 6.9: Estimation of ^{1}H NMR of Asparagine

Table 6.3: Shift Prediction Protocol of Asparagine using NMRPrediction

Node	Shift	Base + Inc.	Comment (ppm rel. to TMS)
CH2	2.80, 2.55	1.37	methylene
		0.85	1 alpha -C(=O)N
		0.22	1 beta -N
		0.23	1 beta -C(=O)O
CH	3.72	1.50	methine
		1.13	1 alpha -N
		0.87	1 alpha -C(=O)O
		0.22	1 beta -C=O
NH2	7.21	6.00	prim. amide
		1.21	general corrections
NH2	8.81	2.00	amine
		6.81	general corrections
OH	12.34	11.00	carboxylic acid
		1.34	general corrections

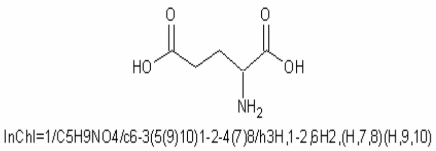

O
O
HO
OH
NH2
InChI=1/C5H9NO4/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H,7,8)(H,9,10)

Figure 6.10: Glutamine on ACD/ChemSketch editor

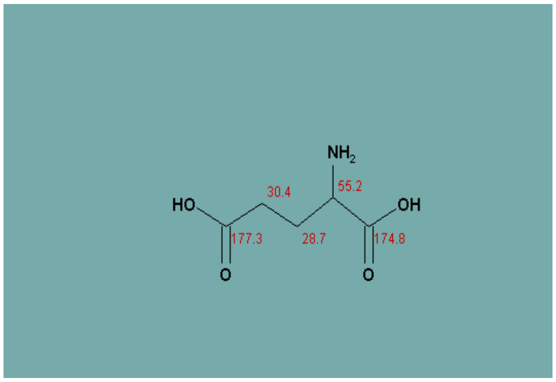


Figure 6.11: ^{13}C NMR of Glutamine

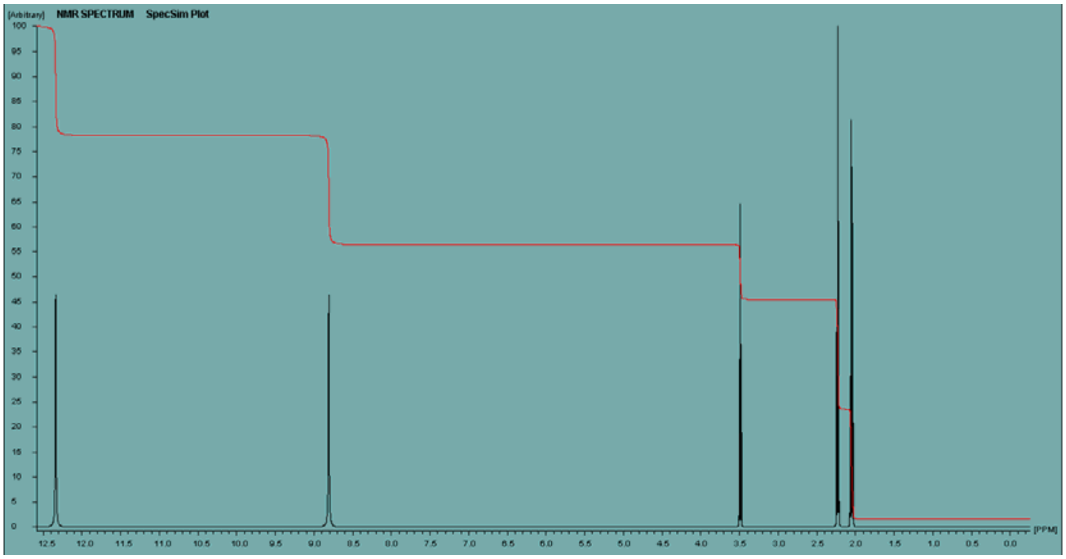
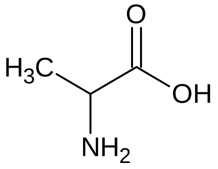
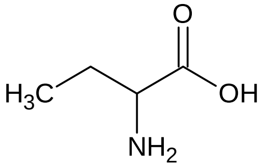
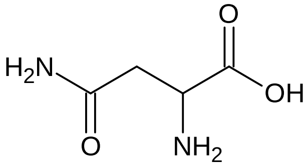
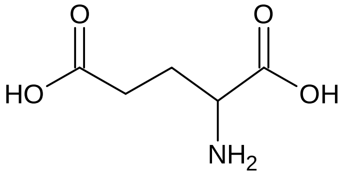


Figure 6.12: Estimation of ^{1}H NMR of Glutamine

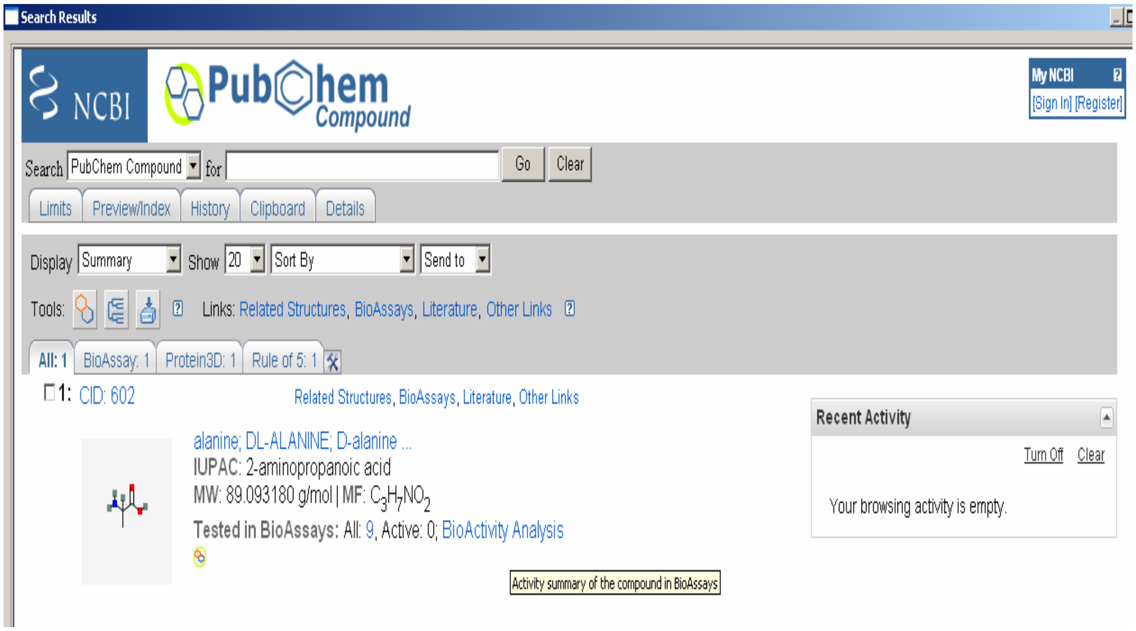
Table 6.4: Shift Prediction Protocol of Glutamine using NMRPrediction

```
Node      Shift    Base + Inc.   Comment (ppm rel. to TMS)

 CH2        2.23         1.37      methylene
                         0.90      1 alpha -C(=O)O
                        -0.04      1 beta -C
 CH2        2.05         1.37      methylene
                         0.23      1 beta -C(=O)O
                         0.22      1 beta -N
                         0.23      1 beta -C(=O)O
 CH         3.49         1.50      methine
                         1.13      1 alpha -N
                         0.87      1 alpha -C(=O)O
                        -0.01      1 beta -C
 OH        12.34        11.00      carboxylic acid
                         1.34      general corrections
 OH        12.34        11.00      carboxylic acid
                         1.34      general corrections
 NH2        8.81         2.00      amine
                         6.81      general corrections
```

Table 6.5: SMILE Code of various structures

Structure	**SMILE Code**
 Alanine	CC(N)C(O)=O
 Amino butyric Acid	CCC(N)C(O)=O
 Asparagine	NC(CC(N)=O)C(O)=O
 Glutamine	NC(CCC(O)=O)C(O)=O

After that web based structure search queries have been performed on these compounds using Pubchem/NCBI. Here activities like Bioactivity Analysis by Structure & Activity Similarity of molecule , Bioactivity Analysis by Structure & Activity Similarity of molecule from Normalized score Percentile , Bioactivity Analysis by Activity & protein target Similarity of molecule from Normalized score Percentile , Bioactivity Analysis by concise Data Table of molecule , Bioactivity Analysis by addition of similar compounds of molecule and Revised compound selection after addition of similar compounds of molecule have been performed.



Search PubChem Compound

Display Summary | Show 20 | Sort By | Send to

1: CID: 602

alanine; DL-ALANINE; D-alanine ...

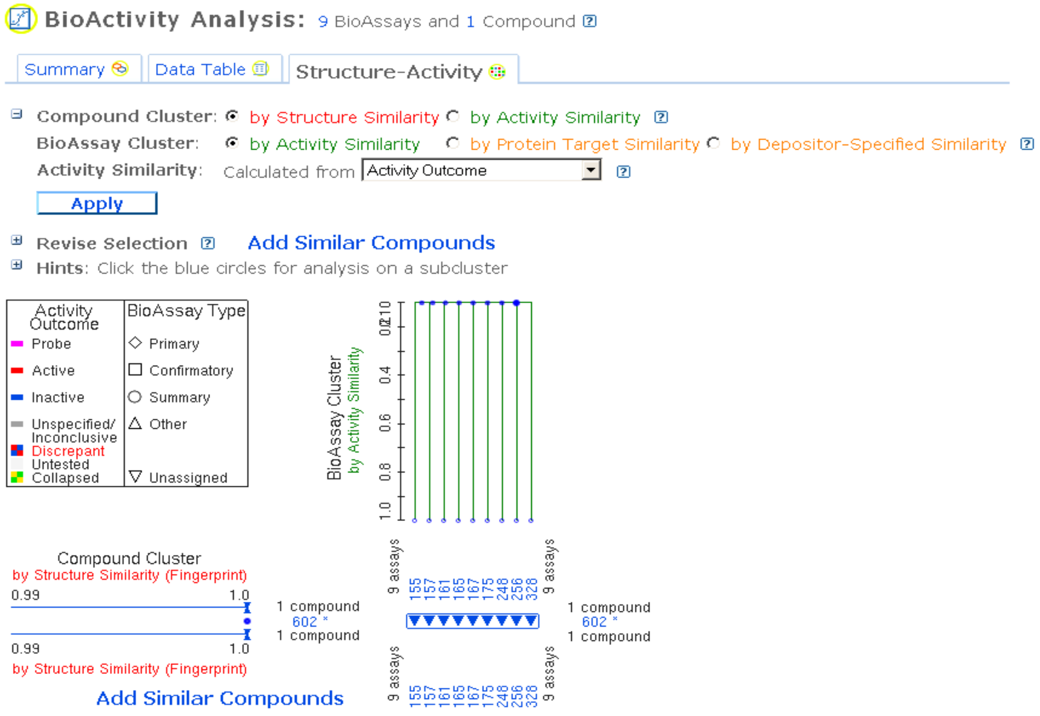
IUPAC: 2-aminopropanoic acid

MW: 89.093180 g/mol | MF: $C_3H_7NO_2$

Tested in BioAssays: All: 9, Active: 0; BioActivity Analysis

Your browsing activity is empty.

Figure 6.13: Search query using Pubchem/NCBI of Analine



BioActivity Analysis: 9 BioAssays and 1 Compound

Summary | Data Table | Structure-Activity

Compound Cluster: ● by Structure Similarity ○ by Activity Similarity

BioAssay Cluster: ● by Activity Similarity ○ by Protein Target Similarity ○ by Depositor-Specified Similarity

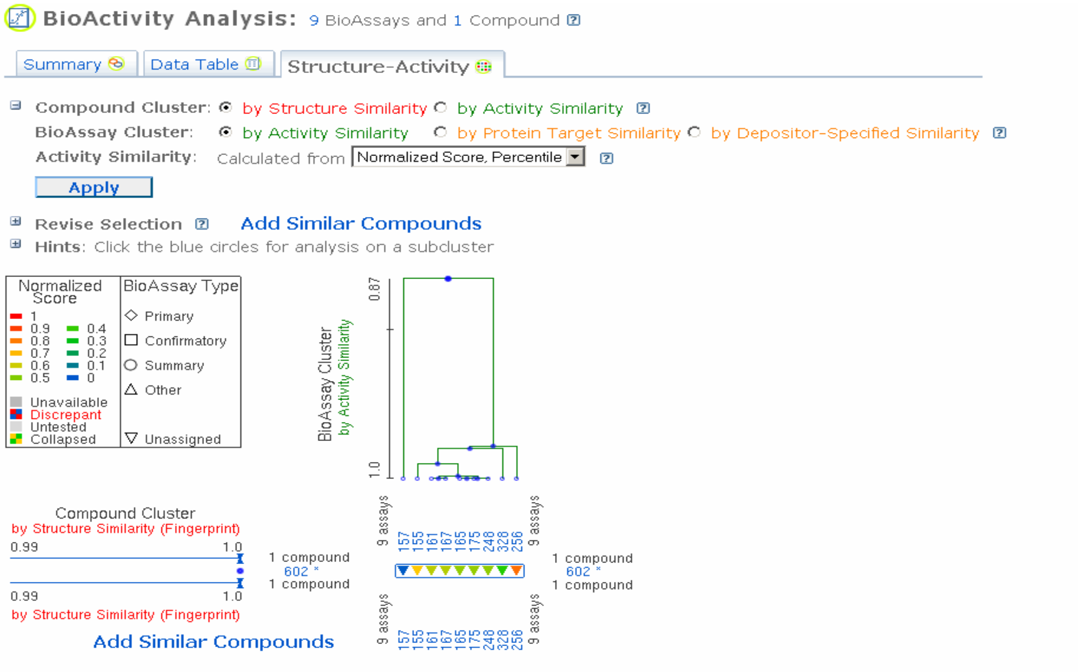
Activity Similarity: Calculated from Activity Outcome

Apply

Revise Selection Add Similar Compounds

Hints: Click the blue circles for analysis on a subcluster

Figure 6.14: Bioactivity Analysis by Structure & Activity Similarity of Analine



BioActivity Analysis: 9 BioAssays and 1 Compound

Summary | Data Table | Structure-Activity

Compound Cluster: ● by Structure Similarity ○ by Activity Similarity

BioAssay Cluster: ● by Activity Similarity ○ by Protein Target Similarity ○ by Depositor-Specified Similarity

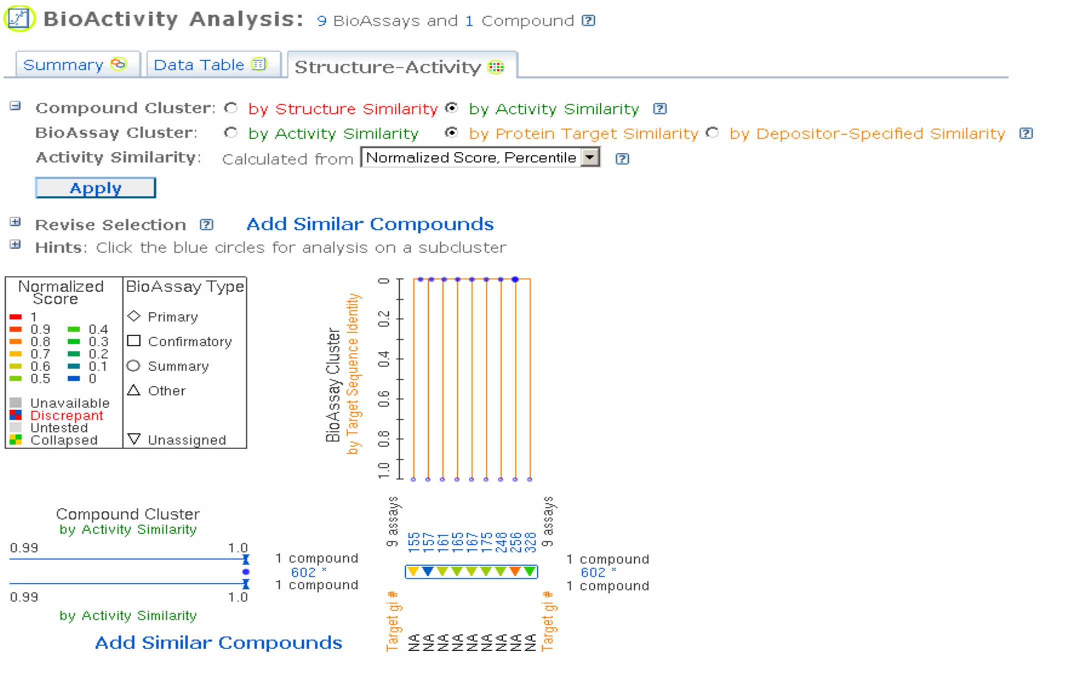
Activity Similarity: Calculated from Normalized Score, Percentile

Apply

Revise Selection Add Similar Compounds

Hints: Click the blue circles for analysis on a subcluster

Figure 6.15: Bioactivity Analysis by Structure & Activity Similarity of Analine from Normalized score Percentile



BioActivity Analysis: 9 BioAssays and 1 Compound

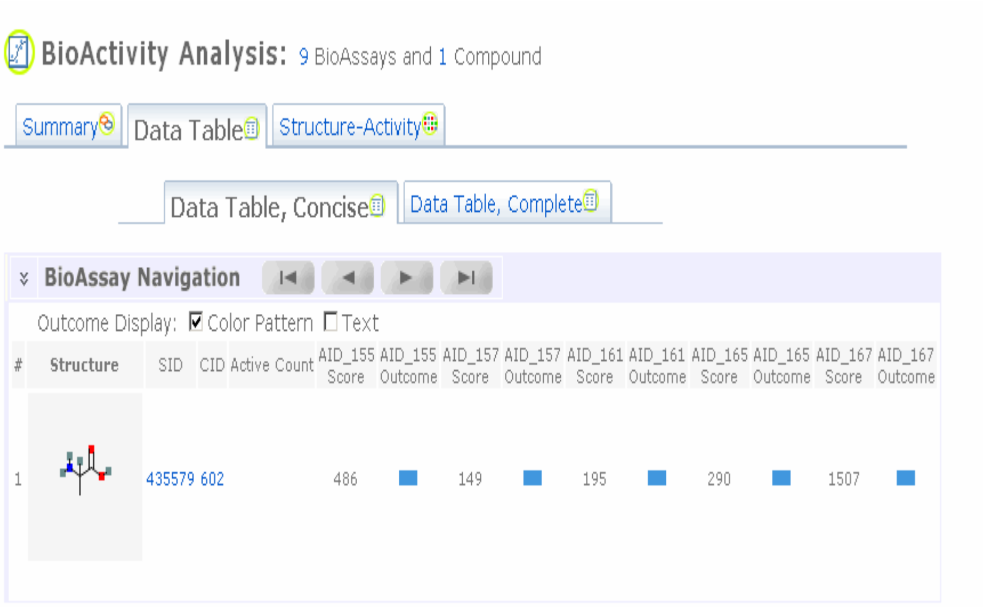
Compound Cluster: ○ by Structure Similarity ● by Activity Similarity

BioAssay Cluster: ○ by Activity Similarity ● by Protein Target Similarity ○ by Depositor-Specified Similarity

Activity Similarity: Calculated from Normalized Score, Percentile

Hints: Click the blue circles for analysis on a subcluster

Figure 6.16: Bioactivity Analysis by Activity & protein target Similarity of Analine from Normalized score Percentile



BioActivity Analysis: 9 BioAssays and 1 Compound

Outcome Display: ☑ Color Pattern □ Text

#	Structure	SID	CID	Active Count	AID_155 Score	AID_155 Outcome	AID_157 Score	AID_157 Outcome	AID_161 Score	AID_161 Outcome	AID_165 Score	AID_165 Outcome	AID_167 Score	AID_167 Outcome
1		435579	602		486	🟦	149	🟦	195	🟦	290	🟦	1507	🟦

Figure 6.17: Bioactivity Analysis by concise Data Table of Analine

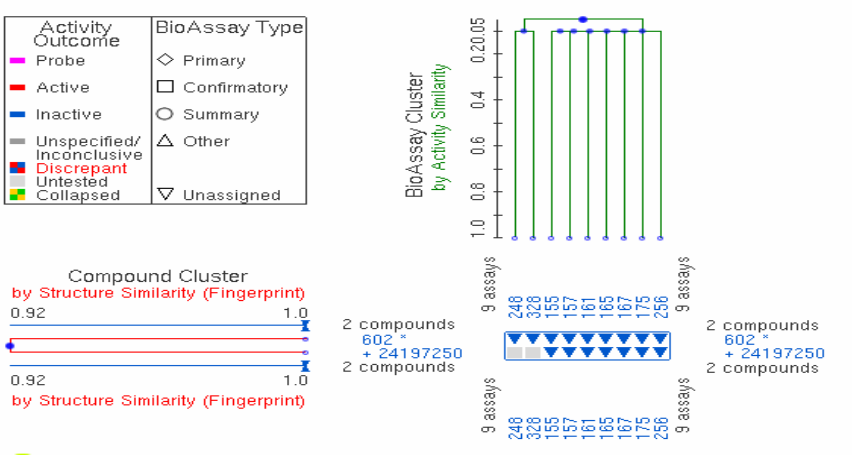


Figure 6.18: Bioactivity Analysis by addition of similar compounds of Analine

Table 6.6: 9- assays by addition of similar compound of Analine

CID/ Activity /AID	248	328	155	157	161	165	167	175	256
602	Inact	Inact	Inact	Inact	Inact	Inact	Inact	Inact	Inact
24197250			Inact	Inact	Inact	Inact	Inact	Inact	Inact

Inact = Inactive

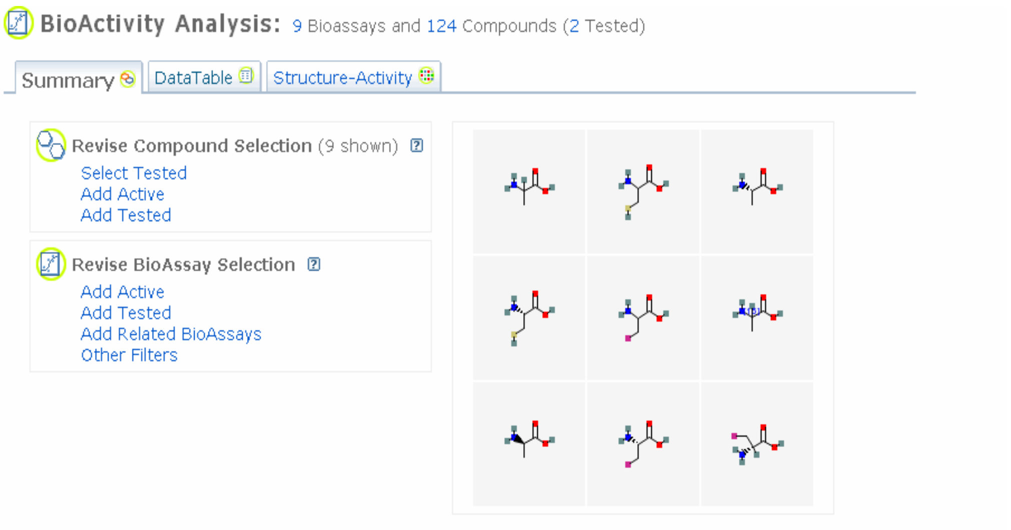
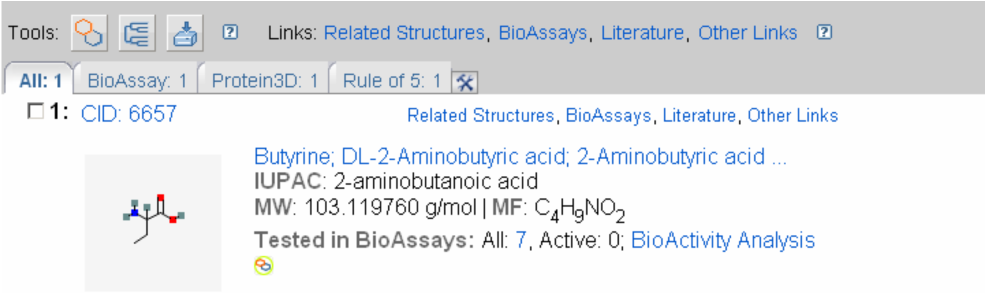

BioActivity Analysis: 9 Bioassays and 124 Compounds (2 Tested)
Summary
DataTable
Structure-Activity
Revise Compound Selection (9 shown)
Select Tested
Add Active
Add Tested
Revise BioAssay Selection
Add Active
Add Tested
Add Related BioAssays
Other Filters

Figure 6.19: Revised compound selection after addition of similar compounds of Analine



Tools: Links: Related Structures, BioAssays, Literature, Other Links

All: 1 | BioAssay: 1 | Protein3D: 1 | Rule of 5: 1

□ 1: CID: 6657 Related Structures, BioAssays, Literature, Other Links

Butyrine; DL-2-Aminobutyric acid; 2-Aminobutyric acid ...
IUPAC: 2-aminobutanoic acid
MW: 103.119760 g/mol | MF: $C_4H_9NO_2$
Tested in BioAssays: All: 7, Active: 0; BioActivity Analysis

Figure 6.20: Search query using Pubchem/NCBI of Amino butyric Acid

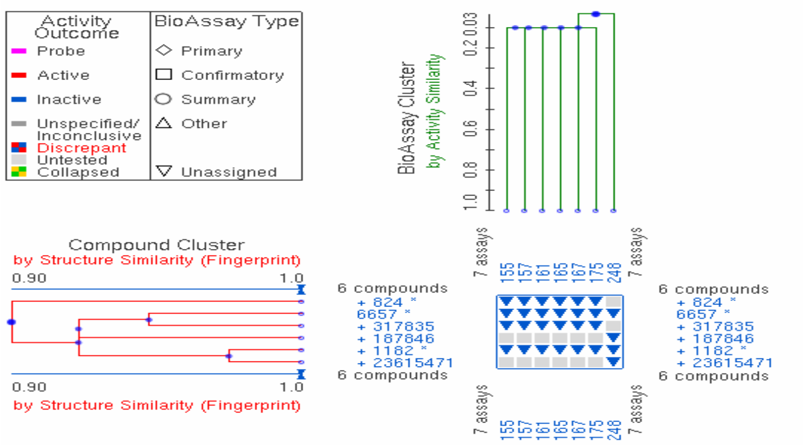


Figure 6.21: Bioactivity Analysis by addition of similar compounds of Amino butyric Acid

Table 6.7: 7- assays by addition of similar compound Amino butyric Acid

CID/ Activity /AID	155	157	161	165	167	175	248
824	Inact	Inact	Inact	Inact	Inact	Inact	
6657	Inact	Inact	Inact	Inact	Inact	Inact	Inact
317835	Inact	Inact	Inact	Inact	Inact	Inact	
187846							Inact
1182	Inact	Inact	Inact	Inact	Inact	Inact	Inact
23615471							Inact

Inact = Inactive

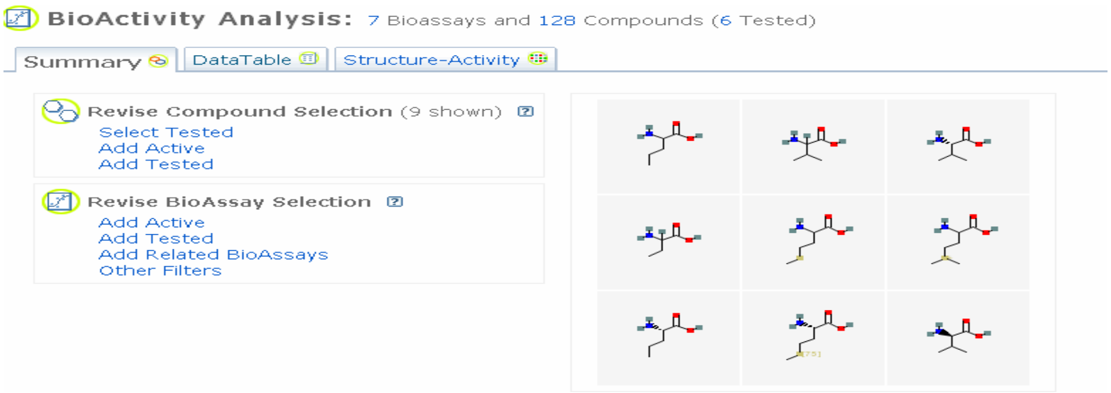
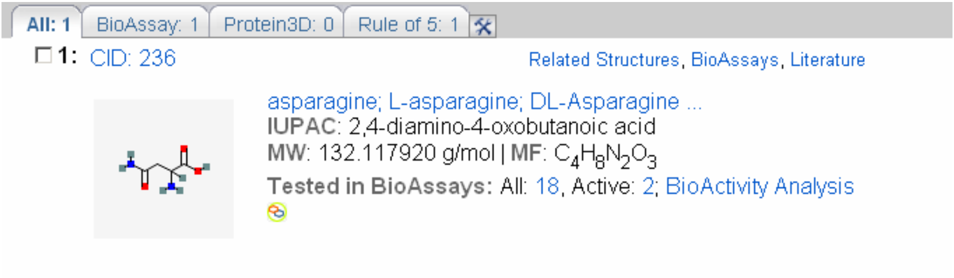

BioActivity Analysis: 7 Bioassays and 128 Compounds (6 Tested)
Summary
DataTable
Structure-Activity
Revise Compound Selection (9 shown)
Select Tested
Add Active
Add Tested
Revise BioAssay Selection
Add Active
Add Tested
Add Related BioAssays
Other Filters

Figure 6.22: Revised compound selection after addition of similar compounds of Amino butyric Acid


All: 1 | BioAssay: 1 | Protein3D: 0 | Rule of 5: 1

□ 1: CID: 236 Related Structures, BioAssays, Literature

asparagine; L-asparagine; DL-Asparagine ...

IUPAC: 2,4-diamino-4-oxobutanoic acid

MW: 132.117920 g/mol | MF: $C_4H_8N_2O_3$

Tested in BioAssays: All: 18, Active: 2; BioActivity Analysis

Figure 6.23: Search query using Pubchem/NBCI of Asparagine

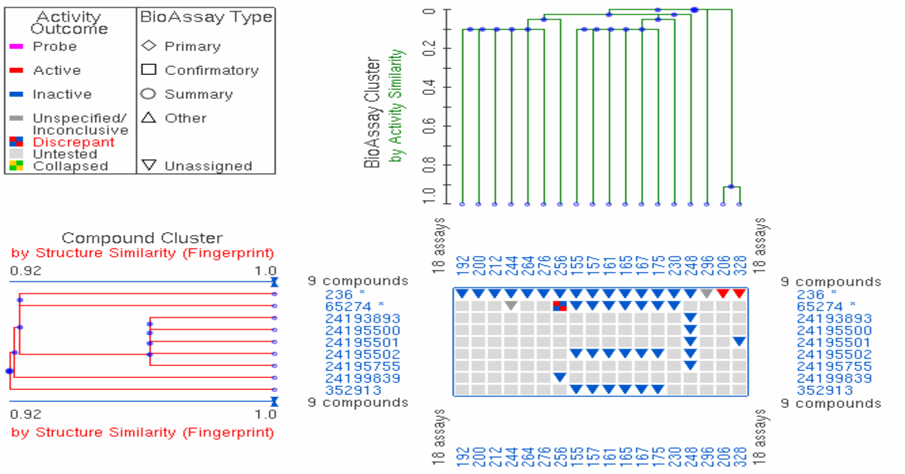
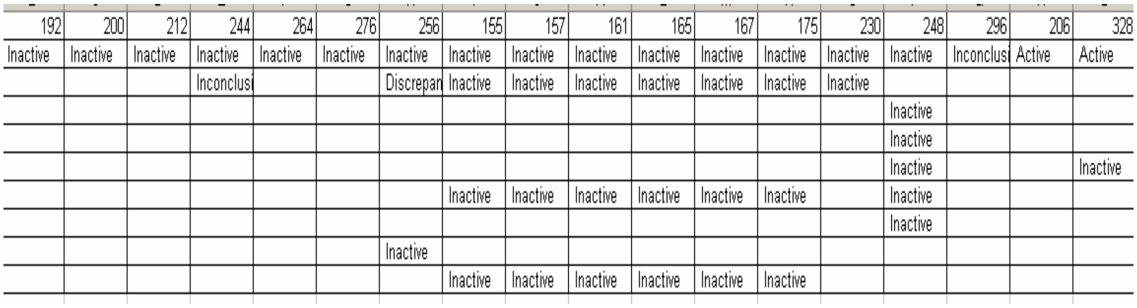


Figure 6.24: Bioactivity Analysis by addition of similar compounds of Asparagine

Table 6.8: 18- assays by addition of similar compound Asparagine



192	200	212	244	264	276	256	155	157	161	165	167	175	230	248	296	206	328
Inactive	Inactive	Inactive	Inactive	Inactive	Inactive	Inactive	Inactive	Inactive	Inactive	Inactive	Inactive	Inactive	Inactive	Inactive	Inconclusi	Active	Active
			Inconclusi			Discrepan	Inactive	Inactive	Inactive	Inactive	Inactive	Inactive	Inactive				
														Inactive			
														Inactive			
														Inactive			Inactive
							Inactive	Inactive	Inactive	Inactive	Inactive	Inactive		Inactive			
														Inactive			
						Inactive											
							Inactive	Inactive	Inactive	Inactive	Inactive	Inactive					

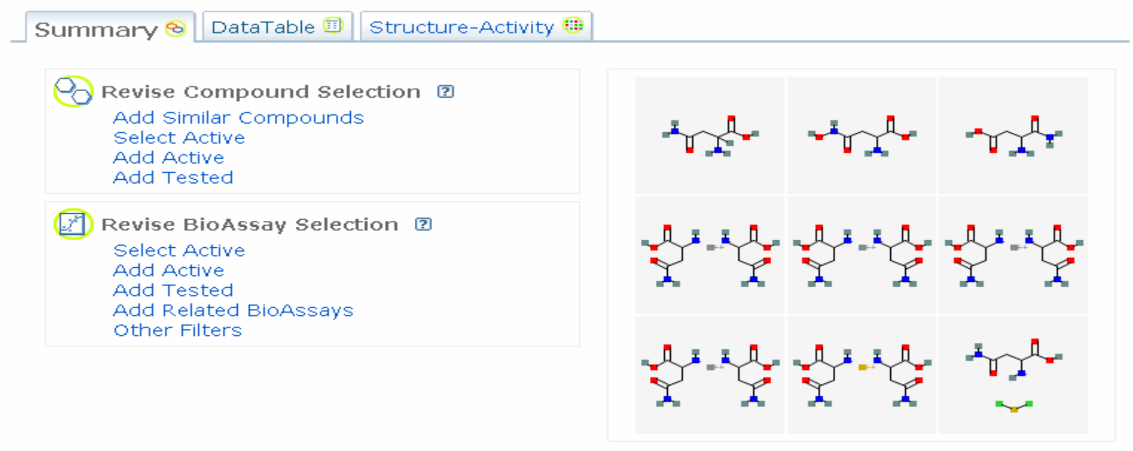
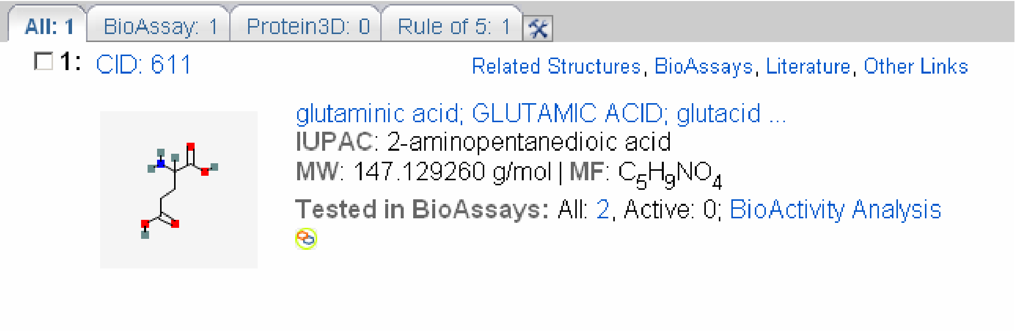


Figure 6.25: Revised compound selection after addition of similar compounds of Asparagine



□ 1: CID: 611 Related Structures, BioAssays, Literature, Other Links

glutaminic acid; GLUTAMIC ACID; glutacid ...

IUPAC: 2-aminopentanedioic acid

MW: 147.129260 g/mol | MF: $C_5H_9NO_4$

Tested in BioAssays: All: 2, Active: 0; BioActivity Analysis

Figure 6.26: Search query using Pubchem/NCBI of Glutamine

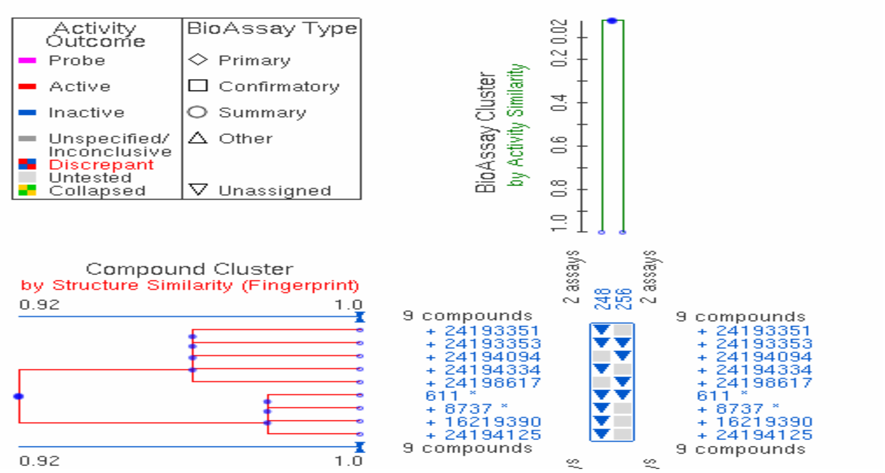


Figure 6.27: Bioactivity Analysis by addition of similar compounds of Glutamine

Table 6.9: 2- assays by addition of similar compound Glutamine

CID/Activity/AID	248	256
24193351	Inactive	
24193353	Inactive	Inactive
24194094		Inactive
24194334	Inactive	
24198617		Inactive
611	Inactive	Inactive
8737	Inactive	
16219390	Inactive	
24194125	Inactive	

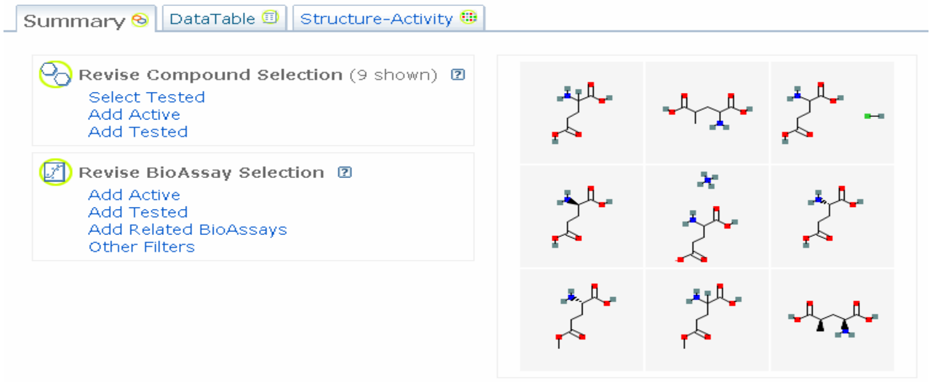


Figure 6.28: Revised compound selection after addition of similar compounds of Glutamine

6.2.2 Activity No -2

In this research work Alanine, Amino butyric acid, Asparagine and Glutamine have been analyzed on ArgusLab tool and calculations like Single Entry Point Calculation, Geometric Optimization and UV Electronic Spectra have been performed.

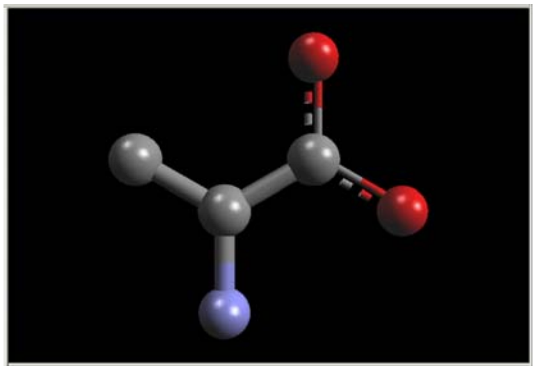


Figure No 6.29 : 3D-Molecular structure of Analine on ArgusLab

Table No 6.10: Single Entry Point Calculation of Alanine Using ArgusLab

```
*********  Validated Experiment & Chemical System Settings  **********

  Calculation started:  Thu Jun 04 10:37:09 2009


  Title:E:\myphdthesis_2009\ACd_results\Ala

  Max. SCF cycles           100
  SCF convergence           1.5936e-013 au. for energy


  Input Atomic Information
  ************************

      1   C   16.904500  -7.535300   0.000000
      2   O   16.904500  -6.205300   0.000000
      3   C   15.752700  -8.200300   0.000000
      4   N   15.752700  -9.530300   0.000000
      5   C   14.600800  -7.535300   0.000000
      6   O   18.056300  -8.200300   0.000000



         Constructing Chemical System(s)


         Basis Set
        ***********

          basis functions : 24
          shells          : 12
          primitives      : 72


         Memory for Main Chemical System
         Max. number 2-ele. ints. = 1596
         Memory Requirements (bytes)
        *****************************

         Core               121248
         Scratch            10368
         System charge      0.000000


***** SCF *****

        Core repulsion    79.3611 au

        Calculating one electron matrix
        Diagonalizing starting one-ele. matrix
        Performing SCF


           Cycle        Energy (au)          Difference
        **********************************************

            1        -36.971063
            2        -32.276391119              4.69467
            3        -33.620242051             -1.34385
            4        -26.283142152               7.3371
            5        -33.505498559             -7.22236
            6        -26.230341192              7.27516
            7        -33.507013214             -7.27667
            8        -26.229486478              7.27753
            9        -33.508071400             -7.27858
           10        -26.229663085              7.27841
           11        -33.508488479             -7.27883
           12        -26.229742535              7.27875
           13        -33.508637619              -7.2789
           14        -26.229770710              7.27887
           15        -33.508689823             -7.27892
           16        -26.229780460              7.27891
           17        -33.508708012             -7.27893
           18        -26.229783833              7.27892
           19        -33.508714344             -7.27893
           20        -26.229785002              7.27893
           21        -33.508716547             -7.27893
           22        -26.229785408              7.27893
           23        -33.508717315             -7.27893
           24        -26.229785549              7.27893
           25        -33.508717582             -7.27893
           26        -26.229785599              7.27893
           27        -33.508717675             -7.27893
           28        -26.229785616              7.27893
           29        -33.508717707             -7.27893
           30        -26.229785622              7.27893
           31        -33.508717718             -7.27893
           32        -26.229785624              7.27893
           33        -33.508717722             -7.27893
           34        -26.229785624              7.27893
           35        -33.508717723             -7.27893
           36        -26.229785625              7.27893
           37        -33.508717724             -7.27893
           38        -26.229785625              7.27893
           39        -33.508717724             -7.27893
           40        -26.229785625              7.27893
           41        -33.508717724             -7.27893
           42        -26.229785625              7.27893
           43        -33.508717724             -7.27893
           44        -26.229785625              7.27893
           45        -33.508717724             -7.27893
           46        -26.229785625              7.27893
           47        -33.508717724             -7.27893
           48        -26.229785625              7.27893
           49        -33.508717724             -7.27893
           50        -26.229785625              7.27893
           51        -33.508717724             -7.27893
           52        -26.229785625              7.27893
           53        -33.508717724             -7.27893
           54        -26.229785625              7.27893
           55        -33.508717724             -7.27893
           56        -26.229785625              7.27893
           57        -33.508717724             -7.27893
           58        -26.229785625              7.27893
           59        -33.508717724             -7.27893
           60        -26.229785625              7.27893
           61        -33.508717724             -7.27893
           62         -26.229785625             7.27893
           63         -33.508717724            -7.27893
           64         -26.229785625             7.27893
           65         -33.508717724            -7.27893
           66         -26.229785625             7.27893
           67         -33.508717724            -7.27893
           68         -26.229785625             7.27893
           69         -33.508717724            -7.27893
           70         -26.229785625             7.27893
           71         -33.508717724            -7.27893
           72         -26.229785625             7.27893
           73         -33.508717724            -7.27893
           74         -26.229785625             7.27893
           75         -33.508717724            -7.27893
           76         -26.229785625             7.27893
           77         -33.508717724            -7.27893
           78         -26.229785625             7.27893
           79         -33.508717724            -7.27893
           80         -26.229785625             7.27893
           81         -33.508717724            -7.27893
           82         -26.229785625             7.27893
           83         -33.508717724            -7.27893
           84         -26.229785625             7.27893
           85         -33.508717724            -7.27893
           86         -26.229785625             7.27893
           87         -33.508717724            -7.27893
           88         -26.229785625             7.27893
           89         -33.508717724            -7.27893
           90         -26.229785625             7.27893
           91         -33.508717724            -7.27893
           92         -26.229785625             7.27893
           93         -33.508717724            -7.27893
           94         -26.229785625             7.27893
           95         -33.508717724            -7.27893
           96         -26.229785625             7.27893
           97         -33.508717724            -7.27893
           98         -26.229785625             7.27893
           99         -33.508717724            -7.27893
          100         -26.229785625             7.27893

        Maximum number of iterations reached: SCF NOT CONVERGED!
```

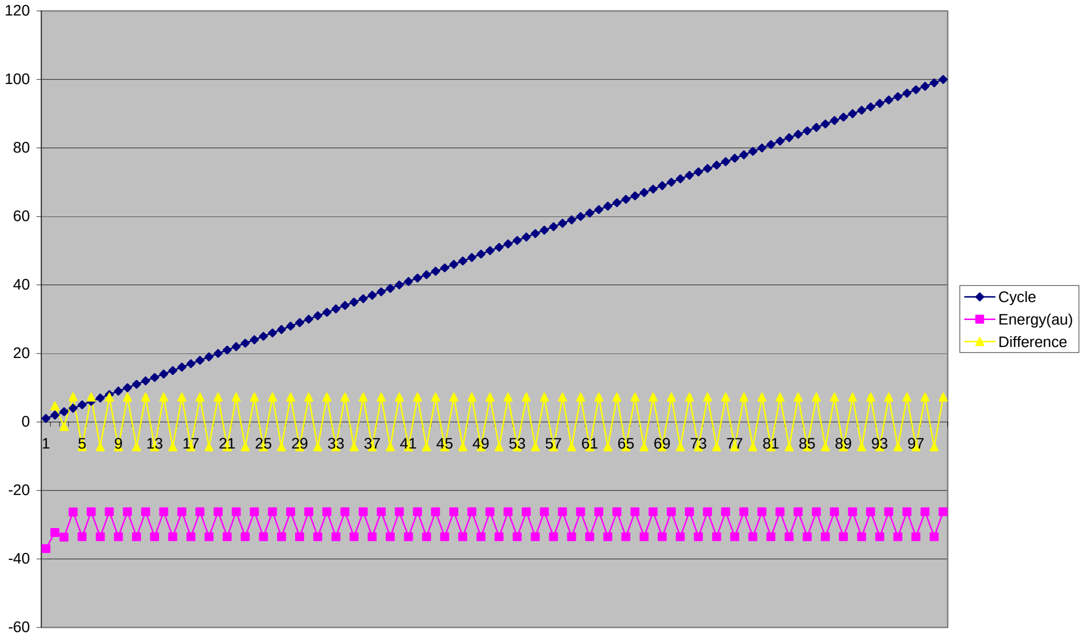


Diagram 6.30: Energy vs. Difference S.E.C of Alanine

`

```
Writing final SCF to disk

       Final SCF Energy =  -26.2297856247 au
       Final SCF Energy =  -16459.4538 kcal/mol


        ***** Heat of Formation *****
           11889.4613 kcal/mol


        Wiberg Atom-Atom Bond Orders
       *****************************


       1           2           3           4           5           6

  1   0.000000
  2   0.331606    0.000000
  3   0.241269    0.010621    0.000000
  4   0.000639    0.000104    0.000185    0.000000
  5   0.003243    0.000301    0.001393    0.000002    0.000000
  6   0.327085    0.009206    0.009767    0.000086    0.000241    0.000000
        Atomic spin densities
       **********************

           1    C    0.6170
           2    O    0.0225
           3    C    0.3284
           4    N    0.0005
           5    C    0.0044
           6    O    0.0273
 S2 operator
 ***********

 exact                 0.750000
 calculated            0.750201


                       Ground State Dipole (debye)

                    X             Y             Z          length
            106.85663544   55.51461066   -0.00000000  120.41682828


        Mulliken Atomic Charges
       ************************

           1    C     2.6734
           2    O     4.0964
           3    C    -3.8542
           4    N    -3.0018
           5    C    -3.9958
           6    O     4.0820


       Properties elapsed time 0 sec.


       Total Elapsed Time 0 sec.
```

Table No 6.11: Geometry Optimization of Analine using ArgusLab.

```
*********  Validated Experiment & Chemical System Settings  **********


  Calculation started:  Thu Jun 04 11:06:01 2009


  Title:E:\myphdthesis_2009\ACd_results\Ala


  Max. SCF cycles           100
  SCF convergence           1.5936e-013 au. for energy

  Max. geom cycles          10
  Convergence criteria:
  max. grad. component <    0.000084   au.
  AM1 param file            C:\Program Files\ArgusLab\params\am1.prm
  SCF saved every           1000 cycles


Input Atomic Information
  ************************

      1   C   16.435771  -7.307230   0.037217
      2   O   17.070817  -6.090942  -0.014829
      3   C   15.266201  -8.194435  -0.019871
      4   N   16.383638  -9.920985   0.002328
      5   C   14.143594  -7.191811  -0.001207
      6   O   18.671478  -8.501396  -0.003638



        Constructing Chemical System(s)


        Basis Set
       ***********

         basis functions : 24
         shells          : 12
         primitives      : 36


        Memory for Main Chemical System
        Max. number 2-ele. ints. = 1596

        Memory Requirements (bytes)
       *****************************

        Core              325080
        Scratch           10368
        System charge     0.000000


***** SCF *****

       Core repulsion    66.1182 au

       Calculating one electron matrix
       Diagonalizing starting one-ele. matrix
       Performing SCF


        Cycle        Energy (au)          Difference
       **********************************************

            1       -36.702730
            2       -31.180576166              5.52215
            3       -34.977490714             -3.79691
            4       -26.267268379              8.71022
            5       -36.048536714             -9.78127
            6       -27.952390982              8.09615
            7       -36.298981727             -8.34659
            8       -27.939103265              8.35988
            9       -35.978322254             -8.03922
           10       -27.902782233              8.07554
           11       -35.690378260             -7.7876
           12       -27.894461588              7.79592
           13       -35.629035870             -7.73457
           14       -27.895280755              7.73376
           15       -35.613901033             -7.71862
           16       -27.896032136              7.71787
           17       -35.609077364             -7.71305
           18       -27.896355285              7.71272
           19       -35.607344655             -7.71099
           20       -27.896481549              7.71086
           21       -35.606693537             -7.71021
           22       -27.896530159              7.71016
           23       -35.606444862             -7.70991
           24       -27.896548859               7.7099
           25       -35.606349333              -7.7098
           26       -27.896556058              7.70979
           27       -35.606312557             -7.70976
           28       -27.896558832              7.70975
           29       -35.606298388             -7.70974
           30       -27.896559901              7.70974
           31       -35.606292928             -7.70973
```

```
 32        -27.896560313              7.70973
 33        -35.606290823             -7.70973
 34        -27.896560472              7.70973
 35        -35.606290012             -7.70973
 36        -27.896560533              7.70973
 37        -35.606289699             -7.70973
 38        -27.896560557              7.70973
 39        -35.606289579             -7.70973
 40        -27.896560566              7.70973
 41        -35.606289532             -7.70973
 42        -27.896560569              7.70973
 43        -35.606289514             -7.70973
 44        -27.896560571              7.70973
 45        -35.606289508             -7.70973
 46        -27.896560571              7.70973
 47        -35.606289505             -7.70973
 48        -27.896560571              7.70973
 49        -35.606289504             -7.70973
 50        -27.896560571              7.70973
 51        -35.606289503             -7.70973
 52        -27.896560572              7.70973
 53        -35.606289503             -7.70973
 54        -27.896560572              7.70973
 55        -35.606289503             -7.70973
 56        -27.896560572              7.70973
 57        -35.606289503             -7.70973
 58        -27.896560572              7.70973
 59        -35.606289503             -7.70973
 60        -27.896560572              7.70973
 61        -35.606289503             -7.70973
 62        -27.896560572              7.70973
 63        -35.606289503             -7.70973
 64        -27.896560572              7.70973
 65        -35.606289503             -7.70973
 66        -27.896560572              7.70973
 67        -35.606289503             -7.70973
 68        -27.896560572              7.70973
 69        -35.606289503             -7.70973
 70        -27.896560572              7.70973
 71        -35.606289503             -7.70973
 72        -27.896560572              7.70973
 73        -35.606289503             -7.70973
 74        -27.896560572              7.70973
 75        -35.606289503             -7.70973
 76        -27.896560572              7.70973
 77        -35.606289503             -7.70973
 78        -27.896560572              7.70973
 79        -35.606289503             -7.70973
 80        -27.896560572              7.70973
 81        -35.606289503             -7.70973
 82        -27.896560572              7.70973
 83        -35.606289503             -7.70973
 84        -27.896560572              7.70973
 85        -35.606289503             -7.70973
 86        -27.896560572              7.70973
 87        -35.606289503             -7.70973
 88        -27.896560572              7.70973
 89        -35.606289503             -7.70973
 90        -27.896560572              7.70973
 91        -35.606289503             -7.70973
 92        -27.896560572              7.70973
 93        -35.606289503             -7.70973
 94        -27.896560572              7.70973
 95        -35.606289503             -7.70973
 96        -27.896560572              7.70973
 97        -35.606289503             -7.70973
 98        -27.896560572              7.70973
 99        -35.606289503             -7.70973
100        -27.896560572              7.70973
```

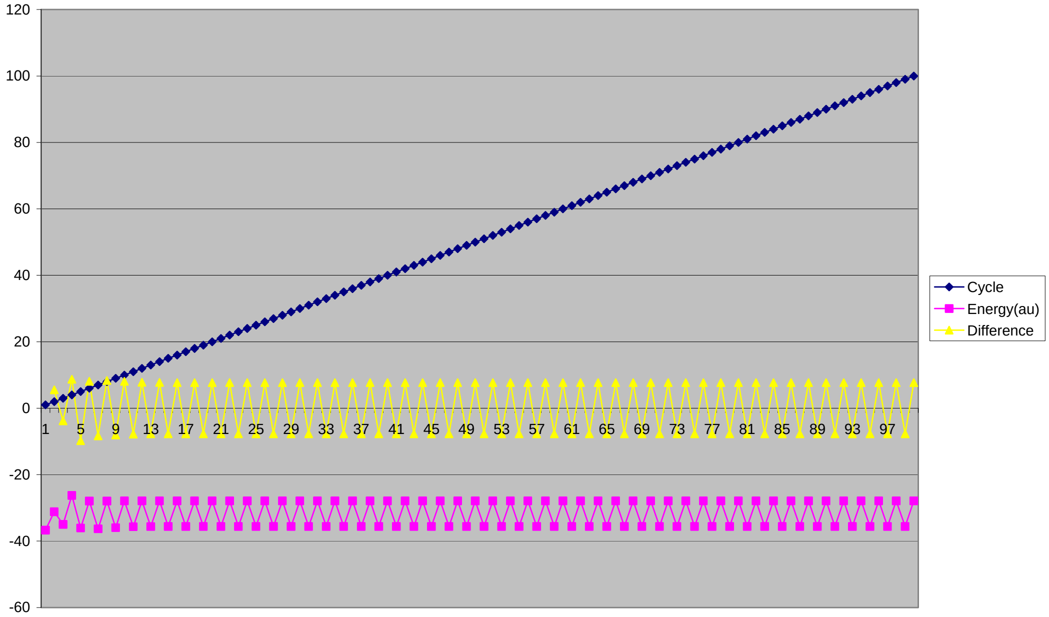


Figure 6.31: Energy vs. Difference of geometric optimization of Analine

```
        Maximum number of iterations reached: SCF NOT CONVERGED!


        Writing final SCF to disk

        Final SCF Energy =  -27.8965605715 au
        Final SCF Energy =  -17505.3718 kcal/mol

        Saving      the      final      SCF      to      the      restart      file
E:\myphdthesis_2009\ACd_results\Ala.restartscf


        SCF elapsed time 1 sec.
```

Table 6.12: Geometric for different components of Analine

```
***** Geometry Optimization *****

        Checkpointing coordinate to E:\myphdthesis_2009\ACd_results\Ala.cor


        Geometry Search using BFGS update


         Cycle    Energy(au)    delE (au)     Grad Norm   |Max Grad(i)|    alpha
        ************************************************************************

        start    -27.896561     0.0000e+000     2.662552     1.270036
          1      -36.561033    -8.6645e+000     1.698515     1.124240    1.5023e-001
          2      -37.477903    -9.1687e-001     1.227344     0.768170    1.1775e-001
          3      -29.966887     7.5110e+000     3.034984     1.893725    2.5710e-031
          4      -36.328840    -6.3620e+000     1.450727     0.862822    1.3180e-001
          5      -29.337689     6.9912e+000     1.962616     1.270998    2.1751e-031
          6      -34.996943    -5.6593e+000     1.393716     0.687540    2.0381e-001
          7      -35.809607    -8.1266e-001     1.285275     0.586843    1.4350e-001
          8      -30.962245     4.8474e+000     2.617189     1.509296    2.4551e-031
          9      -37.516635    -6.5544e+000     1.251350     0.814439    1.5284e-001
         10      -30.178555     7.3381e+000     1.996299     1.450952    2.5216e-031
```

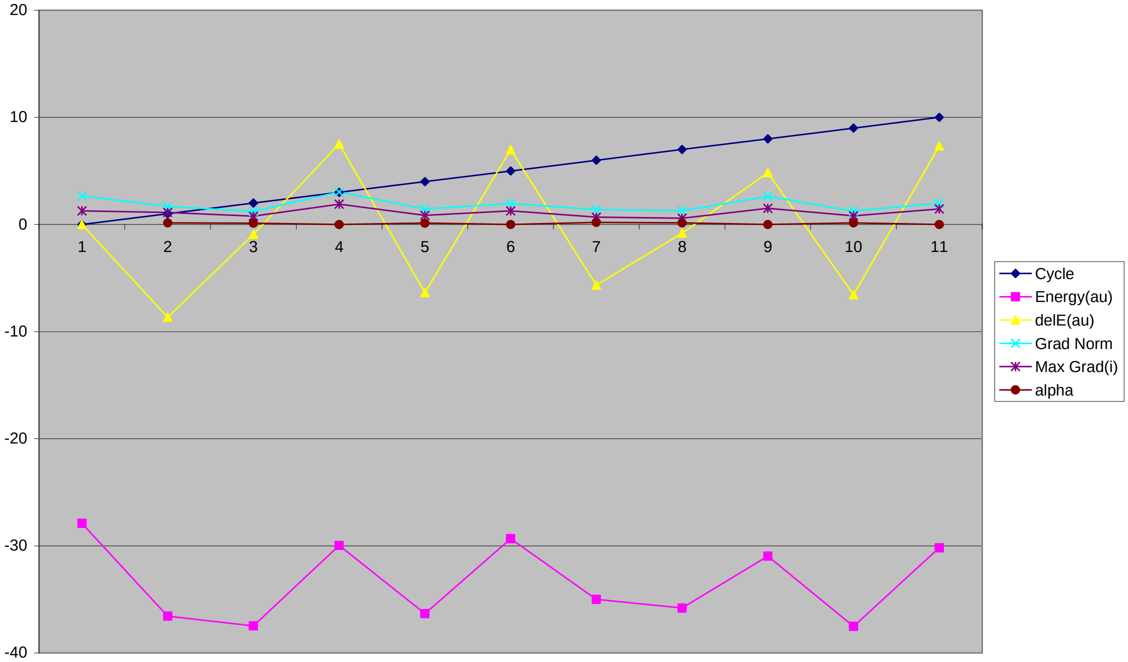


Figure 6.32: Geometric for different components of Analine

```
        >>>Geometry optimization did not converge<<<


        Maximum cycles reached, optimization terminated

         ****************  Final Geometry ****************

          C   16.58012803   -7.44007254    0.04973112      6
          O   16.95387408   -6.07861932   -0.02119125      8
          C   15.45848448   -8.29244282   -0.02981828      6
          N   16.27002874   -9.85642611    0.00468287      7
          C   14.07596372   -7.09355715    0.00047493      6
          O   18.63302097   -8.44568207   -0.00387938      8


        Final Geom Energy =  -30.1785551555 au
        Final Geom Energy =  -18937.3464 kcal/mol

        Geometry Optimization elapsed time 1 min. 54 sec.

         ***** Heat of Formation *****
             6661.9791 kcal/mol

         Atomic spin densities
        **********************

            1    C    0.5857
            2    O    0.0239
            3    C    0.3962
            4    N    0.0000
            5    C   -0.0059
            6    O    0.0000

 S2 operator
 ***********
 exact                     0.750000
 calculated                0.751934

        Properties elapsed time 0 sec.

        Total Elapsed Time 1 min. 55 sec.
```

Quick Plot HOMO : Analine

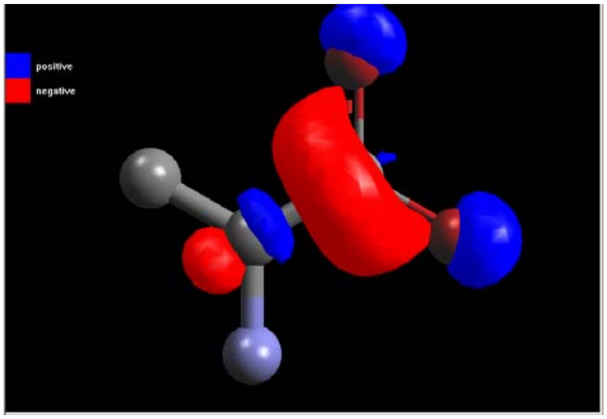


Diagram 6.33: Quick Plot HOMO of Analine

Table 6.13 : Calculating Molecular Orbitals on grids for plotting HOMO of Analine

```
*********  Validated Experiment & Chemical System Settings  **********


  Calculation started:  Thu Jun 04 15:34:36 2009

  Max. SCF cycles            200
  SCF convergence            1.5936e-009 au. for energy

  AM1 param file             C:\Program Files\ArgusLab\params\am1.prm
  SCF saved every            1000 cycles

  Two-electron integrals
    buffer size              1000
    storage                  random list in core

  Property integrals         one center
  Dipole integrals           length operator


  Input Atomic Information
************************

      1   C   16.904500  -7.535300   0.000000
      2   O   16.904500  -6.205300   0.000000
      3   C   15.752700  -8.200300   0.000000
      4   N   15.752700  -9.530300   0.000000
      5   C   14.600800  -7.535300   0.000000
      6   O   18.056300  -8.200300   0.000000
 Plotting the following orbitals to grid files:14


         Constructing Chemical System(s)

         Basis Set
        ***********

         basis functions : 24
         shells          : 12
         primitives      : 72


         Memory for Main Chemical System
         Max. number 2-ele. ints. = 1596

         Memory Requirements (bytes)
        *****************************

         Core              973208
         Scratch            10368
        System charge       0.000000


        Total number of 2-ele integrals 870

        Integrals elapsed time 0 sec.


***** SCF *****

        Core repulsion    79.3611 au

        Calculating one electron matrix
        Diagonalizing starting one-ele. matrix
        Performing SCF


         Cycle        Energy (au)           Difference
        **********************************************
            1       -36.971063
            2       -32.276391119               4.69467
            3       -33.620242051              -1.34385
            4       -26.283142152                7.3371
            5       -33.505498559              -7.22236
            6       -26.230341192               7.27516
            7       -33.507013214              -7.27667
            8       -26.229486478               7.27753
            9       -33.508071400              -7.27858
           10       -26.229663085               7.27841
           11       -33.508488479              -7.27883
           12       -26.229742535               7.27875
           13       -33.508637619               -7.2789
           14       -26.229770710               7.27887
           15       -33.508689823              -7.27892
           16       -26.229780460               7.27891
           17       -33.508708012              -7.27893
           18       -26.229783833               7.27892
           19       -33.508714344              -7.27893
           20       -26.229785002               7.27893
           21       -33.508716547              -7.27893
           22       -26.229785408               7.27893
           23       -33.508717315              -7.27893
           24       -26.229785549               7.27893
           25       -33.508717582              -7.27893
           26       -26.229785599               7.27893
           27       -33.508717675              -7.27893
           28       -26.229785616               7.27893
           29       -33.508717707              -7.27893
           30       -26.229785622               7.27893
           31       -33.508717718              -7.27893
           32       -26.229785624               7.27893
           33       -33.508717722              -7.27893
           34       -26.229785624               7.27893
           35       -33.508717723              -7.27893
           36       -26.229785625               7.27893
           37       -33.508717724              -7.27893
           38       -26.229785625               7.27893
           39       -33.508717724              -7.27893
           40       -26.229785625               7.27893
           41       -33.508717724              -7.27893
           42       -26.229785625               7.27893
           43       -33.508717724              -7.27893
           44       -26.229785625               7.27893
           45       -33.508717724              -7.27893
           46       -26.229785625               7.27893
           47       -33.508717724              -7.27893
           48       -26.229785625               7.27893
           49       -33.508717724              -7.27893
           50       -26.229785625               7.27893
           51       -33.508717724              -7.27893
           52       -26.229785625               7.27893
           53       -33.508717724              -7.27893
           54       -26.229785625               7.27893
           55       -33.508717724              -7.27893
           56       -26.229785625               7.27893
           57       -33.508717724              -7.27893
           58       -26.229785625               7.27893
           59       -33.508717724              -7.27893
           60       -26.229785625               7.27893
           61       -33.508717724              -7.27893
```

```
   62        -26.229785625              7.27893
   63        -33.508717724             -7.27893
   64        -26.229785625              7.27893
   65        -33.508717724             -7.27893
   66        -26.229785625              7.27893
   67        -33.508717724             -7.27893
   68        -26.229785625              7.27893
   69        -33.508717724             -7.27893
   70        -26.229785625              7.27893
   71        -33.508717724             -7.27893
   72        -26.229785625              7.27893
   73        -33.508717724             -7.27893
   74        -26.229785625              7.27893
   75        -33.508717724             -7.27893
   76        -26.229785625              7.27893
   77        -33.508717724             -7.27893
   78        -26.229785625              7.27893
   79        -33.508717724             -7.27893
   80        -26.229785625              7.27893
   81        -33.508717724             -7.27893
   82        -26.229785625              7.27893
   83        -33.508717724             -7.27893
   84        -26.229785625              7.27893
   85        -33.508717724             -7.27893
   86        -26.229785625              7.27893
   87        -33.508717724             -7.27893
   88        -26.229785625              7.27893
   89        -33.508717724             -7.27893
   90        -26.229785625              7.27893
   91        -33.508717724             -7.27893
   92        -26.229785625              7.27893
   93        -33.508717724             -7.27893
   94        -26.229785625              7.27893
   95        -33.508717724             -7.27893
   96        -26.229785625              7.27893
   97        -33.508717724             -7.27893
   98        -26.229785625              7.27893
   99        -33.508717724             -7.27893
  100        -26.229785625              7.27893
  101        -33.508717724             -7.27893
  102        -26.229785625              7.27893
  103        -33.508717724             -7.27893
  104        -26.229785625              7.27893
  105        -33.508717724             -7.27893
  106        -26.229785625              7.27893
  107        -33.508717724             -7.27893
  108        -26.229785625              7.27893
  109        -33.508717724             -7.27893
  110        -26.229785625              7.27893
  111        -33.508717724             -7.27893
  112        -26.229785625              7.27893
  113        -33.508717724             -7.27893
  114        -26.229785625              7.27893
  115        -33.508717724             -7.27893
  116        -26.229785625              7.27893
  117        -33.508717724             -7.27893
  118        -26.229785625              7.27893
  119        -33.508717724             -7.27893
  120        -26.229785625              7.27893
  121        -33.508717724             -7.27893
  122        -26.229785625              7.27893
  123        -33.508717724             -7.27893
  124        -26.229785625              7.27893
  125        -33.508717724             -7.27893
  126        -26.229785625              7.27893
  127        -33.508717724             -7.27893
  128        -26.229785625              7.27893
  129        -33.508717724             -7.27893
  130        -26.229785625              7.27893
  131        -33.508717724             -7.27893
  132        -26.229785625              7.27893
  133        -33.508717724             -7.27893
  134        -26.229785625              7.27893
  135        -33.508717724             -7.27893
  136        -26.229785625              7.27893
  137        -33.508717724             -7.27893
  138        -26.229785625              7.27893
  139        -33.508717724             -7.27893
  140        -26.229785625              7.27893
  141        -33.508717724             -7.27893
  142        -26.229785625              7.27893
```

```
143       -33.508717724            -7.27893
144       -26.229785625             7.27893
145       -33.508717724            -7.27893
146       -26.229785625             7.27893
147       -33.508717724            -7.27893
148       -26.229785625             7.27893
149       -33.508717724            -7.27893
150       -26.229785625             7.27893
151       -33.508717724            -7.27893
152       -26.229785625             7.27893
153       -33.508717724            -7.27893
154       -26.229785625             7.27893
155       -33.508717724            -7.27893
156       -26.229785625             7.27893
157       -33.508717724            -7.27893
158       -26.229785625             7.27893
159       -33.508717724            -7.27893
160       -26.229785625             7.27893
161       -33.508717724            -7.27893
162       -26.229785625             7.27893
163       -33.508717724            -7.27893
164       -26.229785625             7.27893
165       -33.508717724            -7.27893
166       -26.229785625             7.27893
167       -33.508717724            -7.27893
168       -26.229785625             7.27893
169       -33.508717724            -7.27893
170       -26.229785625             7.27893
171       -33.508717724            -7.27893
172       -26.229785625             7.27893
173       -33.508717724            -7.27893
174       -26.229785625             7.27893
175       -33.508717724            -7.27893
176       -26.229785625             7.27893
177       -33.508717724            -7.27893
178       -26.229785625             7.27893
179       -33.508717724            -7.27893
180       -26.229785625             7.27893
181       -33.508717724            -7.27893
182       -26.229785625             7.27893
183       -33.508717724            -7.27893
184       -26.229785625             7.27893
185       -33.508717724            -7.27893
186       -26.229785625             7.27893
187       -33.508717724            -7.27893
188       -26.229785625             7.27893
189       -33.508717724            -7.27893
190       -26.229785625             7.27893
191       -33.508717724            -7.27893
192       -26.229785625             7.27893
193       -33.508717724            -7.27893
194       -26.229785625             7.27893
195       -33.508717724            -7.27893
196       -26.229785625             7.27893
197       -33.508717724            -7.27893
198       -26.229785625             7.27893
199       -33.508717724            -7.27893
200       -26.229785625             7.27893
```

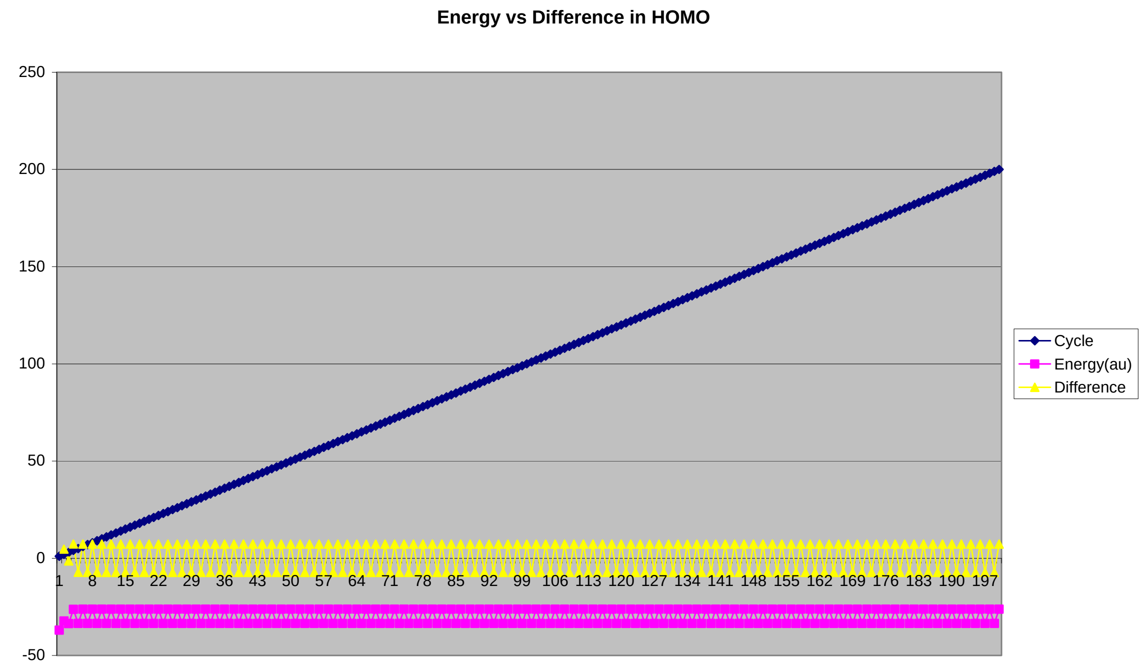


Figure 6.34: Energy vs. Difference in HOMO of Analine

```
        Maximum number of iterations reached: SCF NOT CONVERGED!


        Writing final SCF to disk

        Final SCF Energy =  -26.2297856247 au
        Final SCF Energy =  -16459.4538 kcal/mol

        Saving      the      final      SCF      to      the      restart      file
E:\myphdthesis_2009\ACd_results\Ala.restartscf


        SCF elapsed time 1 sec.


         ***** Heat of Formation *****
            11889.4613 kcal/mol


            Calculating Molecular Orbitals on grids for plotting.


         Atomic spin densities
        **********************

            1    C    0.6170
            2    O    0.0225
            3    C    0.3284
            4    N    0.0005
            5    C    0.0044
            6    O    0.0273


 S2 operator
 ***********

 exact                    0.750000
 calculated               0.750201

        Properties elapsed time 1 sec.
        Total Elapsed Time 2 sec.
```

Quick Plot LUMO: Alanine

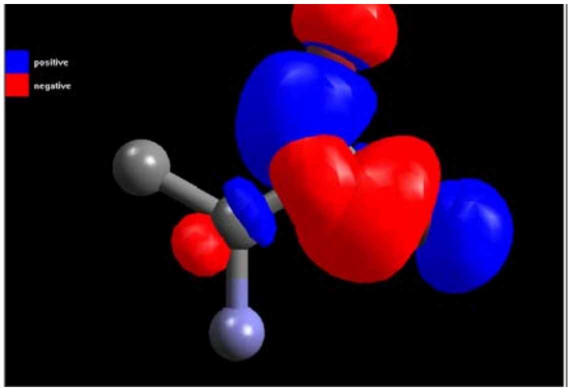


Diagram 6.35: Quick Plot LUMO of Alanine

Table 6.14 : Calculating Molecular Orbitals on grids for plotting LUMO of Alanine

```
*********  Validated Experiment & Chemical System Settings  **********


  Calculation started:  Thu Jun 04 15:53:38 2009


  Max. SCF cycles            200
  SCF convergence            1.5936e-009 au. for energy

  AM1 param file             C:\Program Files\ArgusLab\params\am1.prm
  SCF saved every            1000 cycles

    Input Atomic Information
  ************************

      1   C   16.904500  -7.535300   0.000000
      2   O   16.904500  -6.205300   0.000000
      3   C   15.752700  -8.200300   0.000000
      4   N   15.752700  -9.530300   0.000000
      5   C   14.600800  -7.535300   0.000000
      6   O   18.056300  -8.200300   0.000000


 Plotting the following orbitals to grid files:15

        Constructing Chemical System(s)

        Basis Set
       ***********

         basis functions : 24
         shells          : 12
         primitives      : 72


        Memory for Main Chemical System
        Max. number 2-ele. ints. = 1596

        Memory Requirements (bytes)
       *****************************

        Core                  973208
        Scratch                10368
        System charge          0.000000
```

```
***** SCF *****

        Core repulsion    79.3611 au

        Calculating one electron matrix
        Diagonalizing starting one-ele. matrix
        Performing SCF

         Cycle        Energy (au)            Difference
        **********************************************
           1        -36.971063
           2        -32.276391119               4.69467
           3        -33.620242051              -1.34385
           4        -26.283142152                7.3371
           5        -33.505498559              -7.22236
           6        -26.230341192               7.27516
           7        -33.507013214              -7.27667
           8        -26.229486478               7.27753
           9        -33.508071400              -7.27858
          10        -26.229663085               7.27841
          11        -33.508488479              -7.27883
          12        -26.229742535               7.27875
          13        -33.508637619               -7.2789
          14        -26.229770710               7.27887
          15        -33.508689823              -7.27892
          16        -26.229780460               7.27891
          17        -33.508708012              -7.27893
          18        -26.229783833               7.27892
          19        -33.508714344              -7.27893
          20        -26.229785002               7.27893
          21        -33.508716547              -7.27893
          22        -26.229785408               7.27893
          23        -33.508717315              -7.27893
          24        -26.229785549               7.27893
          25        -33.508717582              -7.27893
          26        -26.229785599               7.27893
          27        -33.508717675              -7.27893
          28        -26.229785616               7.27893
          29        -33.508717707              -7.27893
          30        -26.229785622               7.27893
          31        -33.508717718              -7.27893
          32        -26.229785624               7.27893
          33        -33.508717722              -7.27893
          34        -26.229785624               7.27893
          35        -33.508717723              -7.27893
          36        -26.229785625               7.27893
          37        -33.508717724              -7.27893
          38        -26.229785625               7.27893
          39        -33.508717724              -7.27893
          40        -26.229785625               7.27893
          41        -33.508717724              -7.27893
          42        -26.229785625               7.27893
          43        -33.508717724              -7.27893
          44        -26.229785625               7.27893
          45        -33.508717724              -7.27893
          46        -26.229785625               7.27893
          47        -33.508717724              -7.27893
          48        -26.229785625               7.27893
          49        -33.508717724              -7.27893
          50        -26.229785625               7.27893
          51        -33.508717724              -7.27893
          52        -26.229785625               7.27893
          53        -33.508717724              -7.27893
          54        -26.229785625               7.27893
          55        -33.508717724              -7.27893
          56        -26.229785625               7.27893
          57        -33.508717724              -7.27893
          58        -26.229785625               7.27893
          59        -33.508717724              -7.27893
          60        -26.229785625               7.27893
          61        -33.508717724              -7.27893
          62        -26.229785625               7.27893
          63        -33.508717724              -7.27893
          64        -26.229785625               7.27893
          65        -33.508717724              -7.27893
          66        -26.229785625               7.27893
          67        -33.508717724              -7.27893
          68        -26.229785625               7.27893
          69        -33.508717724              -7.27893
          70        -26.229785625               7.27893
          71        -33.508717724              -7.27893
```

```
     72      -26.229785625          7.27893
     73      -33.508717724         -7.27893
     74      -26.229785625          7.27893
     75      -33.508717724         -7.27893
     76      -26.229785625          7.27893
     77      -33.508717724         -7.27893
     78      -26.229785625          7.27893
     79      -33.508717724         -7.27893
     80      -26.229785625          7.27893
     81      -33.508717724         -7.27893
     82      -26.229785625          7.27893
     83      -33.508717724         -7.27893
     84      -26.229785625          7.27893
     85      -33.508717724         -7.27893
     86      -26.229785625          7.27893
     87      -33.508717724         -7.27893
     88      -26.229785625          7.27893
     89      -33.508717724         -7.27893
     90      -26.229785625          7.27893
     91      -33.508717724         -7.27893
     92      -26.229785625          7.27893
     93      -33.508717724         -7.27893
     94      -26.229785625          7.27893
     95      -33.508717724         -7.27893
     96      -26.229785625          7.27893
     97      -33.508717724         -7.27893
     98      -26.229785625          7.27893
     99      -33.508717724         -7.27893
    100      -26.229785625          7.27893
    101      -33.508717724         -7.27893
    102      -26.229785625          7.27893
    103      -33.508717724         -7.27893
    104      -26.229785625          7.27893
    105      -33.508717724         -7.27893
    106      -26.229785625          7.27893
    107      -33.508717724         -7.27893
    108      -26.229785625          7.27893
    109      -33.508717724         -7.27893
    110      -26.229785625          7.27893
    111      -33.508717724         -7.27893
    112      -26.229785625          7.27893
    113      -33.508717724         -7.27893
    114      -26.229785625          7.27893
    115      -33.508717724         -7.27893
    116      -26.229785625          7.27893
    117      -33.508717724         -7.27893
    118      -26.229785625          7.27893
    119      -33.508717724         -7.27893
    120      -26.229785625          7.27893
    121      -33.508717724         -7.27893
    122      -26.229785625          7.27893
    123      -33.508717724         -7.27893
    124      -26.229785625          7.27893
    125      -33.508717724         -7.27893
    126      -26.229785625          7.27893
    127      -33.508717724         -7.27893
    128      -26.229785625          7.27893
    129      -33.508717724         -7.27893
    130      -26.229785625          7.27893
    131      -33.508717724         -7.27893
    132      -26.229785625          7.27893
    133      -33.508717724         -7.27893
    134      -26.229785625          7.27893
    135      -33.508717724         -7.27893
    136      -26.229785625          7.27893
    137      -33.508717724         -7.27893
    138      -26.229785625          7.27893
    139      -33.508717724         -7.27893
    140      -26.229785625          7.27893
    141      -33.508717724         -7.27893
    142      -26.229785625          7.27893
    143      -33.508717724         -7.27893
    144      -26.229785625          7.27893
    145      -33.508717724         -7.27893
    146      -26.229785625          7.27893
    147      -33.508717724         -7.27893
    148      -26.229785625          7.27893
    149      -33.508717724         -7.27893
    150      -26.229785625          7.27893
    151      -33.508717724         -7.27893
    152      -26.229785625          7.27893
```

153	-33.508717724	-7.27893
154	-26.229785625	7.27893
155	-33.508717724	-7.27893
156	-26.229785625	7.27893
157	-33.508717724	-7.27893
158	-26.229785625	7.27893
159	-33.508717724	-7.27893
160	-26.229785625	7.27893
161	-33.508717724	-7.27893
162	-26.229785625	7.27893
163	-33.508717724	-7.27893
164	-26.229785625	7.27893
165	-33.508717724	-7.27893
166	-26.229785625	7.27893
167	-33.508717724	-7.27893
168	-26.229785625	7.27893
169	-33.508717724	-7.27893
170	-26.229785625	7.27893
171	-33.508717724	-7.27893
172	-26.229785625	7.27893
173	-33.508717724	-7.27893
174	-26.229785625	7.27893
175	-33.508717724	-7.27893
176	-26.229785625	7.27893
177	-33.508717724	-7.27893
178	-26.229785625	7.27893
179	-33.508717724	-7.27893
180	-26.229785625	7.27893
181	-33.508717724	-7.27893
182	-26.229785625	7.27893
183	-33.508717724	-7.27893
184	-26.229785625	7.27893
185	-33.508717724	-7.27893
186	-26.229785625	7.27893
187	-33.508717724	-7.27893
188	-26.229785625	7.27893
189	-33.508717724	-7.27893
190	-26.229785625	7.27893
191	-33.508717724	-7.27893
192	-26.229785625	7.27893
193	-33.508717724	-7.27893
194	-26.229785625	7.27893
195	-33.508717724	-7.27893
196	-26.229785625	7.27893
197	-33.508717724	-7.27893
198	-26.229785625	7.27893
199	-33.508717724	-7.27893
200	-26.229785625	7.27893

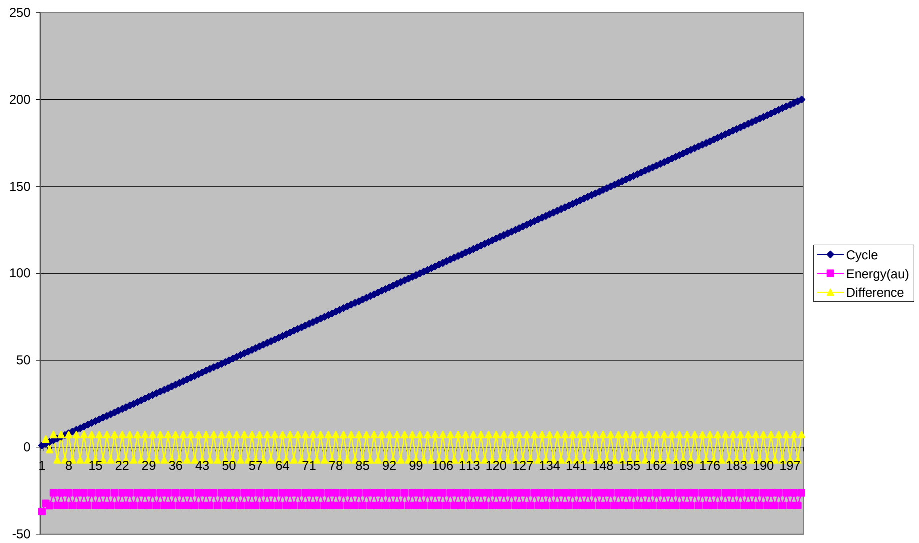


Figure 6.36: Energy vs. Difference by LUMO of Analine

```
       Maximum number of iterations reached: SCF NOT CONVERGED!


       Writing final SCF to disk

       Final SCF Energy =  -26.2297856247 au
       Final SCF Energy =  -16459.4538 kcal/mol

       Saving       the       final       SCF       to       the       restart       file
E:\myphdthesis_2009\ACd_results\Ala.restartscf


       SCF elapsed time 0 sec.

        ***** Heat of Formation *****
           11889.4613 kcal/mol


           Calculating Molecular Orbitals on grids for plotting.

        Atomic spin densities
       **********************

           1    C    0.6170
           2    O    0.0225
           3    C    0.3284
           4    N    0.0005
           5    C    0.0044
           6    O    0.0273


 S2 operator
 ***********

 exact                 0.750000
 calculated            0.750201

       Properties elapsed time 1 sec.
       Total Elapsed Time 1 sec.
```

Quick Plot ESP Mapped Density : Analine

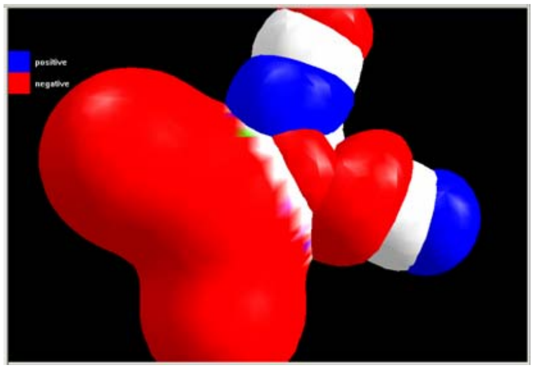


Figure 6.37: Quick Plot ESP Mapped Density of Analine

Table 6.15 : Calculating Molecular Orbitals on grids for plotting ESP Mapped Density of Analine

```
*********  Validated Experiment & Chemical System Settings  **********


   Max. SCF cycles          200
  SCF convergence          1.5936e-009 au. for energy
```

```
   Two-electron integrals
    buffer size             1000
  Input Atomic Information
  ************************

      1   C   16.904500  -7.535300   0.000000
      2   O   16.904500  -6.205300   0.000000
      3   C   15.752700  -8.200300   0.000000
      4   N   15.752700  -9.530300   0.000000
      5   C   14.600800  -7.535300   0.000000
      6   O   18.056300  -8.200300   0.000000



 Plotting electron density for the following states to grid files:0
 Plotting electrostatic potential for the following states to grid files:0


        Constructing Chemical System(s)


        Basis Set
       ***********

         basis functions : 24
         shells          : 12
         primitives      : 72


        Memory for Main Chemical System
        Max. number 2-ele. ints. = 1596

        Memory Requirements (bytes)
       *****************************

        Core              973208
        Scratch           10368
        System charge     0.000000


***** SCF *****
       Core repulsion    79.3611 au

       Calculating one electron matrix
       Diagonalizing starting one-ele. matrix
       Performing SCF

        Cycle        Energy (au)            Difference
       **********************************************
           1       -36.971063
           2       -32.276391119               4.69467
           3       -33.620242051              -1.34385
           4       -26.283142152                7.3371
           5       -33.505498559              -7.22236
           6       -26.230341192               7.27516
           7       -33.507013214              -7.27667
           8       -26.229486478               7.27753
           9       -33.508071400              -7.27858
          10       -26.229663085               7.27841
          11       -33.508488479              -7.27883
          12       -26.229742535               7.27875
          13       -33.508637619               -7.2789
          14       -26.229770710               7.27887
          15       -33.508689823              -7.27892
          16       -26.229780460               7.27891
          17       -33.508708012              -7.27893
          18       -26.229783833               7.27892
          19       -33.508714344              -7.27893
          20       -26.229785002               7.27893
          21       -33.508716547              -7.27893
          22       -26.229785408               7.27893
          23       -33.508717315              -7.27893
          24       -26.229785549               7.27893
          25       -33.508717582              -7.27893
          26       -26.229785599               7.27893
          27       -33.508717675              -7.27893
          28       -26.229785616               7.27893
          29       -33.508717707              -7.27893
          30       -26.229785622               7.27893
          31       -33.508717718              -7.27893
          32       -26.229785624               7.27893
          33        -33.508717722             -7.27893
          34        -26.229785624              7.27893
          35        -33.508717723             -7.27893
          36        -26.229785625              7.27893
          37        -33.508717724             -7.27893
          38        -26.229785625              7.27893
          39        -33.508717724             -7.27893
          40        -26.229785625              7.27893
          41        -33.508717724             -7.27893
          42        -26.229785625              7.27893
          43        -33.508717724             -7.27893
          44        -26.229785625              7.27893
          45        -33.508717724             -7.27893
          46        -26.229785625              7.27893
          47        -33.508717724             -7.27893
          48        -26.229785625              7.27893
          49        -33.508717724             -7.27893
          50        -26.229785625              7.27893
          51        -33.508717724             -7.27893
          52        -26.229785625              7.27893
          53        -33.508717724             -7.27893
          54        -26.229785625              7.27893
          55        -33.508717724             -7.27893
          56        -26.229785625              7.27893
          57        -33.508717724             -7.27893
          58        -26.229785625              7.27893
          59        -33.508717724             -7.27893
          60        -26.229785625              7.27893
          61        -33.508717724             -7.27893
          62        -26.229785625              7.27893
          63        -33.508717724             -7.27893
          64        -26.229785625              7.27893
          65        -33.508717724             -7.27893
          66        -26.229785625              7.27893
          67        -33.508717724             -7.27893
          68        -26.229785625              7.27893
          69        -33.508717724             -7.27893
          70        -26.229785625              7.27893
          71        -33.508717724             -7.27893
          72        -26.229785625              7.27893
          73        -33.508717724             -7.27893
          74        -26.229785625              7.27893
          75        -33.508717724             -7.27893
          76        -26.229785625              7.27893
          77        -33.508717724             -7.27893
          78        -26.229785625              7.27893
          79        -33.508717724             -7.27893
          80        -26.229785625              7.27893
          81        -33.508717724             -7.27893
          82        -26.229785625              7.27893
          83        -33.508717724             -7.27893
          84        -26.229785625              7.27893
          85        -33.508717724             -7.27893
          86        -26.229785625              7.27893
          87        -33.508717724             -7.27893
          88        -26.229785625              7.27893
          89        -33.508717724             -7.27893
          90        -26.229785625              7.27893
          91        -33.508717724             -7.27893
          92        -26.229785625              7.27893
          93        -33.508717724             -7.27893
          94        -26.229785625              7.27893
          95        -33.508717724             -7.27893
          96        -26.229785625              7.27893
          97        -33.508717724             -7.27893
          98        -26.229785625              7.27893
          99        -33.508717724             -7.27893
         100        -26.229785625              7.27893
         101        -33.508717724             -7.27893
         102        -26.229785625              7.27893
         103        -33.508717724             -7.27893
         104        -26.229785625              7.27893
         105        -33.508717724             -7.27893
         106        -26.229785625              7.27893
         107        -33.508717724             -7.27893
         108        -26.229785625              7.27893
         109        -33.508717724             -7.27893
         110        -26.229785625              7.27893
         111        -33.508717724             -7.27893
         112        -26.229785625              7.27893
         113        -33.508717724             -7.27893
```

```
114        -26.229785625             7.27893
115        -33.508717724            -7.27893
116        -26.229785625             7.27893
117        -33.508717724            -7.27893
118        -26.229785625             7.27893
119        -33.508717724            -7.27893
120        -26.229785625             7.27893
121        -33.508717724            -7.27893
122        -26.229785625             7.27893
123        -33.508717724            -7.27893
124        -26.229785625             7.27893
125        -33.508717724            -7.27893
126        -26.229785625             7.27893
127        -33.508717724            -7.27893
128        -26.229785625             7.27893
129        -33.508717724            -7.27893
130        -26.229785625             7.27893
131        -33.508717724            -7.27893
132        -26.229785625             7.27893
133        -33.508717724            -7.27893
134        -26.229785625             7.27893
135        -33.508717724            -7.27893
136        -26.229785625             7.27893
137        -33.508717724            -7.27893
138        -26.229785625             7.27893
139        -33.508717724            -7.27893
140        -26.229785625             7.27893
141        -33.508717724            -7.27893
142        -26.229785625             7.27893
143        -33.508717724            -7.27893
144        -26.229785625             7.27893
145        -33.508717724            -7.27893
146        -26.229785625             7.27893
147        -33.508717724            -7.27893
148        -26.229785625             7.27893
149        -33.508717724            -7.27893
150        -26.229785625             7.27893
151        -33.508717724            -7.27893
152        -26.229785625             7.27893
153        -33.508717724            -7.27893
154        -26.229785625             7.27893
155        -33.508717724            -7.27893
156        -26.229785625             7.27893
157        -33.508717724            -7.27893
158        -26.229785625             7.27893
159        -33.508717724            -7.27893
160        -26.229785625             7.27893
161        -33.508717724            -7.27893
162        -26.229785625             7.27893
163        -33.508717724            -7.27893
164        -26.229785625             7.27893
165        -33.508717724            -7.27893
166        -26.229785625             7.27893
167        -33.508717724            -7.27893
168        -26.229785625             7.27893
169        -33.508717724            -7.27893
170        -26.229785625             7.27893
171        -33.508717724            -7.27893
172        -26.229785625             7.27893
173        -33.508717724            -7.27893
174        -26.229785625             7.27893
175        -33.508717724            -7.27893
176        -26.229785625             7.27893
177        -33.508717724            -7.27893
178        -26.229785625             7.27893
179        -33.508717724            -7.27893
180        -26.229785625             7.27893
181        -33.508717724            -7.27893
182        -26.229785625             7.27893
183        -33.508717724            -7.27893
184        -26.229785625             7.27893
185        -33.508717724            -7.27893
186        -26.229785625             7.27893
187        -33.508717724            -7.27893
188        -26.229785625             7.27893
189        -33.508717724            -7.27893
190        -26.229785625             7.27893
191        -33.508717724            -7.27893
192        -26.229785625             7.27893
193        -33.508717724            -7.27893
194        -26.229785625             7.27893
```

```
            195        -33.508717724              -7.27893
            196        -26.229785625               7.27893
            197        -33.508717724              -7.27893
            198        -26.229785625               7.27893
            199        -33.508717724              -7.27893
            200        -26.229785625               7.27893
```

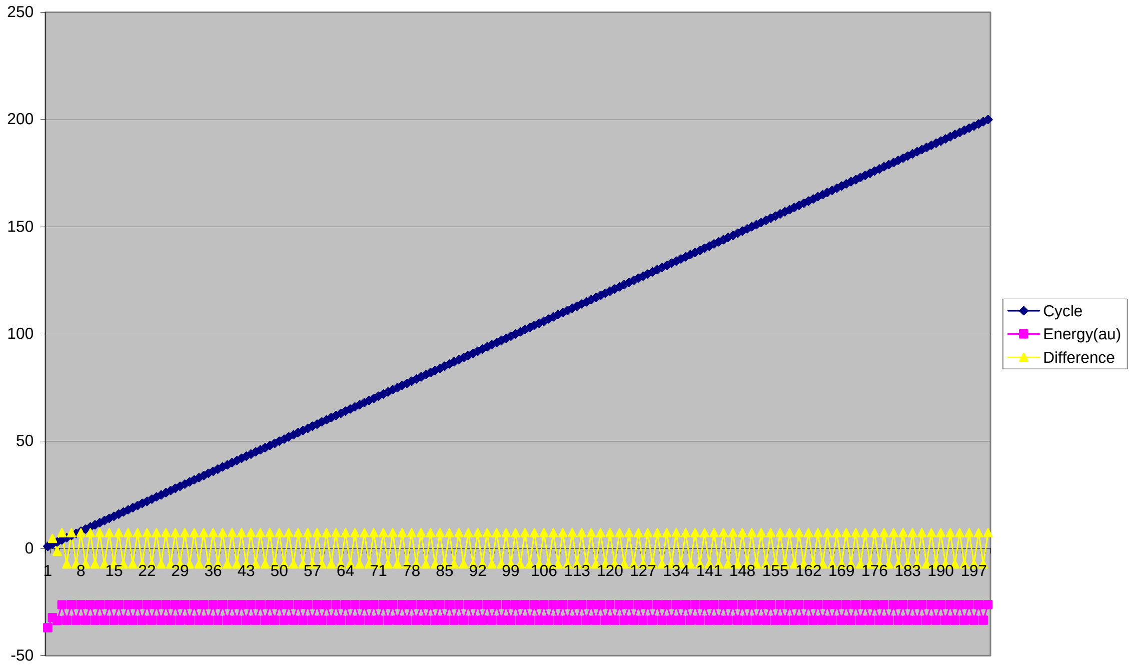


Figure 6.38: Energy vs. Difference /cycle by ESP Mapped Density of Analine

```
          Final SCF Energy =  -26.2297856247 au
          Final SCF Energy =  -16459.4538 kcal/mol

          SCF elapsed time 1 sec.


           ***** Heat of Formation *****
              11889.4613 kcal/mol

              Calculating ground-state density on grid.

              Calculating ground-state electrostatic potential on grid.

           Atomic spin densities
          **********************
              1    C    0.6170
              2    O    0.0225
              3    C    0.3284
              4    N    0.0005
              5    C    0.0044
              6    O    0.0273

    S2 operator
    ***********
    exact                   0.750000
    calculated              0.750201

          Properties elapsed time 17 sec.
          Total Elapsed Time 18 sec.
```

6.2.2.1 Data Analysis of Experimental Work

In this research work like Alanine other compounds amino butyric acid, Asparagine and Glutamine have been analyzed on ArgusLab tool and calculations like Single Entry Point Calculation and Geometric Optimization have been performed.

(A) Amino butyric acid

Single Entry Point Calculation: Amino butyric acid

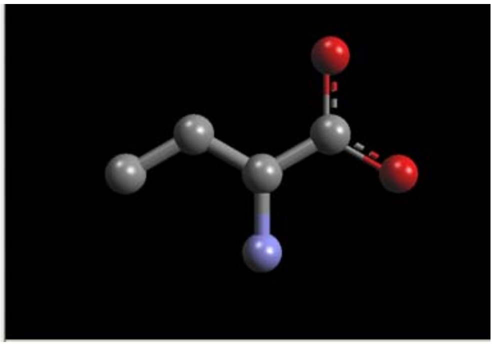


Figure 6.39: Structure of Amino butyric acid

Table 6.16: Single Entry Point Calculation of Amino butyric acid

```
*********  Validated Experiment & Chemical System Settings  **********


  Input Atomic Information
  ************************

      1   C   16.727200  -9.538300   0.000000
      2   O   16.727200  -8.208300   0.000000
      3   C   15.575500 -10.203300   0.000000
      4   N   15.575500 -11.533300   0.000000
      5   C   14.423600  -9.538300   0.000000
      6   O   17.879100 -10.203300   0.000000
      7   C   13.271800 -10.203400   0.000000


         Constructing Chemical System(s)

         Basis Set
        ***********

          basis functions : 28
          shells          : 14
          primitives      : 84


         Memory for Main Chemical System
         Max. number 2-ele. ints. = 2212


         Memory Requirements (bytes)
        *****************************

         Core               154592
         Scratch             13888
         System charge       0.000000
```

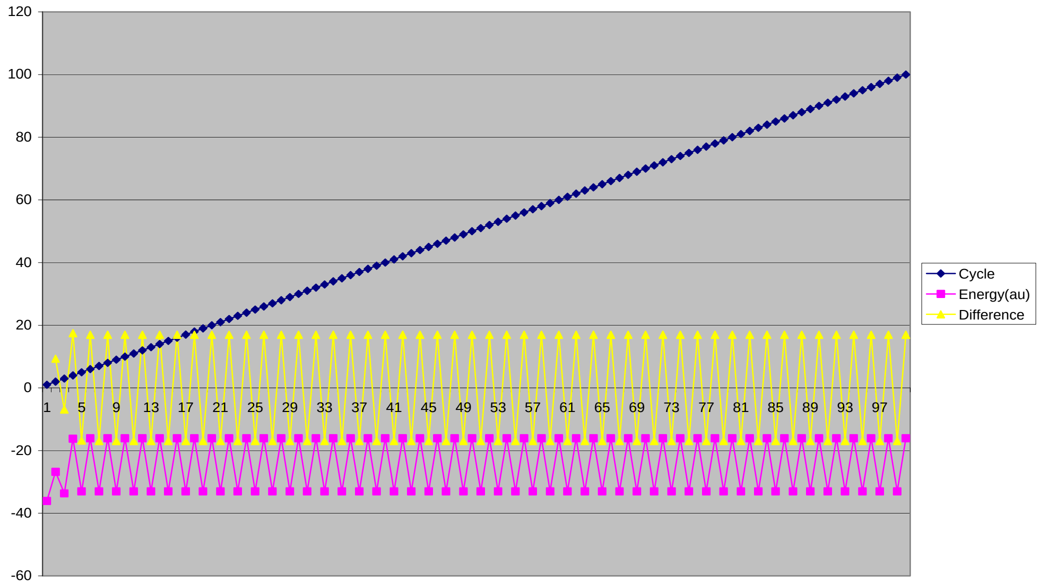


Figure 6.40: Energy vs. Difference /cycle by SEC of Amino butyric acid

```
        ***** SCF *****

        Writing final SCF to disk

        Final SCF Energy =  -16.0505499405 au
        Final SCF Energy =  -10071.8812 kcal/mol

               SCF elapsed time 1 sec.


  ***** Heat of Formation *****
            21233.9419 kcal/mol

Wiberg Atom-Atom Bond Orders
*****************************

       1           2           3           4           5           6

  1   0.000000
  2   0.159277    0.000000
  3   0.593365    0.132542    0.000000
  4   0.001599    0.002860    0.003560    0.000000
  5   0.000804    0.005289    0.002427    0.000046    0.000000
  6   0.052133    0.006354    0.140044    0.004958    0.000753    0.000000
  7   0.000011    0.000003    0.000014    0.000000    0.000000    0.000006


         Atomic spin densities
        **********************

            1    C    0.0616
            2    O    0.7115
            3    C    0.1654
            4    N    0.0033
            5    C    0.0068
            6    O    0.0514
            7    C    0.0000
S2 operator
 ***********

 exact                     0.750000
 calculated                0.751744

                           Ground State Dipole (debye)

                        X             Y             Z           length
            179.55334737   64.31200335   -0.00000000  190.72346034
         Mulliken Atomic Charges
        ************************

            1    C    3.7788
            2    O    4.7938
            3    C   -3.4233
            4    N   -2.9980
            5    C   -4.0014
            6    O    5.8506
            7    C   -4.0005

        Properties elapsed time 0 sec.

        Total Elapsed Time 1 sec.
```

Geometry Optimization: Amino butyric acid

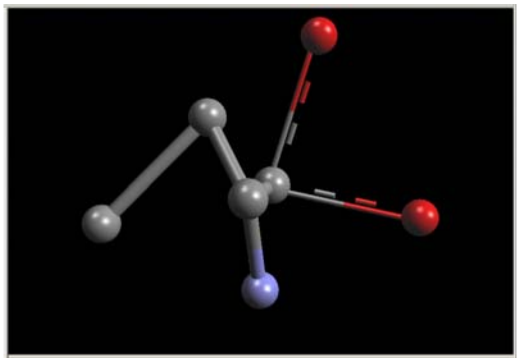


Figure 6.41: Geometry Optimization of Amino butyric acid

Table 6.9 : Geometry Optimization calculation of Amino butyric acid

```
*******  Validated Experiment & Chemical System Settings  **********


  Calculation started:  Thu Jun 04 16:50:46 2009


  Title:E:\myphdthesis_2009\ACd_results\Abu
  Max. geom cycles          100
  Convergence criteria:
  max. grad. component <    0.000084   au.

  Input Atomic Information
  ************************

      1   C   16.727200  -9.538300   0.000000
      2   O   16.727200  -8.208300   0.000000
      3   C   15.575500 -10.203300   0.000000
      4   N   15.575500 -11.533300   0.000000
      5   C   14.423600  -9.538300   0.000000
      6   O   17.879100 -10.203300   0.000000
      7   C   13.271800 -10.203400   0.000000

 Constructing Chemical System(s)


         Basis Set
        ***********

          basis functions : 28
          shells          : 14
          primitives      : 42
         Memory for Main Chemical System
         Max. number 2-ele. ints. = 2212

         Memory Requirements (bytes)
        *****************************
         Core                 440544
         Scratch              13888
         System charge        0.000000

Total number of 2-ele integrals 1232

        Integrals elapsed time 0 sec.
```

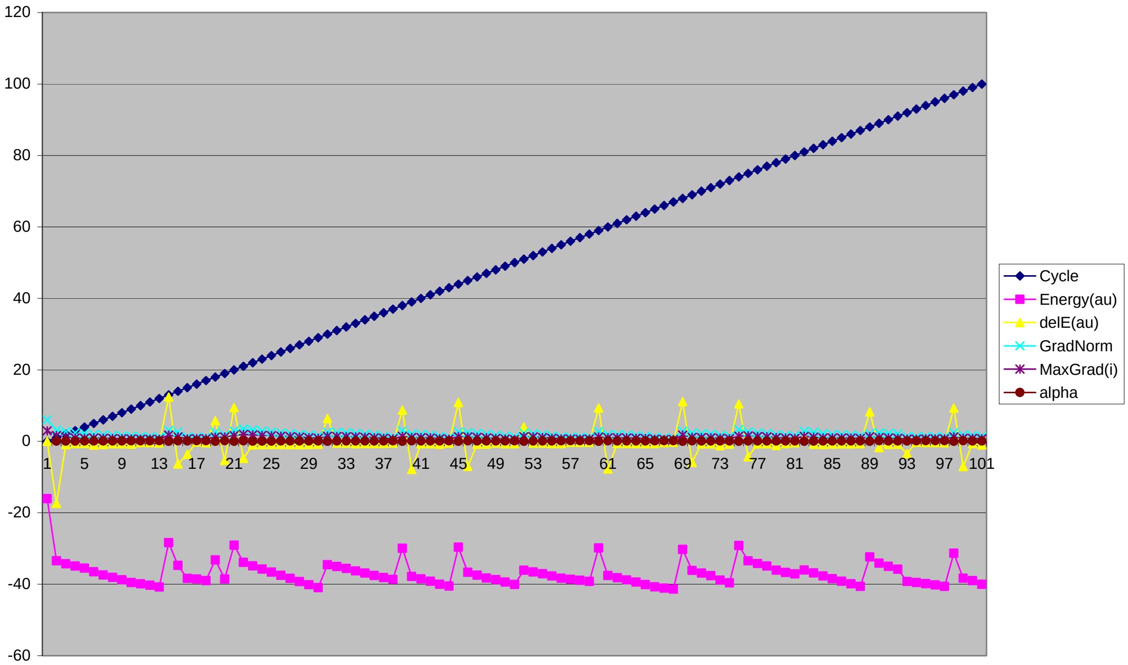


Figure 6.42: Geometry Optimization Energy for various components of Amino butyric acid

```
***** SCF *****

        Core repulsion    99.4321 au
        Final SCF Energy =  -16.0321304196 au
        Final SCF Energy =  -10060.3228 kcal/mol

***************  Final Geometry ****************

          C   15.94062971    -9.99263718   -0.48994553      6
          O   16.76806159    -7.34951479    0.07850090      8
          C   15.52407585   -10.21360278    0.64044515      6
          N   15.72470082   -11.86193116   -0.07280711      7
          C   14.77876094    -8.77341644   -0.10464275      6
          O   18.56438623   -10.55937950    0.03787747      8
          C   12.87928488   -10.67771816   -0.08942813      6


        Final Geom Energy =  -39.9881072545 au
        Final Geom Energy =  -25092.9388 kcal/mol

        Geometry Optimization elapsed time 4 min. 53 sec.

***** Heat of Formation *****

            10076.7315 kcal/mol
```

```
       Atomic spin densities
      **********************

          1    C    1.0005
          2    O    0.0005
          3    C   -0.0225
          4    N    0.0189
          5    C    0.0019
          6    O    0.0008
          7    C   -0.0000

S2 operator
***********

exact                  0.750000
calculated             0.790950

      Properties elapsed time 0 sec.

      Total Elapsed Time 4 min. 54 sec.
```

Quick Plot HOMO: Amino butyric acid

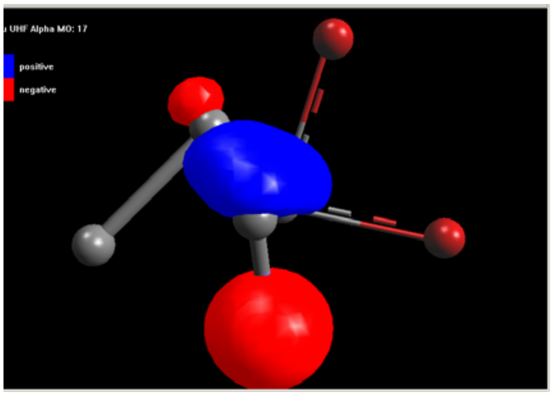


Figure 6.43: Quick Plot HOMO of Amino butyric acid

Table 6.17: Quick Plot HOMO calculation of Amino butyric acid

```
 Calculation started:  Thu Jun 04 17:29:12 2009

 Max. SCF cycles           200
 SCF convergence           1.5936e-009 au. for energy

 Input Atomic Information
 ************************

     1   C   15.940630  -9.992637  -0.489946
     2   O   16.768062  -7.349515   0.078501
     3   C   15.524076 -10.213603   0.640445
     4   N   15.724701 -11.861931  -0.072807
     5   C   14.778761  -8.773416  -0.104643
     6   O   18.564386 -10.559379   0.037877
     7   C   12.879285 -10.677718  -0.089428

Plotting the following orbitals to grid files:16
```

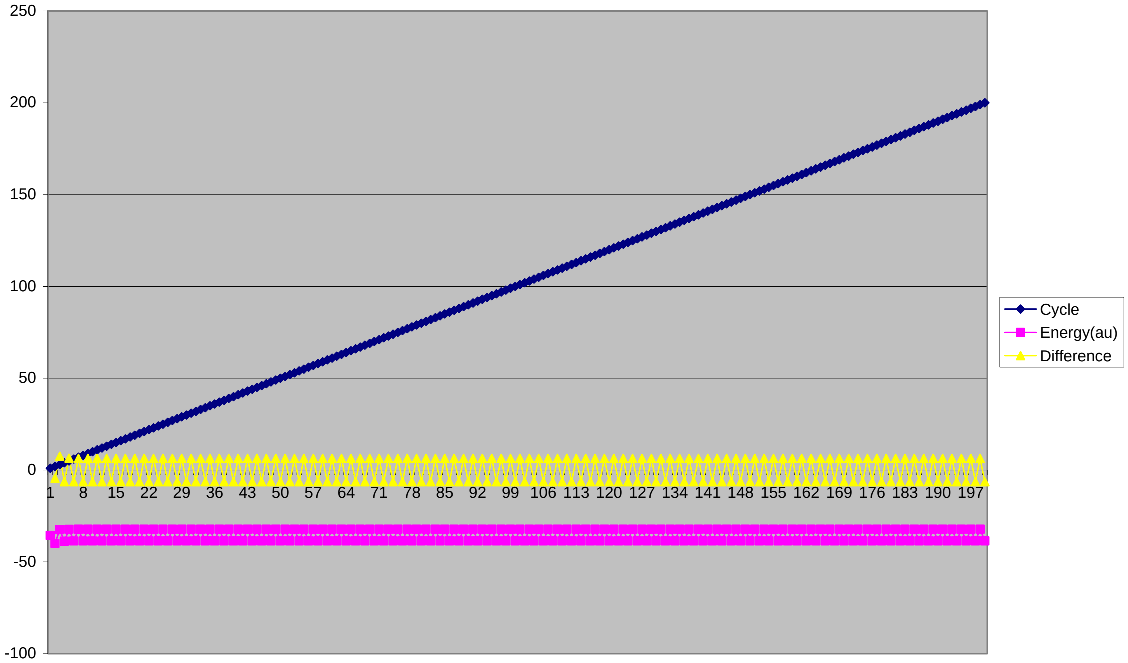


Figure 6.44: Energy vs. Difference /cycle Quick Plot HOMO calculation of Amino butyric acid

```
Constructing Chemical System(s)
        Basis Set
       ***********

         basis functions : 28
         shells          : 14
         primitives      : 84


        Memory for Main Chemical System
        Max. number 2-ele. ints. = 2212

        Memory Requirements (bytes)
       *****************************

        Core             1002304
        Scratch            13888
        System charge       0.000000

***** SCF *****

       Core repulsion    79.7563 au
       Final SCF Energy =  -38.5635015264 au
       Final SCF Energy =  -24198.9844 kcal/mol

***** Heat of Formation *****
             7106.8388 kcal/mol

            Calculating Molecular Orbitals on grids for plotting.

        Atomic spin densities
       **********************

            1    C   -0.0184
            2    O    0.0000
            3    C    0.0197
            4    N    0.0012
            5    C    0.9972
            6    O   -0.0000
            7    C    0.0002
```

```
 S2 operator
 ***********

 exact                    0.750000
 calculated               0.751080

        Properties elapsed time 1 sec.

        Total Elapsed Time 2 sec.
```

Quick Plot LUMO : Amino butyric acid

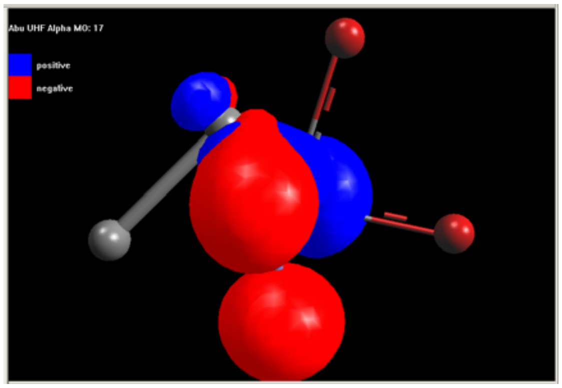


Figure 6.45: Quick Plot LUMO of Amino butyric acid

Table 6.18 : Quick Plot LUMO calculation of Amino butyric acid

```
 ********  Validated Experiment & Chemical System Settings  **********


   Calculation started:  Thu Jun 04 17:50:51 2009
   Max. SCF cycles           200
   SCF convergence           1.5936e-009 au. for energy

   Input Atomic Information
   ************************

       1   C   15.940630  -9.992637  -0.489946
       2   O   16.768062  -7.349515   0.078501
       3   C   15.524076 -10.213603   0.640445
       4   N   15.724701 -11.861931  -0.072807
       5   C   14.778761  -8.773416  -0.104643
       6   O   18.564386 -10.559379   0.037877
       7   C   12.879285 -10.677718  -0.089428

 Constructing Chemical System(s)


          Basis Set
         ***********

          basis functions : 28
          shells          : 14
          primitives      : 84


          Memory for Main Chemical System
          Max. number 2-ele. ints. = 2212

          Memory Requirements (bytes)
         *****************************

          Core               1002304
          Scratch              13888
          System charge        0.000000
```

```
               Total number of 2-ele integrals 2212

       ***** SCF *****


               Core repulsion    79.7563 au
                Final SCF Energy =  -38.5635015264 au
               Final SCF Energy  =  -24198.9844 kcal/mol


               SCF elapsed time 1 sec.


                ***** Heat of Formation *****
                    7106.8388 kcal/mol
```

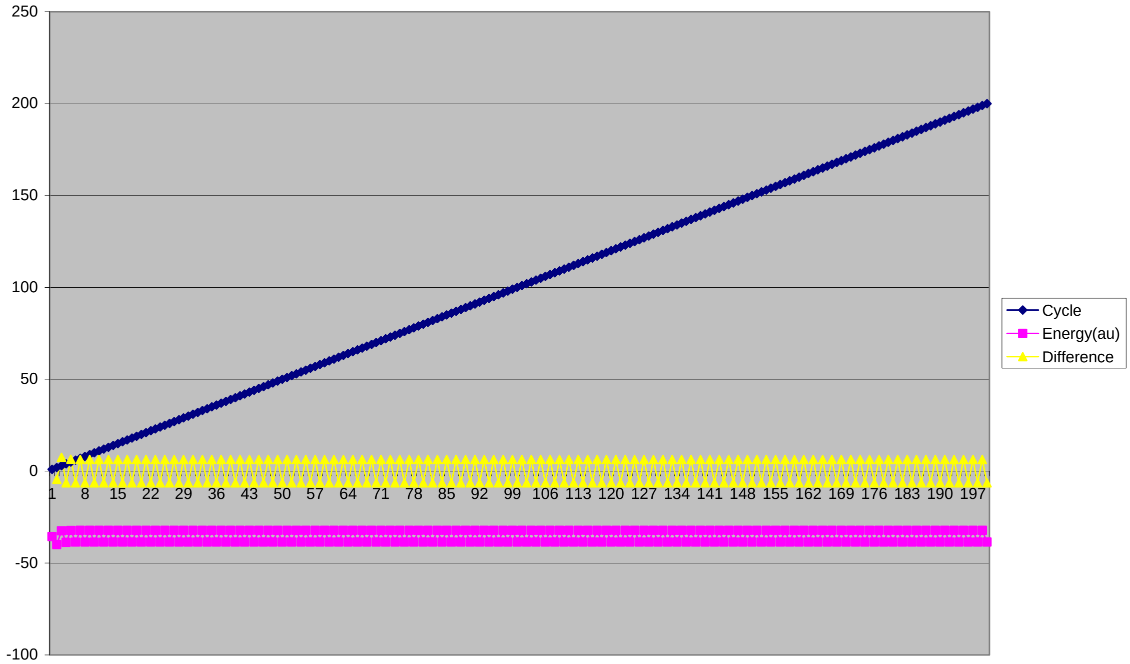


Figure 6.46: Energy vs. Difference/ cycle by Quick Plot LUMO of Amino butyric acid

```
                Calculating Molecular Orbitals on grids for plotting.
                Atomic spin densities
               **********************

                   1    C   -0.0184
                   2    O    0.0000
                   3    C    0.0197
                   4    N    0.0012
                   5    C    0.9972
                   6    O   -0.0000
                   7    C    0.0002


               S2 operator
               ***********

               exact                  0.750000
               calculated             0.751080

               Properties elapsed time 1 sec.
```

Quick Plot ESP Mapped Density: Amino butyric acid

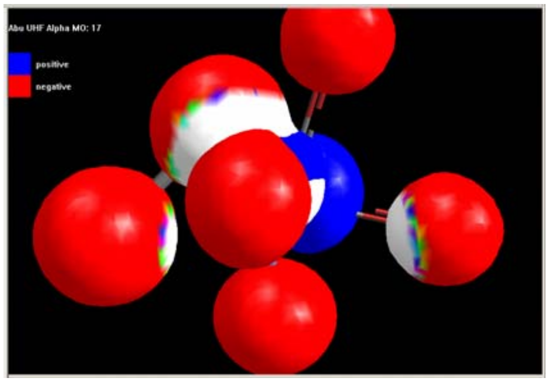


Figure 6.47: Quick Plot ESP Mapped Density of Amino butyric acid

Table 6.19: Quick Plot ESP Mapped Density calculation of Amino butyric acid

```
*********  Validated Experiment & Chemical System Settings  **********


  Calculation started:  Thu Jun 04 18:40:01 2009


  Title:
  SCF convergence           1.5936e-009 au. for energy

    Input Atomic Information
  ************************

      1   C   15.940630  -9.992637  -0.489946
      2   O   16.768062  -7.349515   0.078501
      3   C   15.524076 -10.213603   0.640445
      4   N   15.724701 -11.861931  -0.072807
      5   C   14.778761  -8.773416  -0.104643
      6   O   18.564386 -10.559379   0.037877
      7   C   12.879285 -10.677718  -0.089428


        Constructing Chemical System(s)


        Basis Set
       ***********

         basis functions : 28
         shells          : 14
         primitives      : 84


        Memory for Main Chemical System
        Max. number 2-ele. ints. = 2212

        Memory Requirements (bytes)
       *****************************

        Core                  1002304
        Scratch               13888
        System charge         0.000000


       Total number of 2-ele integrals 2212

       Integrals elapsed time 0 sec.
```

```
***** SCF *****

       Core repulsion    79.7563 au

       Calculating one electron matrix
       Diagonalizing starting one-ele. matrix
       Performing SCF
       Final SCF Energy =  -38.5635015264 au
       Final SCF Energy =  -24198.9844 kcal/mol

       SCF elapsed time 1 sec.
```

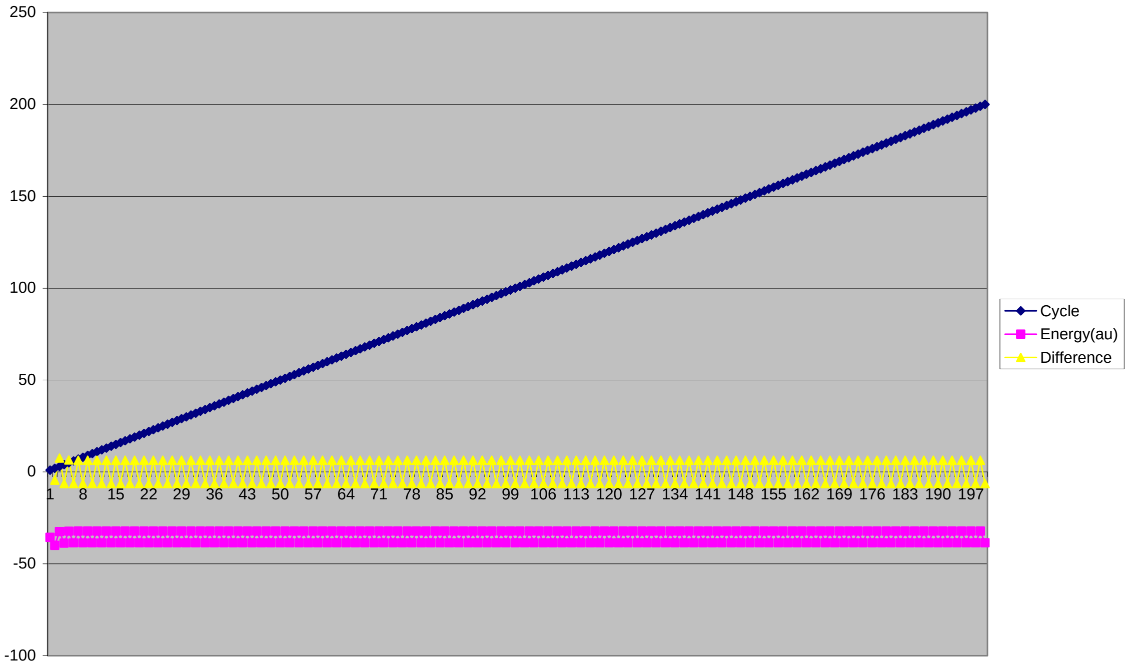


Figure 6.48: Energy vs. Difference / cycle by Quick Plot ESP Mapped Density of Amino butyric acid

```
            ***** Heat of Formation *****
            7106.8388 kcal/mol


        Atomic spin densities
       **********************

           1    C   -0.0184
           2    O    0.0000
           3    C    0.0197
           4    N    0.0012
           5    C    0.9972
           6    O   -0.0000
           7    C    0.0002


 S2 operator
 ***********

 exact                 0.750000
 calculated            0.751080


       Properties elapsed time 19 sec.

       Total Elapsed Time 20 sec.
```

Asparagine

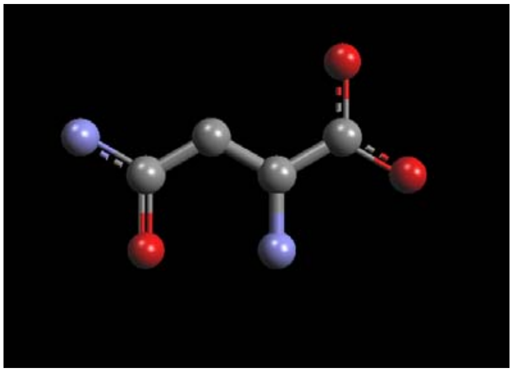


Figure 6.49: Structure of Asparagine

Single Entry Point Calculation: Asparagine

Table 6.20: Single Entry Point Calculation of Asparagine

```
Input Atomic Information
************************

      1   C   21.569300  -8.224400   0.000000
      2   C   23.873000  -8.224400   0.000000
      3   C   20.417600  -8.889400   0.000000
      4   C   22.721200  -8.889400   0.000000
      5   O   23.873000  -6.894400   0.000000
      6   O   20.417600 -10.219400   0.000000
      7   N   19.265800  -8.224400   0.000000
      8   N   22.721200 -10.219400   0.000000
      9   O   25.024800  -8.889400   0.000000

Constructing Chemical System(s)


Basis Set
***********

         basis functions : 36
         shells          : 18
         primitives      : 108


        Memory for Main Chemical System
        Max. number 2-ele. ints. = 3744

***** SCF *****

       Core repulsion    165.108 au
       Final SCF Energy =  -24.2129529102 au
       Final SCF Energy =     -15193.8710 kcal/mol

 ***** Heat of Formation *****

           28241.3138 kcal/mol
```

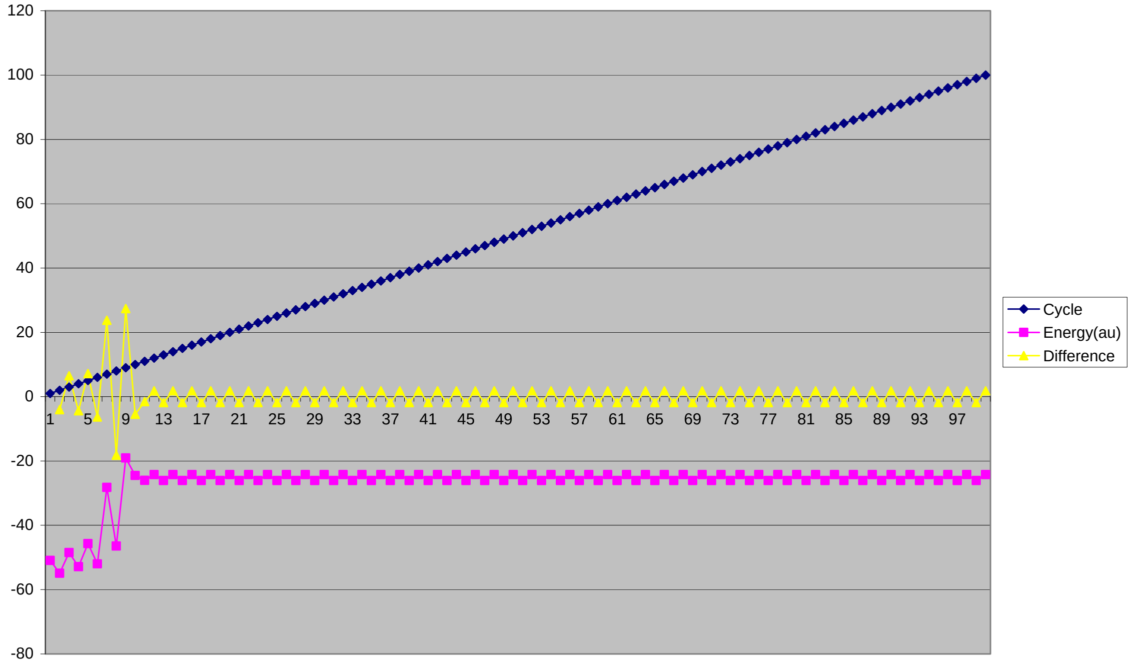


Figure 6.50: Energy vs. Difference of Single Entry Point Calculation of Asparagine

```
Properties elapsed time 0 sec.

Total Elapsed Time 1 sec.
```

Geometry Optimization: Asparagine

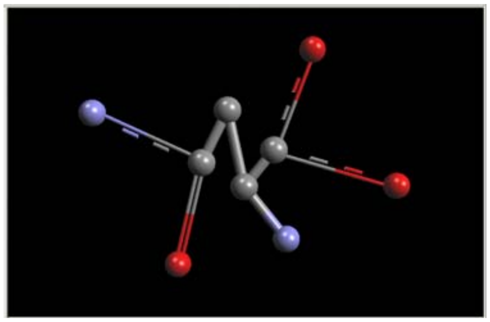


Figure 6.51: Geometry Optimization of Asparagine

Table 6.21 : Geometry Optimization calculation of Asparagine

```
Max. SCF cycles             50
SCF convergence             1.5936e-013 au. for energy

Max. geom cycles            100
Convergence criteria:
max. grad. component <      0.000084   au.
```

```
Input Atomic Information
************************

      1   C   21.569300  -8.224400   0.000000
      2   C   23.873000  -8.224400   0.000000
      3   C   20.417600  -8.889400   0.000000
      4   C   22.721200  -8.889400   0.000000
      5   O   23.873000  -6.894400   0.000000
      6   O   20.417600 -10.219400   0.000000
      7   N   19.265800  -8.224400   0.000000
      8   N   22.721200 -10.219400   0.000000
      9   O   25.024800  -8.889400   0.000000

Constructing Chemical System(s)


Basis Set
***********

         basis functions : 36
         shells          : 18
         primitives      : 54


        Memory for Main Chemical System
        Max. number 2-ele. ints. = 3744

Memory Requirements (bytes)
*****************************

        Core              727504
        Scratch            22464
***** SCF *****


       Core repulsion    165.108 au
       Final SCF Energy =  -24.2078555875 au
       Final SCF Energy =  -15190.6724 kcal/mol

       Final Geom Energy =  -49.2464605506 au
       Final Geom Energy =  -30902.6484 kcal/mol
       Geometry Optimization elapsed time 10 min. 41 sec.

***** Heat of Formation *****
                   15290.0969 kcal/mol

Total Elapsed Time 10 min. 42 sec.
```

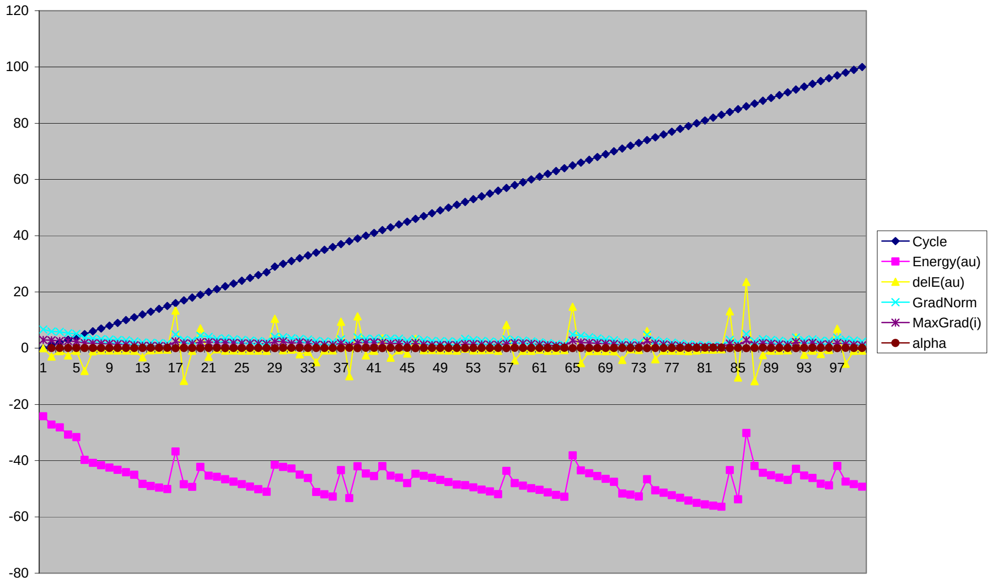


Figure 6.52: Geometry Optimization Energy for various components of Asparagine

QuickPlot HOMO: Asparagine

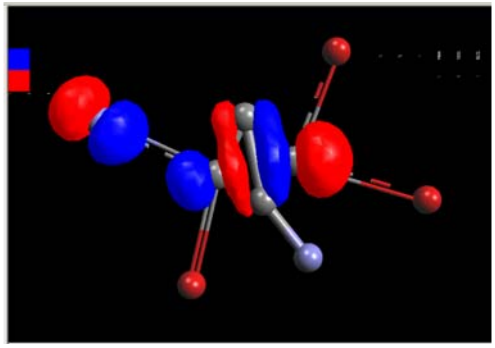


Figure 6.53: QuickPlot HOMO of Asparagine

Table 6.22: QuickPlot HOMO of Asparagine

```
*********  Validated Experiment & Chemical System Settings  **********


        Max. SCF cycles           200
        SCF convergence           1.5936e-009 au. for energy


Input Atomic Information
************************

              1   C   21.754545  -7.442867   0.000000
              2   C   22.902106  -8.435586  -0.000002
              3   C   21.117400  -8.750163  -0.000003
              4   C   22.167307  -9.347993   0.000005
              5   O   23.854067  -5.931479   0.000000
              6   O   20.535689 -11.277524   0.000000
              7   N   18.418534  -7.512176   0.000000
              8   N   23.206285 -10.652366  -0.000000
              9   O   25.927568  -9.324447   0.000000

         Plotting the following orbitals to grid files:21

        Constructing Chemical System(s)

Basis Set
***********

                 basis functions : 36
                 shells          : 18
                 primitives      : 108


                Memory for Main Chemical System
                Max. number 2-ele. ints. = 3744

Memory Requirements (bytes)
*****************************

                Core             1051640
                Scratch            22464


***** SCF *****
               Core repulsion    133.616 au
                Final SCF Energy =  -67.3585267794 au
               Final SCF Energy  =  -42268.1518 kcal/mol
***** Heat of Formation *****
             1167.0330 kcal/mol
Total Elapsed Time 2 sec.
```

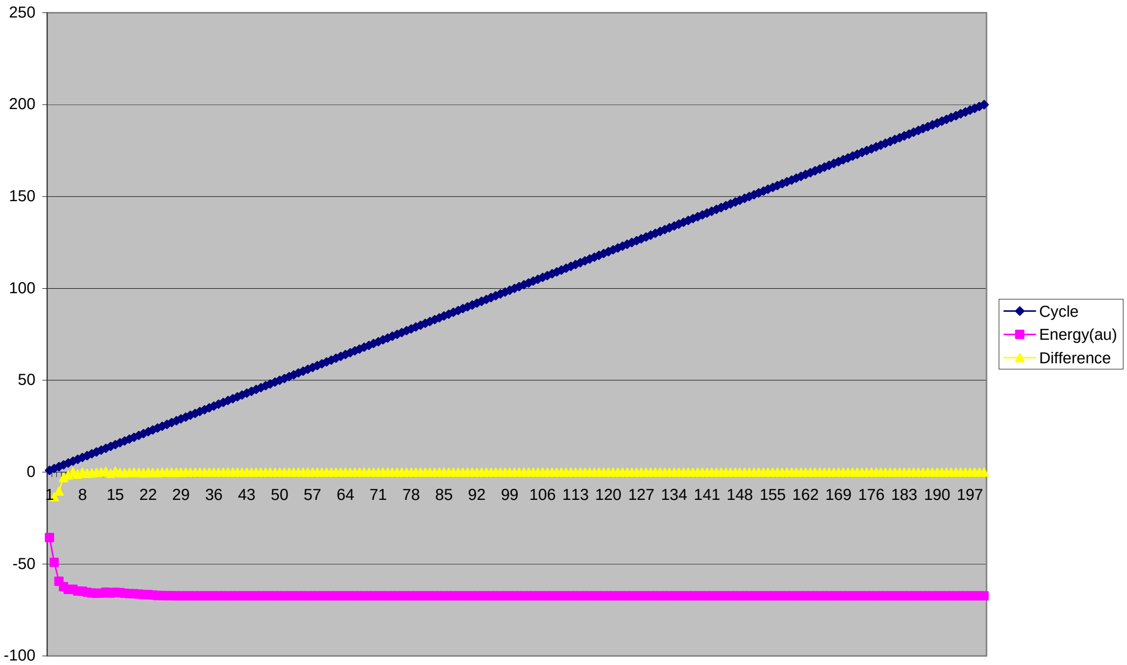


Figure 6.54 : Energy vs. Difference /cycle for QuickPlot HOMO of Asparagine

QuickPlot LUMO: Asparagine

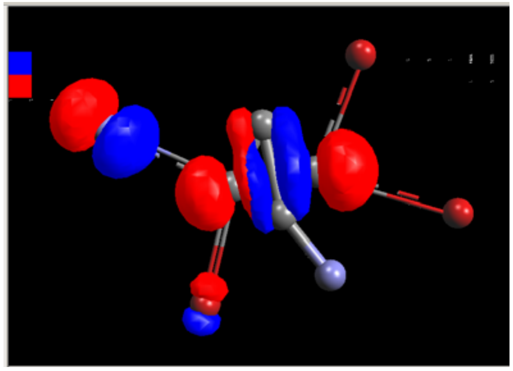


Figure 6.55: QuickPlot LUMO of Asparagine

Table 6.23: QuickPlot LUMO calculation of Asparagine

```
********  Validated Experiment & Chemical System Settings  **********

       Max. SCF cycles            200
       SCF convergence            1.5936e-009 au. for energy


Input Atomic Information
************************

             1   C   21.754545  -7.442867   0.000000
             2   C   22.902106  -8.435586  -0.000002
             3   C   21.117400  -8.750163  -0.000003
             4   C   22.167307  -9.347993   0.000005
             5   O   23.854067  -5.931479   0.000000
             6   O   20.535689 -11.277524   0.000000
```

```
              7   N   18.418534  -7.512176   0.000000
              8   N   23.206285 -10.652366  -0.000000
              9   O   25.927568  -9.324447   0.000000


                Constructing Chemical System(s)


Basis Set
***********

                basis functions : 36
                shells          : 18
                primitives      : 108


                Memory for Main Chemical System
                Max. number 2-ele. ints. = 3744

Memory Requirements (bytes)
*****************************

                Core              1051640
                Scratch             22464


***** SCF *****

               Core repulsion    133.616 au
               Final SCF Energy =   -67.3585267794 au
               Final SCF Energy =   -42268.1518 kcal/mol


***** Heat of Formation *****
                1167.0330 kcal/mol

        Total Elapsed Time 2 sec.
```

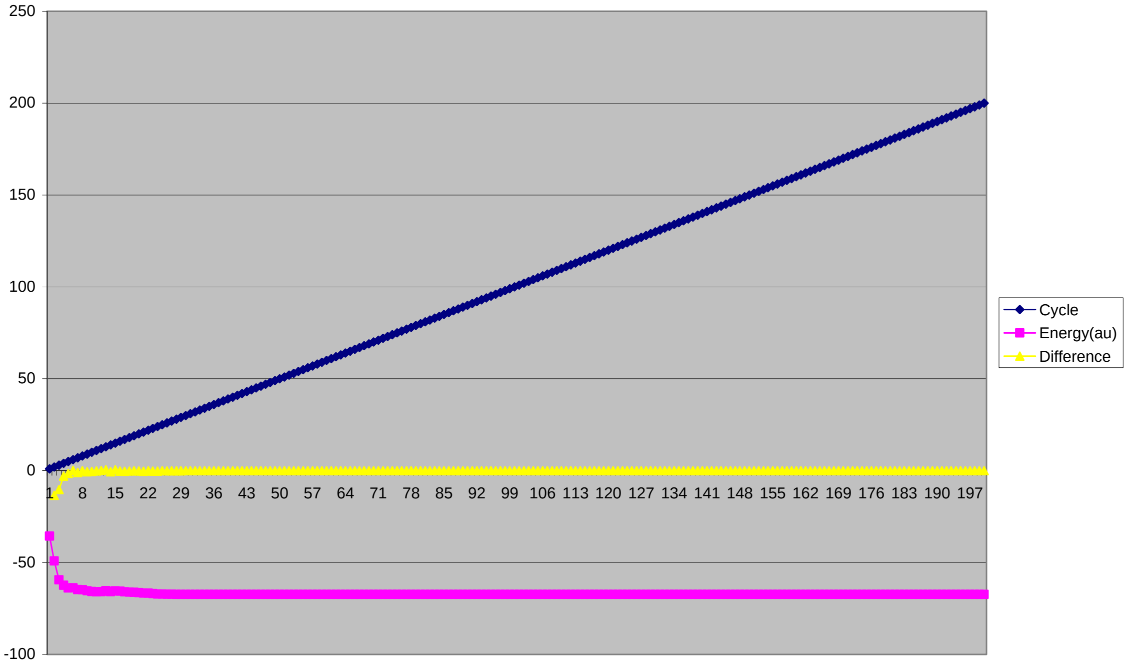


Figure 6.56: Energy vs. Difference /cycle for QuickPlot LUMO of Asparagine

Quick Plot ESP Mapped Density: Asparagine

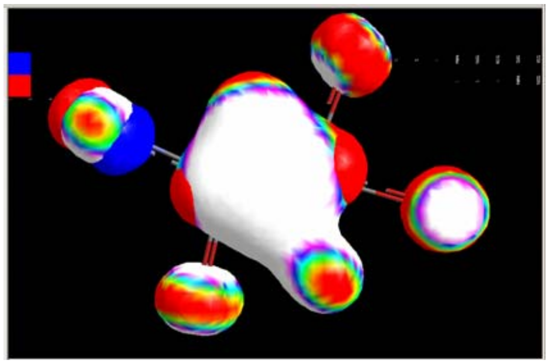


Figure 6.57: Quick Plot ESP Mapped Density of Asparagine

Table 6.24: Quick Plot ESP Mapped Density calculation of Asparagine

```
*********  Validated Experiment & Chemical System Settings  **********

Max. SCF cycles           200
SCF convergence           1.5936e-009 au. for energy


Input Atomic Information
************************

      1   C   21.754545  -7.442867   0.000000
      2   C   22.902106  -8.435586  -0.000002
      3   C   21.117400  -8.750163  -0.000003
      4   C   22.167307  -9.347993   0.000005
      5   O   23.854067  -5.931479   0.000000
      6   O   20.535689 -11.277524   0.000000
      7   N   18.418534  -7.512176   0.000000
      8   N   23.206285 -10.652366  -0.000000
      9   O   25.927568  -9.324447   0.000000


Constructing Chemical System(s)


Basis Set
***********

         basis functions : 36
         shells          : 18
         primitives      : 108

        Memory for Main Chemical System
        Max. number 2-ele. ints. = 3744

Memory Requirements (bytes)
*****************************

        Core             1051640
        Scratch            22464

Total number of 2-ele integrals 2935

***** SCF *****

       Core repulsion    133.616 au
       Final SCF Energy =  -67.3585267794 au
       Final SCF Energy =  -42268.1518 kcal/mol

***** Heat of Formation *****
            1167.0330 kcal/mol
Total Elapsed Time 26 sec.
```

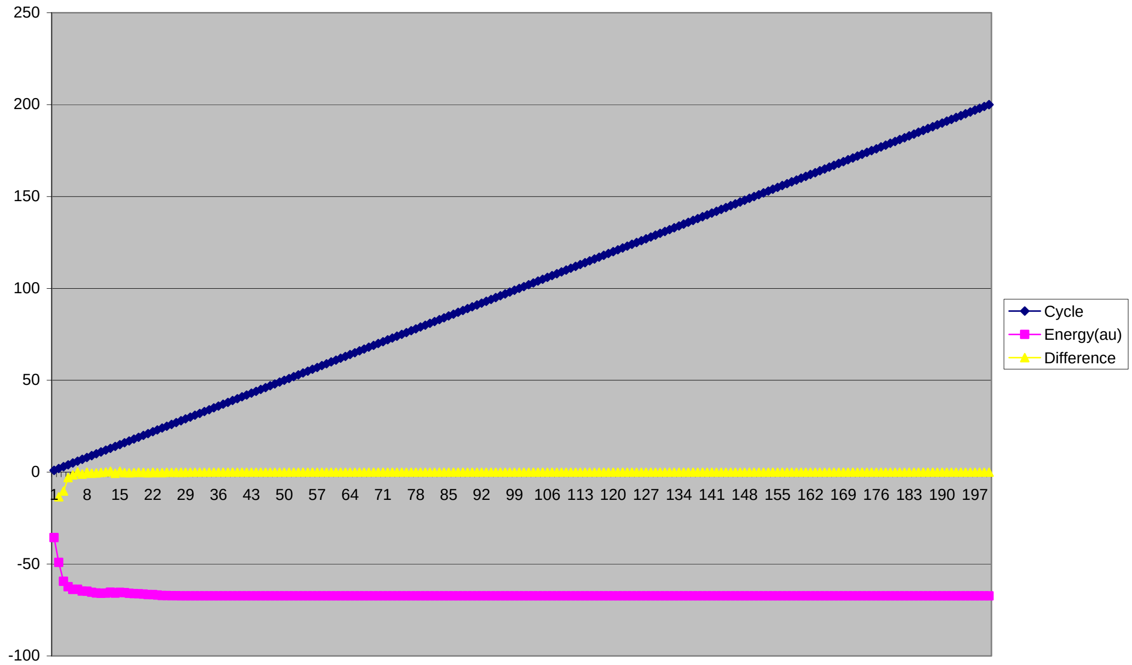


Figure 6.58: Energy vs. Difference for Quick Plot ESP Mapped Density calculation of Asparagine

Glutamine

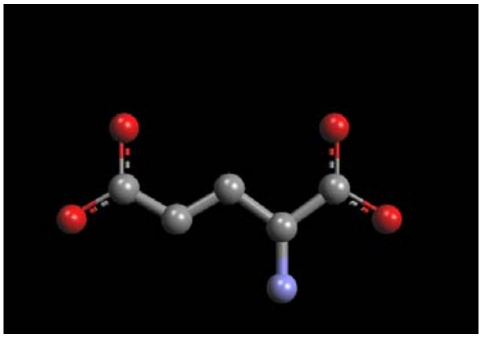


Figure 6.59: Structure of Glutamine

Single Entry Point Calculation: Glutamine

Table 6.25: Single Entry Point Calculation of Glutamine

```
*********  Validated Experiment & Chemical System Settings  **********

Max. SCF cycles            100
SCF convergence            1.5936e-013 au. for energy

Basis Set
***********

         basis functions : 40
         shells          : 20
         primitives      : 120
```

```
        Memory for Main Chemical System
        Max. number 2-ele. ints. = 4660

Memory Requirements (bytes)
*****************************

        Core              251216
        Scratch            27520

Total number of 2-ele integrals 2554

***** SCF *****

       Core repulsion    188.429 au
       Final SCF Energy =  -34.1405108556 au
       Final SCF Energy =  -21423.5133 kcal/mol

***** Heat of Formation *****
           27536.8764 kcal/mol

Atomic spin densities
**********************

          1    C    0.0000
          2    C    0.0006
          3    O    0.0000
          4    C    0.0000
          5    O    0.0006
          6    C    0.0002
          7    C    0.0379
          8    O    0.0001
          9    O    0.0000
         10    N    0.9605

 S2 operator
 ***********
 exact                    0.750000
 calculated               0.750000

       Total Elapsed Time 1 sec.
```

Geometry Optimization: Glutamine

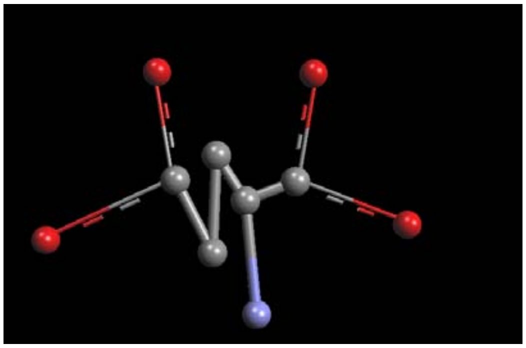


Figure 6.60: Geometry Optimization of Glutamine

Table 6.26: Geometry Optimization calculation of Glutamine

```
Constructing Chemical System(s)


Basis Set
***********

         basis functions : 40
         shells          : 20
         primitives      : 60
```

```
        Memory for Main Chemical System
        Max. number 2-ele. ints. = 4660

Memory Requirements (bytes)
*****************************

        Core              899000
        Scratch            27520


Total number of 2-ele integrals 2770

       Final SCF Energy =  -17.6840713073 au
       Final SCF Energy =  -11096.9323 kcal/mol


***************  Final Geometry ****************

                C   15.31615076  -13.95376286    0.00000000      6
                C   18.08684663  -13.98614029    0.00000000      6
                O   14.89461756  -11.46192091    0.00000000      8
                C   16.17329576  -15.62707299    0.00000000      6
                O   18.50404616  -11.50082702   -0.00000000      8
                C   16.26399478  -13.38874488    0.00000000      6
                C   16.93275916  -14.42568040   -0.00000000      6
                O   20.66644582  -14.98357984   -0.00000000      8
                O   12.32919477  -15.33565085    0.00000000      8
                N   17.19304862  -17.05401998    0.00000000      7


               Final Geom Energy =  -59.9845308085 au
               Final Geom Energy =  -37640.8953 kcal/mol


***** Heat of Formation *****
                   11319.6093 kcal/mol


               Total Elapsed Time 12 min. 16 sec.
```

QuickPlot HOMO: Glutamine

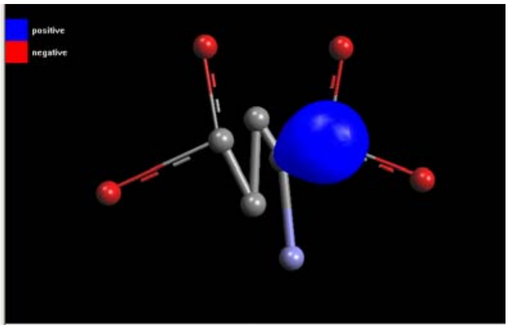


Figure 6.61: QuickPlot HOMO calculation of Glutamine

Table 6.27: QuickPlot HOMO calculation of Glutamine

```
                Constructing Chemical System(s)


Basis Set
***********

                basis functions : 40
                shells          : 20
                primitives      : 120
                Memory for Main Chemical System
                Max. number 2-ele. ints. = 4660

Memory Requirements (bytes)
*****************************

                Core             1115224
                Scratch            27520


***** SCF *****

       Core repulsion    159.976 au
       Final SCF Energy =  -59.1523636153 au
       Final SCF Energy =  -37118.7021 kcal/mol


S2 operator
***********

 exact                    0.750000
 calculated               0.750001


       Total Elapsed Time 4 sec.
```

QuickPlot LUMO: Glutamine

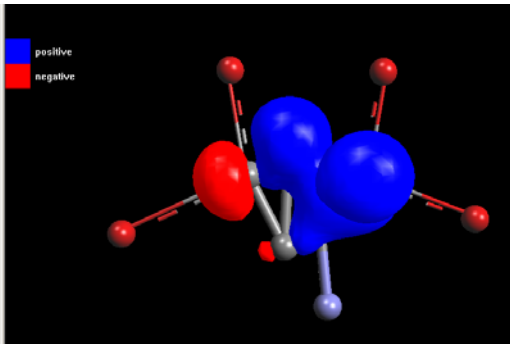


Figure 6.62: QuickPlot LUMO calculation of Glutamine

Table 6.28 QuickPlot LUMO calculation of Glutamine

```
       Input Atomic Information
       ************************

             1   C   15.316151 -13.953763   0.000000
             2   C   18.086847 -13.986140   0.000000
             3   O   14.894618 -11.461921   0.000000
             4   C   16.173296 -15.627073   0.000000
             5   O   18.504046 -11.500827  -0.000000
             6   C   16.263995 -13.388745   0.000000
             7   C   16.932759 -14.425680  -0.000000
             8   O   20.666446 -14.983580  -0.000000
             9   O   12.329195 -15.335651   0.000000
            10   N   17.193049 -17.054020   0.000000

       constructing Chemical System(s)


Basis Set
***********

                 basis functions : 40
                 shells          : 20
                 primitives      : 120
               Final SCF Energy =  -59.1523636153 au
               Final SCF Energy =  -37118.7021 kcal/mol
```

```
                Memory for Main Chemical System
                Max. number 2-ele. ints. = 4660

Memory Requirements (bytes)
*****************************
                Core             1115224
                Scratch            27520

S2 operator
***********

        exact                  0.750000
        calculated             0.750001

               Properties elapsed time 1 sec.

               Total Elapsed Time 3 sec.
```

Quick Plot ESP Mapped Density: Glutamine

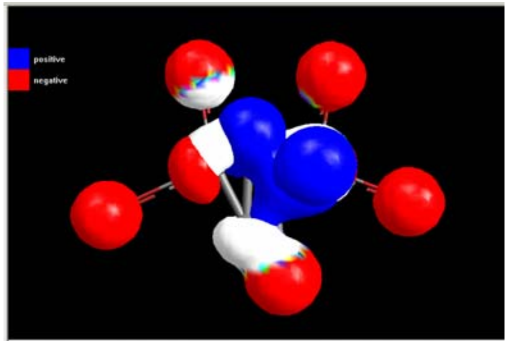


Figure 6.63: Quick Plot ESP Mapped Density of Glutamine

Table 6.29: Quick Plot ESP Mapped Density of Glutamine

```
Constructing Chemical System(s)

Basis Set
***********

         basis functions : 40
         shells          : 20
         primitives      : 120


        Memory for Main Chemical System
        Max. number 2-ele. ints. = 4660

Memory Requirements (bytes)
*****************************

        Core             1115224
        Scratch            27520

***** SCF *****

       Core repulsion    159.976 au

       Final SCF Energy =  -59.1523636153 au
       Final SCF Energy =  -37118.7021 kcal/mol

               SCF elapsed time 2 sec.

**** Heat of Formation *****
               11841.6877 kcal/mol
```

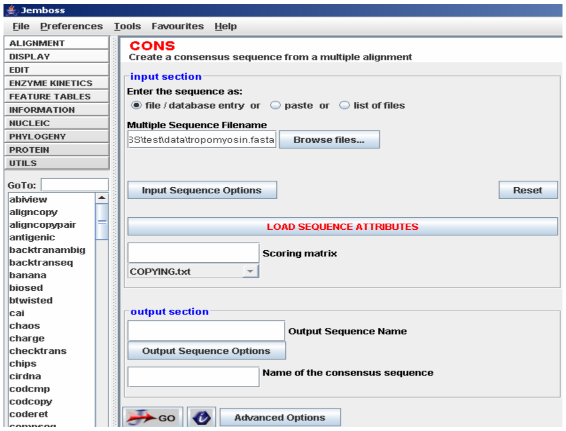
6.2.3 Activity No -3

In the second phase of research work experiment has been performed using tools like DAMBE and Jumboss. Under this research work molecular sequence of Nucleotides and Proteins has been analyzed.

6.2.3.1 Sequence Analysis Using Jemboss

Creation of Sequence from Multiple Alignments

In this phase of my research work I have created nucleotide sequence from multiple alignments using *tropomyosin.fasta* file using Jemboss software.



Jemboss

CONS

Create a consensus sequence from a multiple alignment

input section

Enter the sequence as: ● file / database entry or ○ paste or ○ list of files

Multiple Sequence Filename: 3S\test\data\tropomyosin.fasta

Scoring matrix: COPYING.txt

output section

Output Sequence Name

Name of the consensus sequence

Figure 6.64 : Creation of Nucleotide Sequence from Multiple Alignment using Jemboss

```
>EMBOSS_001
CCGGCCGCCAGCAGCACTAATGTGCTGGAGGCGCAAACTCACCATATGCTCCGGCACCCC
AAGGGTGGGGGGAGGGGGGCGCACAGGAGGCGCAGCGGCTGCAGGAAGAAGAGGGCGAG
AGGGAGGTGATGGAGGGAGGGGCGAGAGCGGCGGGCAAGCAGGAGAGCTAACGGCTGATC
ACGGCGGCGTCAGCGAATGAGAGGAGGGCTGGAACGGCCAGGTGGCGGAGCGAGGACGCG
GAAGCGGAAACTGAGAAGAAGAAGGGGAGAGGGCCGAAGAGCGTAGTAAGAGAGGCAAAG
AACAAGAA---GAAG--G--A--GAAG--G--GA-A----A--A-------AA----G--
GAGGA-G--------AG---G--G-A------------C--G-T---C--G---------
---------G--G-G-A--G-C--G-GG-GG------AA-G--GA-G-G---A-G--GA-
AA-----A----G--A--A------T-------------G-A-------A------TG-T
------G--GA----A-----GAA--A-T-----------T--A-G----A---------
--------T-CT---G-A----AA---A-----A-A----T---GA--A------AC-A-
-------A---C--CC-C--GAAC-A--AGA--C----C---C---AG-----G--TCAG
AGCC--AACTAC---ATCAGTAA-CT-C--AA---GACGAG-AT-A-C--CTCC-ACACT
C--GA-CT--GTG-TGGC--CTTCTC-T--ATGAATGA-CT--AGGTGTTGCC-CTG--C
AACG-C-CAA-C---C-G----TGG---------C--T----GAGGTGTATGAAGGCCAG
GTGTCCGGAATGCCCAACGACCCAAGCCCCCTGCAAGTGGCTGTGAAGACGCTGCCTGAA
GTGTGC
```

Figure 6.65 : Creation of Nucleotide Sequence from Multiple Alignment

Drawing a Threshold Dot Plot of two Sequences

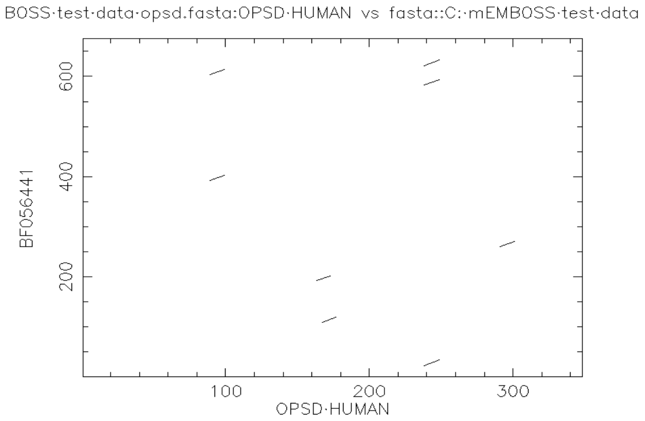
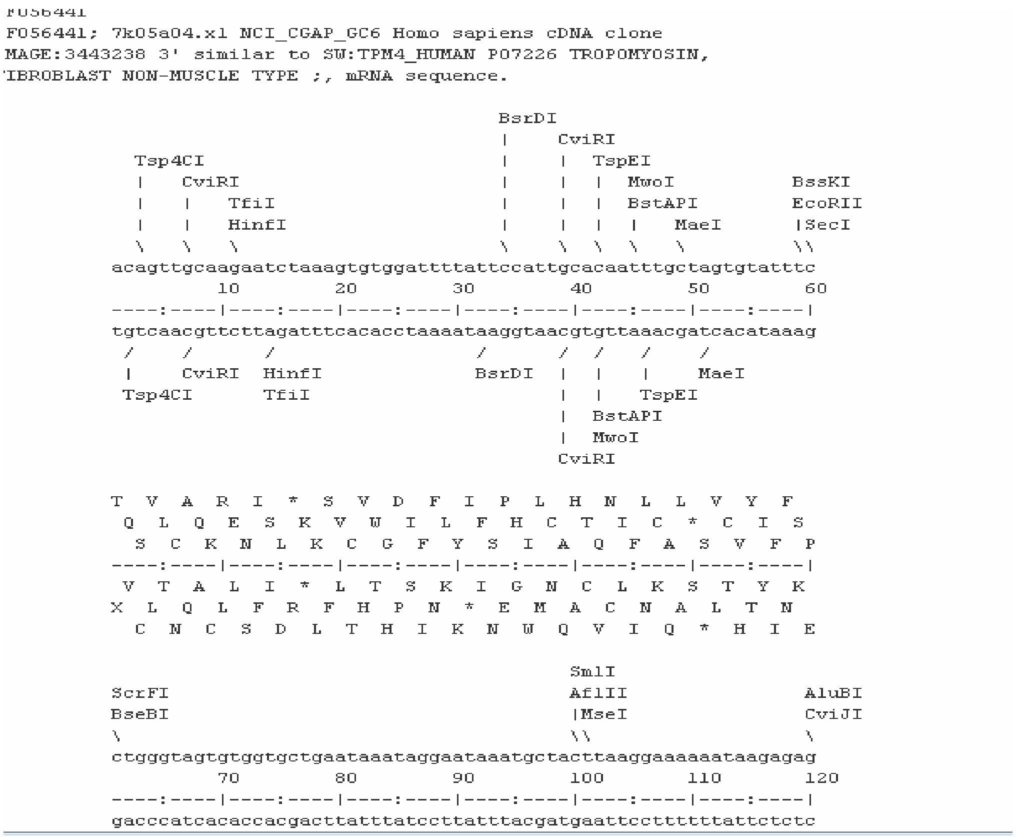


Figure 6.66 : Draw a threshold Dot Plot of two Sequences

Displaying Restriction Enzyme binding site in nucleotide sequence



```
F056441
F056441; 7k05a04.x1 NCI_CGAP_GC6 Homo sapiens cDNA clone
MAGE:3443238 3' similar to SW:TPM4_HUMAN P07226 TROPOMYOSIN,
'IBROBLAST NON-MUSCLE TYPE ;, mRNA sequence.

                                          BsrDI
                                          |    CviRI
           Tsp4CI                         |    |  TspEI
           |   CviRI                      |    |  |  MwoI          BssKI
           |   |   TfiI                   |    |  |  BstAPI        EcoRII
           |   |   HinfI                  |    |  |  |   MaeI      |SecI
           \   \   \                      \    \  \  \   \         \\
         acagttgcaagaatctaaagtgtggattttattccattgcacaatttgctagtgtatttc
                  10        20        30        40        50        60
         ----:----|----:----|----:----|----:----|----:----|----:----|
         tgtcaacgttcttagatttcacacctaaaataaggtaacgtgttaaacgatcacataaag
          /    /      /                 /      /  /   /    /
          |    CviRI  HinfI             BsrDI  |  |   |    MaeI
          Tsp4CI      TfiI                     |  |   TspEI
                                               |  BstAPI
                                               |  MwoI
                                               CviRI

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

                                                SmlI
         ScrFI                                  AflII               AluBI
         BseBI                                  |MseI               CviJI
         \                                      \\                  \
         ctgggtagtgtggtgctgaataaataggaataaatgctacttaaggaaaaaataagagag
                  70        80        90        100       110       120
         ----:----|----:----|----:----|----:----|----:----|----:----|
         gacccatcacaccacgacttatttatccttatttacgatgaattcctttttattctctc
```

Figure 6.67: Display restriction Enzyme binding site in nucleotide sequence

Calculation of Codon adaptation Index

```
Sequence: BF056441 CAI: 0.208
Sequence: BE848719 CAI: 0.223
Sequence: BF022813 CAI: 0.139
Sequence: BF452255 CAI: 0.136
Sequence: BG089808 CAI: 0.161
Sequence: BG147728 CAI: 0.128
Sequence: BI817778 CAI: 0.107
Sequence: AF186109 CAI: 0.184
Sequence: AF186110 CAI: 0.171
Sequence: AF310722 CAI: 0.176
Sequence: AF362886 CAI: 0.189
Sequence: AF362887 CAI: 0.191
Sequence: AF087679 CAI: 0.176
```

Figure 6.68 : Calculation of Codon adaptation Index

Calculation of isochores in DNA Sequence

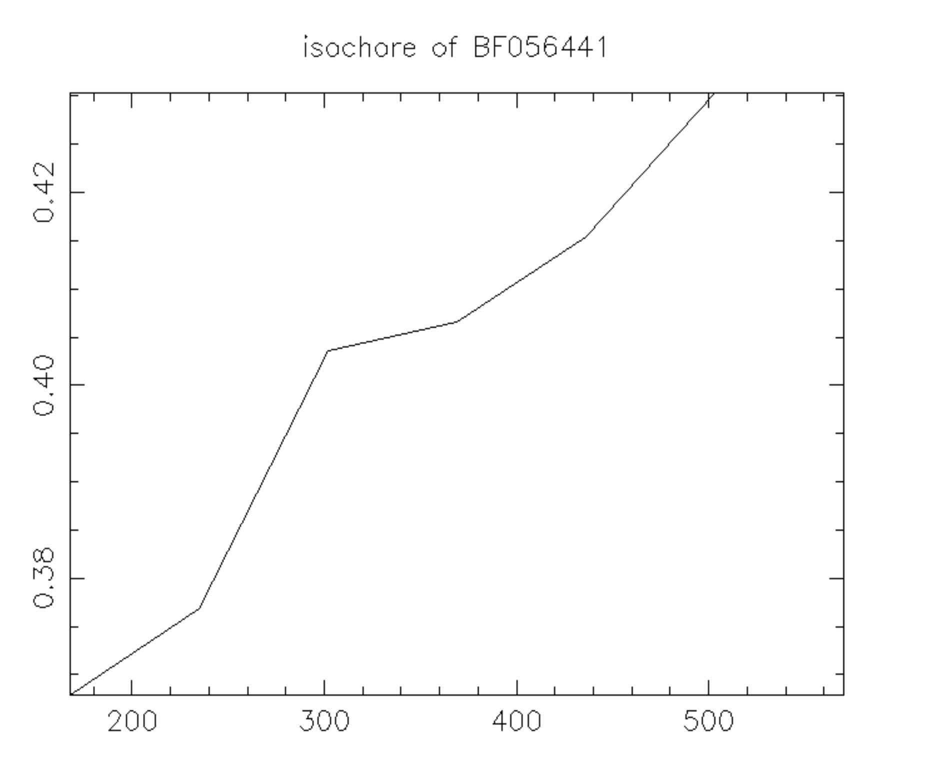
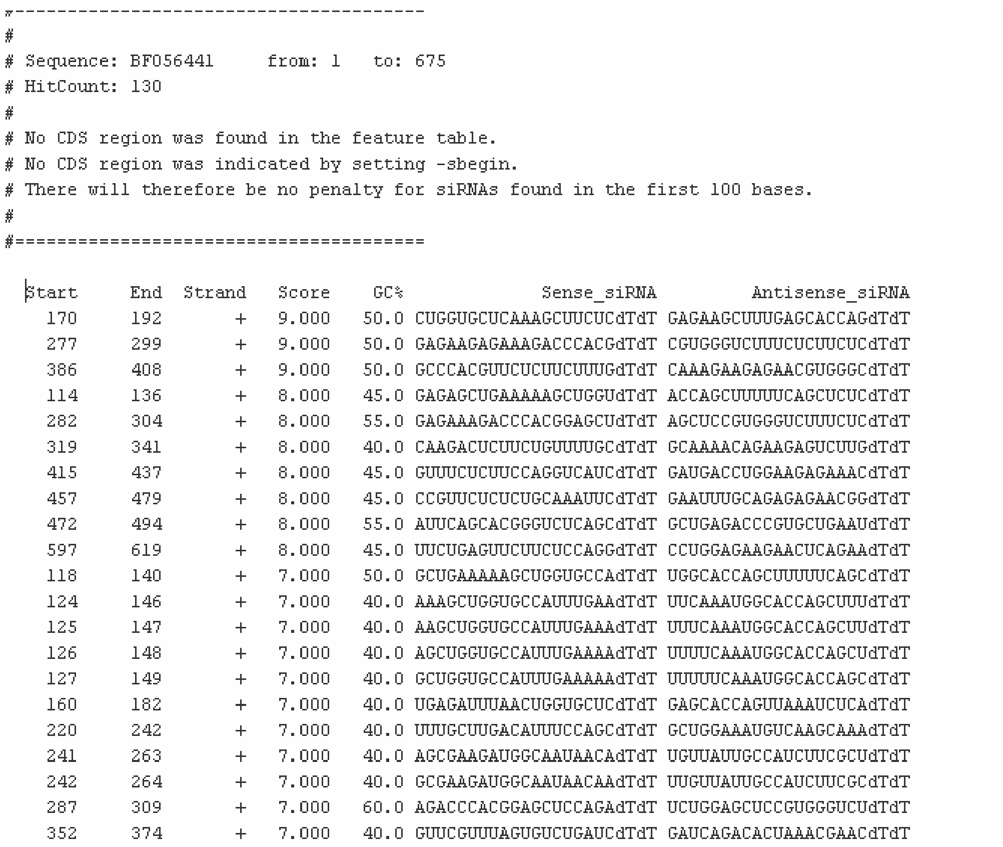


Figure 6.69 : Calculation of isochores in DNA Sequence

Finding of siRNA duplexes in mRNA

Table 6.30 : Finding of siRNA duplexes in mRNA



```
#---------------------------------------
#
# Sequence: BF056441     from: 1   to: 675
# HitCount: 130
#
# No CDS region was found in the feature table.
# No CDS region was indicated by setting -sbegin.
# There will therefore be no penalty for siRNAs found in the first 100 bases.
#
#=======================================

 Start     End  Strand    Score    GC%           Sense_siRNA       Antisense_siRNA
   170     192       +    9.000   50.0 CUGGUGCUCAAAGCUUCUCdTdT GAGAAGCUUUGAGCACCAGdTdT
   277     299       +    9.000   50.0 GAGAAGAGAAAGACCCACGdTdT CGUGGGUCUUUCUCUUCUCdTdT
   386     408       +    9.000   50.0 GCCCACGUUCUCUUCUUUGdTdT CAAAGAAGAGAACGUGGGCdTdT
   114     136       +    8.000   45.0 GAGAGCUGAAAAAGCUGGUdTdT ACCAGCUUUUUCAGCUCUCdTdT
   282     304       +    8.000   55.0 GAGAAAGACCCACGGAGCUdTdT AGCUCCGUGGGUCUUUCUCdTdT
   319     341       +    8.000   40.0 CAAGACUCUUCUGUUUUGCdTdT GCAAAACAGAAGAGUCUUGdTdT
   415     437       +    8.000   45.0 GUUUCUCUUCCAGGUCAUCdTdT GAUGACCUGGAAGAGAAACdTdT
   457     479       +    8.000   45.0 CCGUUCUCUCUGCAAAUUCdTdT GAAUUUGCAGAGAGAACGGdTdT
   472     494       +    8.000   55.0 AUUCAGCACGGGUCUCAGCdTdT GCUGAGACCCGUGCUGAAUdTdT
   597     619       +    8.000   45.0 UUCUGAGUUCUUCUCCAGGdTdT CCUGGAGAAGAACUCAGAAdTdT
   118     140       +    7.000   50.0 GCUGAAAAAGCUGGUGCCAdTdT UGGCACCAGCUUUUUCAGCdTdT
   124     146       +    7.000   40.0 AAAGCUGGUGCCAUUUGAAdTdT UUCAAAUGGCACCAGCUUUdTdT
   125     147       +    7.000   40.0 AAGCUGGUGCCAUUUGAAAdTdT UUUCAAAUGGCACCAGCUUdTdT
   126     148       +    7.000   40.0 AGCUGGUGCCAUUUGAAAAdTdT UUUUCAAAUGGCACCAGCUdTdT
   127     149       +    7.000   40.0 GCUGGUGCCAUUUGAAAAAdTdT UUUUUCAAAUGGCACCAGCdTdT
   160     182       +    7.000   40.0 UGAGAUUUAACUGGUGCUCdTdT GAGCACCAGUUAAAUCUCAdTdT
   220     242       +    7.000   40.0 UUUGCUUGACAUUUCCAGCdTdT GCUGGAAAUGUCAAGCAAAdTdT
   241     263       +    7.000   40.0 AGCGAAGAUGGCAAUAACAdTdT UGUUAUUGCCAUCUUCGCUdTdT
   242     264       +    7.000   40.0 GCGAAGAUGGCAAUAACAAdTdT UUGUUAUUGCCAUCUUCGCdTdT
   287     309       +    7.000   60.0 AGACCCACGGAGCUCCAGAdTdT UCUGGAGCUCCGUGGGUCUdTdT
   352     374       +    7.000   40.0 GUUCGUUUAGUGUCUGAUCdTdT GAUCAGACACUAAACGAACdTdT
```

Calculation of fractional GC Content of Nucleic Acid sequences

Table 6.31: Calculation of fractional GC Content of Nucleic Acid sequences

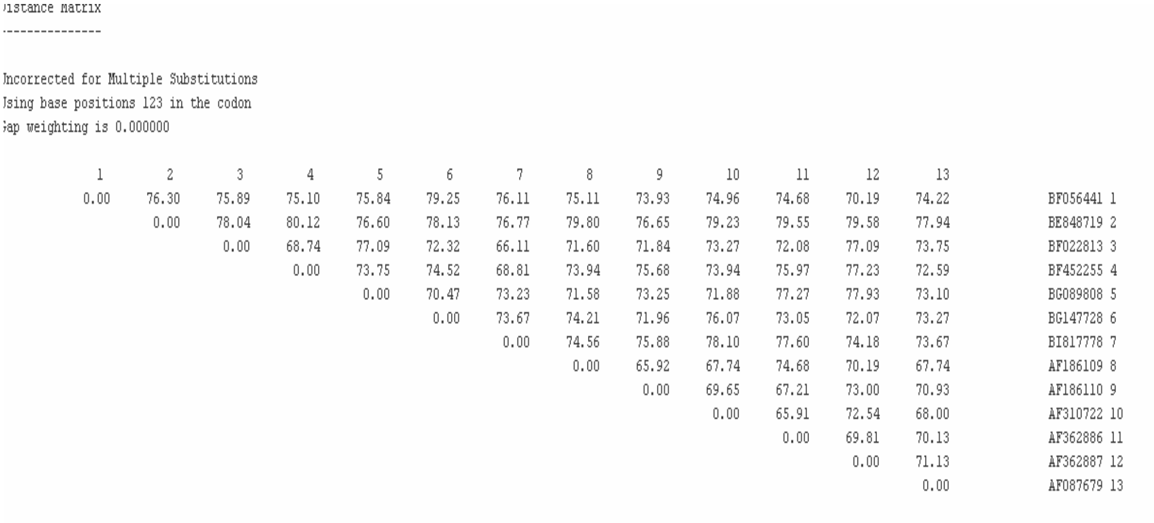
```
#Sequence   GC content
  BF056441      0.40
  BE848719      0.46
  BF022813      0.59
  BF452255      0.57
  BG089808      0.54
  BG147728      0.55
  BI817778      0.62
  AF186109      0.55
  AF186110      0.54
  AF310722      0.56
  AF362886      0.49
  AF362887      0.50
  AF087679      0.52
```

Back -translate to Protein Sequence to ambiguous nucleotide sequence

```
>BF056441 BF056441; 7k05a04.x1 NCI_CGAP_GC6 Homo sapiens cDNA clone IMAGE:3443238 3'
similar to SW:TPM4_HUMAN P07226 TROPOMYOSIN, FIBROBLAST NON-MUSCLE TYPE ;, mRNA sequence.
GCNTGYGCNGGNACNACNGGNTGYGCNGCNGGNGCNGCNACNTGYACNGCNGCNGCNGGN
ACNGGNACNGGNGGNGCNACNACNACNACNGCNACNACNTGYTGYGCNACNACNGGNTGY
GCNTGYGCNGCNACNACNACNGGNTGYACNGCNGGNACNGGNACNGCNACNACNACNTGY
TGYACNGGNGGNGGNACNGCNGGNACNGGNACNGGNGGNACNGGNTGYACNGGNGCNGCN
ACNGCNGCNGCNACNGCNGGNGGNGCNGCNACNGCNGCNGCNACNGGNTGYACNGCNTGY
ACNACNGCNGCNGGNGGNGCNGCNGCNGCNGCNGCNACNGCNGCNGGNGCNGGNGCNGGN
TGYACNGGNGCNGCNGCNGCNGCNGGNTGYACNGGNGGNACNGGNTGYTGYGCNACNACN
ACNGGNGCNGCNGCNGCNGCNGCNGCNGCNGCNGCNGGNGGNGGNGCNGCNGGNGGNGCN
GCNACNGGNGCNGGNGCNACNACNACNGCNGCNTGYACNGGNGGNACNGGNTGYACNTGY
GCNGCNGCNGGNTGYACNACNTGYACNTGYTGYGGNGCNACNGCNTGYGCNGCNGCNGCN
ACNGCNACNACNACNGGNGGNACNTGYGCNACNGGNACNGCNACNACNTGYGCNACNGCN
GCNACNACNACNGGNTGYACNACNGGNGCNTGYGCNACNACNACNTGYTGYGCNGGNTGY
GCNGCNGCNGGNTGYGGNGCNGCNGGNGCNACNGGNGGNTGYGCNGCNACNGCNGCNTGY
GCNGCNGCNGCNGGNGGNGCNGCNTGYACNACNTGYACNACNGCNTGYGCNGCNGGNGCN
GGNGCNGCNGGNGCNGGNGCNGCNGCNGGNGCNTGYTGYTGYGCNTGYGGNGGNGCNGGN
TGYACNTGYTGYGCNGGNGCNGGNACNACNACNTGYACNGGNACNACNGGNGGNGCNGCN
TGYGCNGCNGGNGCNTGYACNTGYACNACNTGYACNGGNACNACNACNACNGGNTGYACN
ACNGCNACNGCNACNGCNTGYGCNGGNACNACNGCNGCNGGNACNACNTGYGGNACNACN
ACNGCNGGNACNGGNACNTGYACNGGNGCNACNTGYTGYGCNGGNACNGGNACNTGYACN
GGNGCNACNGGNACNGCNGCNGGNTGYTGYTGYGCNTGYGGNACNACNTGYACNTGYACN
ACNTGYACNACNACNGGNGGNTGYTGYACNGGNGGNGGNTGYGCNGCNGGNACNACNACN
TGYACNTGYACNACNTGYTGYGCNGGNGGNACNTGYGCNACNTGYGCNGCNACNACNGGN
ACNTGYACNACNACNACNTGYTGYGCNGGNACNACNACNACNGGNTGYGCNGCNTGYTGY
GGNACNACNTGYACNTGYACNTGYACNGGNTGYGCNGCNGCNACNACNTGYGCNGGNTGY
GCNTGYGGNGGNGGNACNTGYACNTGYGCNGGNTGYTGYACNTGYACNACNACNTGYGCN
GGNACNACNACNGGNACNTGYGCNGGNGCNTGYGCNGGNGCNGCNGGNACNACNACNGCN
GCNACNACNACNTGYACNACNTGYACNACNTGYGCNACNGCNACNACNACNGGNACNTGY
TGYACNTGYTGYACNACNACNACNTGYGCNGGNGCNGCNACNGCNTGYACNACNACNACN
TGYGCNGGNGCNACNGGNTGYGCNGGNTGYTGYACNTGYTGYGCNGGNGCNGGNGCNACN
ACNACNTGYGCNGGNGCNACNACNGGNACNGCNGGNACNACNGCNTGYGCNGCNACNACN
TGYACNGGNGCNGGNACNACNTGYACNACNTGYACNTGYTGYGCNGGNGGNACNTGYGCN
TGYTGYGCNTGYGCNACNACNACNACNGCNACNACNTGYGCNGGNGCNTGYGCNTGYTGY
ACNTGYTGYGGNTGYGCNTGYGGNTGYACNACNTGYACNTGYACNGGNTGYTGYTGYACN
TGYACNTGYGCNGGNTGYACNGCNGCNTGYTGYTGYACNACNTGY
```

Figure 6.70 : Back -translate to Protein Sequence to ambiguous nucleotide sequence

Creation of distance matrix from multiple sequence alignment



```
istance Matrix
--------------

ncorrected for Multiple Substitutions
sing base positions 123 in the codon
ap weighting is 0.000000

           1       2       3       4       5       6       7       8       9      10      11      12      13
        0.00   76.30   75.89   75.10   75.84   79.25   76.11   75.11   73.93   74.96   74.68   70.19   74.22          BF056441 1
                0.00   78.04   80.12   76.60   78.13   76.77   79.80   76.65   79.23   79.55   79.58   77.94          BE848719 2
                        0.00   68.74   77.09   72.32   66.11   71.60   71.84   73.27   72.08   77.09   73.75          BF022813 3
                                0.00   73.75   74.52   68.81   73.94   75.68   73.94   75.97   77.23   72.59          BF452255 4
                                        0.00   70.47   73.23   71.58   73.25   71.88   77.27   77.93   73.10          BG089808 5
                                                0.00   73.67   74.21   71.96   76.07   73.05   72.07   73.27          BG147728 6
                                                        0.00   74.56   75.88   78.10   77.60   74.18   73.67          BI817778 7
                                                                0.00   65.92   67.74   74.68   70.19   67.74          AF186109 8
                                                                        0.00   69.65   67.21   73.00   70.93          AF186110 9
                                                                                0.00   65.91   72.54   68.00          AF310722 10
                                                                                        0.00   69.81   70.13          AF362886 11
                                                                                                0.00   71.13          AF362887 12
                                                                                                        0.00          AF087679 13
```

Figure 6.71 : Creation of distance matrix from multiple sequence alignment

6.2.3.2 Nucleotide Sequence Using DAMBE

Under this research work nucleotide sequence has been created from multiple alignments through *tropomyosin.fasta* file using DAMBE software.

Running FASTA algorithm to align locally two sequences

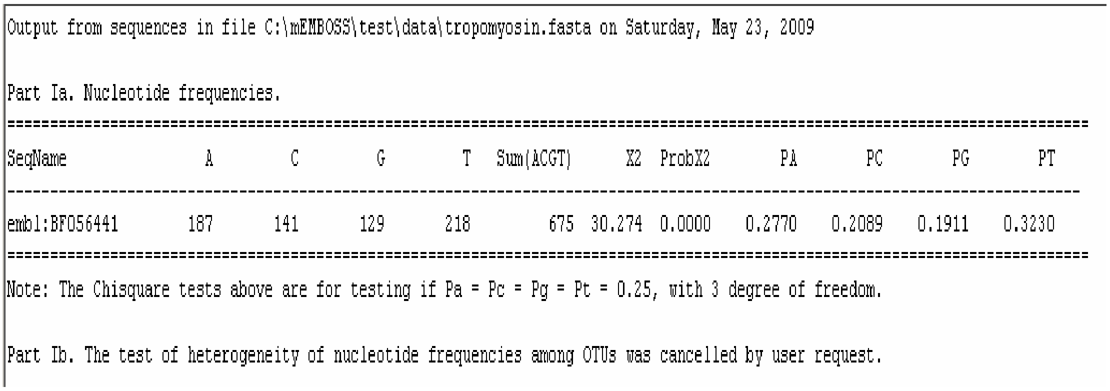
Table 6.32: Running FASTA algorithm to align locally two sequences

```
Query:  ACCGCGATGACGAATA
Target: GAATACGACTGACGATGGA
N-tuple: 1

Result of the local alignment
Number of matched words of length 1: 7
Query:  ACCGCGATGACGAATA
Target: GAATACGACTGACGATGGA

Computational details:

Part I. Hash table of query

0         0         6         9         12        13        15
1         1         2         4         10
2         3         5         8         11
3         7         14


Part II. Target seq. table

0         -3        -5        -8        -11
1         1         -5        -8        -11       -12       -14
2         2         -4        -7        -10       -11       -13
3         -4        -11
4         4         -2        -5        -8        -9        -11
5         4         3         1         -5
6         3         1         -2        -5
7         7         1         -2        -5        -6        -8
8         7         6         4         -2
9         2         -5
10        7         5         2         -1
11        11        5         2         -1        -2        -4
12        11        10        8         2
13        10        8         5         2
14        14        8         5         2         1         -1
15        8         1
16        13        11        8         5
17        14        12        9         6
18        18        12        9         6         5         3


Part III. Frequency table

-15       0
-14       1
-13       1
-12       1
-11       5
-10       1
-9        1
-8        4
-7        1
-6        1
-5        7
-4        3
-3        1
-2        5
-1        3
0         0
1         6
```

```
2          7
3          3
4          3
5          6
6          3
7          3
```

Nucleotide Frequency Calculation

Table 6.33: Nucleotide Frequency Calculation



```
Output from sequences in file C:\mEMBOSS\test\data\tropomyosin.fasta on Saturday, May 23, 2009

Part Ia. Nucleotide frequencies.
==========================================================================================================================
SeqName                A        C        G        T   Sum(ACGT)      X2  ProbX2        PA        PC        PG        PT
--------------------------------------------------------------------------------------------------------------------------
embl:BF056441        187      141      129      218         675  30.274  0.0000    0.2770    0.2089    0.1911    0.3230
==========================================================================================================================
Note: The Chisquare tests above are for testing if Pa = Pc = Pg = Pt = 0.25, with 3 degree of freedom.

Part Ib. The test of heterogeneity of nucleotide frequencies among OTUs was cancelled by user request.
```

Relative CpG, TpG and CpA abundance and GC%

Table 6.34 :Relative CpG, TpG and CpA abundance and GC%

```
Relative CpG, TpG and CpA abundance and GC%

==================================================================
SeqName                     RA(CpG)         RA(TpG+CpA)       GC%
------------------------------------------------------------------
embl:BF056441                0.4688              1.7385    0.4000
embl:BE848719                0.4070              1.7275    0.4642
embl:BF022813                1.7365              3.0724    0.5919
embl:BF452255                1.2272              2.3809    0.5706
embl:BG089808                0.9545              1.8657    0.5380
embl:BG147728                1.0379              2.3310    0.5506
embl:BI817778                1.5579              2.3169    0.6239
embl:AF186109                0.9357              1.6668    0.5503
embl:AF186110                0.5796              1.4454    0.5413
embl:AF310722                0.6665              1.3112    0.5621
embl:AF362886                1.4989              4.1446    0.4903
embl:AF362887                1.0261              2.7307    0.5023
embl:AF087679                0.7944              1.2812    0.5158
==================================================================
```

```
RA is the odds-ratio measure, i.e., F(XY)/[F(X)*F(Y)] for quantifying the relative
abundance of dinucleotides (for nucleotide sequences) or di-aa (for protein sequences).
  For nucleotide sequences, F(X) and F(Y) are the frequencies of X and Y, respectively,
and F(XY) is the frequency of dinucleotide XY.
  For amino acid sequences, they are amino acid and di-aa frequencies.

XY is considered high (or low) when RA > 1.25 (or <0.78). See Hollander, M., and D. A.
Wolfe. 1973. Nonparametric statistical methods. Wiley, New York.
```

```
For an application of RA, see Karlin, S., W. Doerfler, and L. R. Cardon. 1994. Why is CpG
suppressed in the genomes of virtually all small eukaryotic viruses but not in those of
large eukaryotic viruses? J Virol 68:2889-2897.


Part II: Distribution of CpG in individual sequences:

Summary statistics of the inter-CpG distances

SeqName                              N      Mean       STD        CV      Skew      Kurt
-----------------------------------------------------------------------------------------------
embl:BF056441                        7   67.4286   54.7809    0.8124    0.8692    0.5025
embl:BE848719                        9   60.7778   51.2732    0.8436    0.9363    0.9433
embl:BF022813                       14   18.0417   30.2834    1.6785    3.2732   11.6885
embl:BF452255                       15   19.4583   30.9234    1.5892    2.9717    9.8669
embl:BG089808                       15   17.9630   26.4175    1.4707    2.7305    7.9922
embl:BG147728                       12   21.2000   33.9017    1.5991    2.7200    7.8178
embl:BI817778                       21   16.5357   15.1645    0.9171    1.8831    4.0354
embl:AF186109                       20   21.1212   33.3164    1.5774    3.3791   13.5583
embl:AF186110                       25   29.7667   34.7768    1.1683    2.7001    9.3338
embl:AF310722                       28   21.7778   30.5977    1.4050    3.1991   13.1335
embl:AF362886                        7   43.4286   52.4622    1.2080    1.6513    2.6721
embl:AF362887                       10   41.4000   50.7701    1.2263    2.2652    5.7178
embl:AF087679                       21   23.4857   26.0151    1.1077    1.6260    2.0147
-----------------------------------------------------------------------------------------------
```

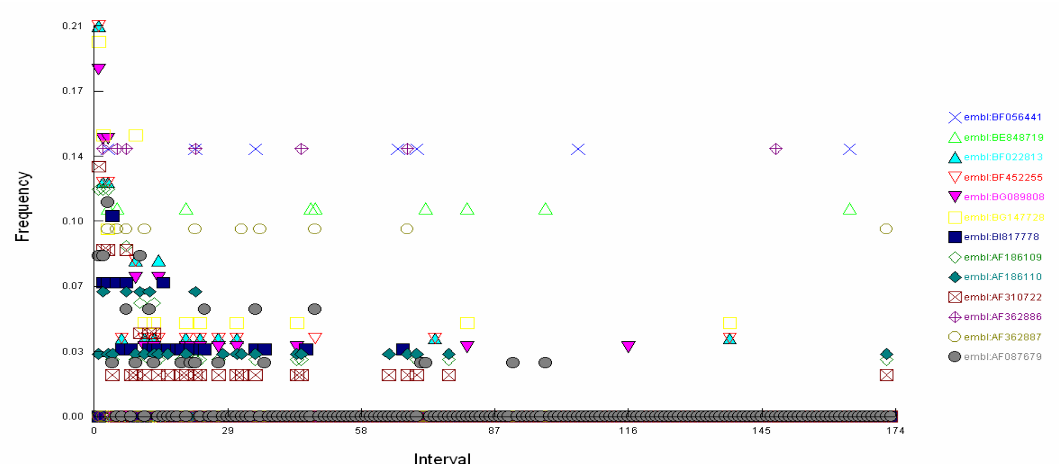
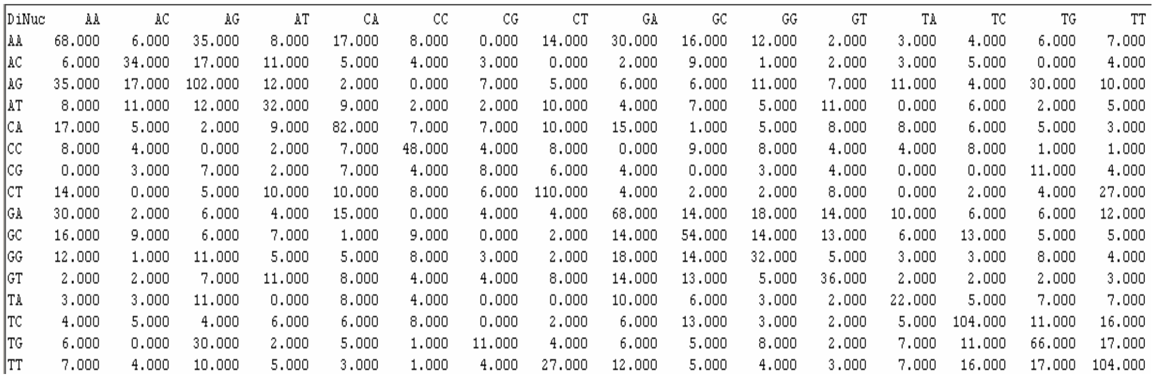


Figure 6.72: Graph of Relative CpG, TpG and CpA abundance and GC%

Di-nucleotide Substitution Pattern

Table 6.35: Di-nucleotide Substitution Pattern



DiNuc	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA	68.000	6.000	35.000	8.000	17.000	8.000	0.000	14.000	30.000	16.000	12.000	2.000	3.000	4.000	6.000	7.000
AC	6.000	34.000	17.000	11.000	5.000	4.000	3.000	0.000	2.000	9.000	1.000	2.000	3.000	5.000	0.000	4.000
AG	35.000	17.000	102.000	12.000	2.000	0.000	7.000	5.000	6.000	6.000	11.000	7.000	11.000	4.000	30.000	10.000
AT	8.000	11.000	12.000	32.000	9.000	2.000	2.000	10.000	4.000	7.000	5.000	11.000	0.000	6.000	2.000	5.000
CA	17.000	5.000	2.000	9.000	82.000	7.000	7.000	10.000	15.000	1.000	5.000	8.000	8.000	6.000	5.000	3.000
CC	8.000	4.000	0.000	2.000	7.000	48.000	4.000	8.000	0.000	9.000	8.000	4.000	4.000	8.000	1.000	1.000
CG	0.000	3.000	7.000	2.000	7.000	4.000	8.000	6.000	4.000	0.000	3.000	4.000	0.000	0.000	11.000	4.000
CT	14.000	0.000	5.000	10.000	10.000	8.000	6.000	110.000	4.000	2.000	2.000	8.000	0.000	2.000	4.000	27.000
GA	30.000	2.000	6.000	4.000	15.000	0.000	4.000	4.000	68.000	14.000	18.000	14.000	10.000	6.000	6.000	12.000
GC	16.000	9.000	6.000	7.000	1.000	9.000	0.000	2.000	14.000	54.000	14.000	13.000	6.000	13.000	5.000	5.000
GG	12.000	1.000	11.000	5.000	5.000	8.000	3.000	2.000	18.000	14.000	32.000	5.000	3.000	3.000	8.000	4.000
GT	2.000	2.000	7.000	11.000	8.000	4.000	4.000	8.000	14.000	13.000	5.000	36.000	2.000	2.000	2.000	3.000
TA	3.000	3.000	11.000	0.000	8.000	4.000	0.000	0.000	10.000	6.000	3.000	2.000	22.000	5.000	7.000	7.000
TC	4.000	5.000	4.000	6.000	6.000	8.000	0.000	2.000	6.000	13.000	3.000	2.000	5.000	104.000	11.000	16.000
TG	6.000	0.000	30.000	2.000	5.000	1.000	11.000	4.000	6.000	5.000	8.000	2.000	7.000	11.000	66.000	17.000
TT	7.000	4.000	10.000	5.000	3.000	1.000	4.000	27.000	12.000	5.000	4.000	3.000	7.000	16.000	17.000	104.000

Graph Stacking Energy vs. Sequence Numbering

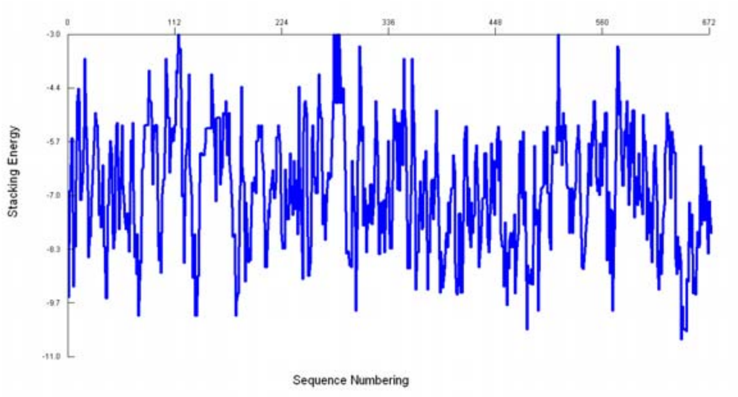
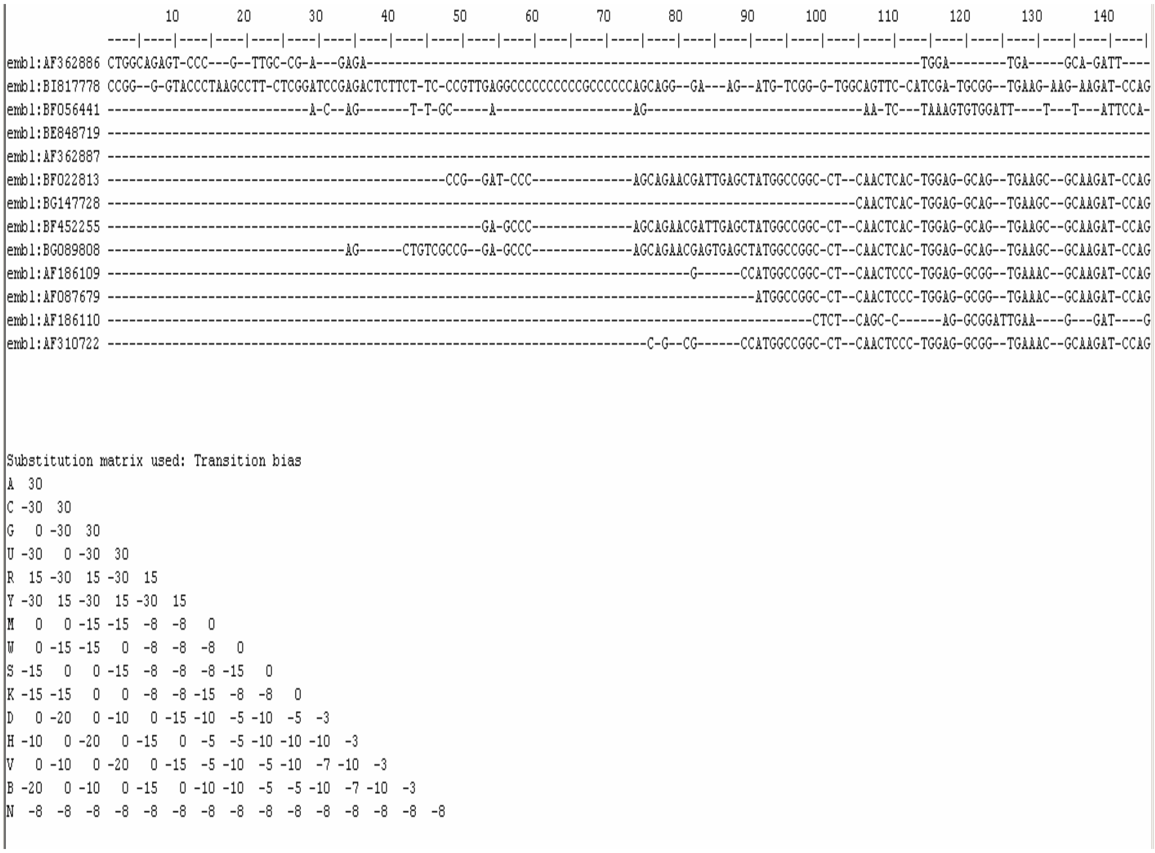


Figure 6.73: Graph Stacking Energy vs Sequence Numbering

Quick Multiple Alignment of Nucleotide Sequence



```
                      10        20        30        40        50        60        70        80        90       100       110       120       130       140
              ----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
embl:AF362886 CTGGCAGAGT-CCC---G--TTGC-CG-A---GAGA-----------------------------------------------------------------------------TGGA--------TGA-----GCA-GATT----
embl:BI817778 CCGG--G-GTACCCTAAGCCTT-CTCGGATCCGAGACTCTTCT-TC-CCGTTGAGGCCCCCCCCCCGCCCCCCAGCAGG--GA---AG--ATG-TCGG-G-TGGCAGTTC-CATCGA-TGCGG--TGAAG-AAG-AAGAT-CCAG
embl:BF056441 ----------------------------A-C--AG-------T-T-GC-----A-------------------AG------------------------------AA-TC---TAAAGTGTGGATT----T---T---ATTCCA-
embl:BE848719 -------------------------------------------------------------------------------------------------------------------------------------------------
embl:AF362887 -------------------------------------------------------------------------------------------------------------------------------------------------
embl:BF022813 -----------------------------------------------CCG--GAT-CCC--------------AGCAGAACGATTGAGCTATGGCCGGC-CT--CAACTCAC-TGGAG-GCAG--TGAAGC--GCAAGAT-CCAG
embl:BG147728 --------------------------------------------------------------------------------------------------------CAACTCAC-TGGAG-GCAG--TGAAGC--GCAAGAT-CCAG
embl:BF452255 ----------------------------------------------------GA-GCCC--------------AGCAGAACGATTGAGCTATGGCCGGC-CT--CAACTCAC-TGGAG-GCAG--TGAAGC--GCAAGAT-CCAG
embl:BG089808 ---------------------------------AG------CTGTCGCCG--GA-GCCC--------------AGCAGAACGAGTGAGCTATGGCCGGC-CT--CAACTCAC-TGGAG-GCAG--TGAAGC--GCAAGAT-CCAG
embl:AF186109 ---------------------------------------------------------------------------------G------CCATGGCCGGC-CT--CAACTCCC-TGGAG-GCGG--TGAAAC--GCAAGAT-CCAG
embl:AF087679 ------------------------------------------------------------------------------------------ATGGCCGGC-CT--CAACTCCC-TGGAG-GCGG--TGAAAC--GCAAGAT-CCAG
embl:AF186110 --------------------------------------------------------------------------------------------------CTCT--CAGC-C------AG-GCGGATTGAA----G---GAT----G
embl:AF310722 ---------------------------------------------------------------------------C-G--CG------CCATGGCCGGC-CT--CAACTCCC-TGGAG-GCGG--TGAAAC--GCAAGAT-CCAG




Substitution matrix used: Transition bias
A  30
C -30  30
G   0 -30  30
U -30   0 -30  30
R  15 -30  15 -30  15
Y -30  15 -30  15 -30  15
M   0   0 -15 -15  -8  -8   0
W   0 -15 -15   0  -8  -8  -8   0
S -15   0   0 -15  -8  -8  -8 -15   0
K -15 -15   0   0  -8  -8 -15  -8  -8   0
D   0 -20   0 -10   0 -15 -10  -5 -10  -5  -3
H -10   0 -20   0 -15   0  -5  -5 -10 -10 -10  -3
V   0 -10   0 -20   0 -15  -5 -10  -5 -10  -7 -10  -3
B -20   0 -10   0 -15   0 -10 -10  -5  -5 -10  -7 -10  -3
N  -8  -8  -8  -8  -8  -8  -8  -8  -8  -8  -8  -8  -8  -8  -8
```

Figure 6.74: Quick Multiple Alignment of Nucleotide Sequence

Creation of Phylogenetics Tree of Nucleotide sequence based on Distance Method

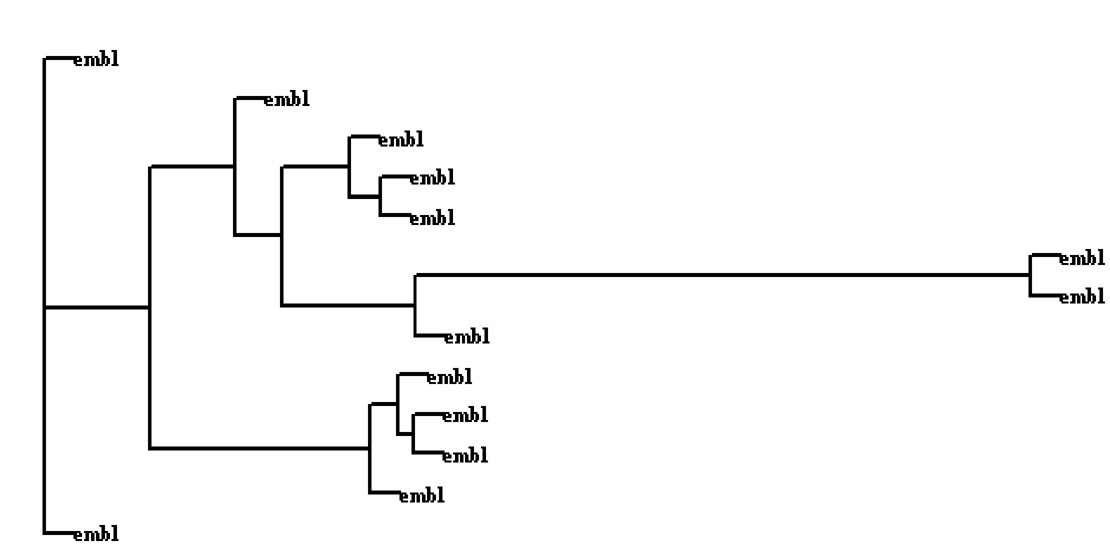
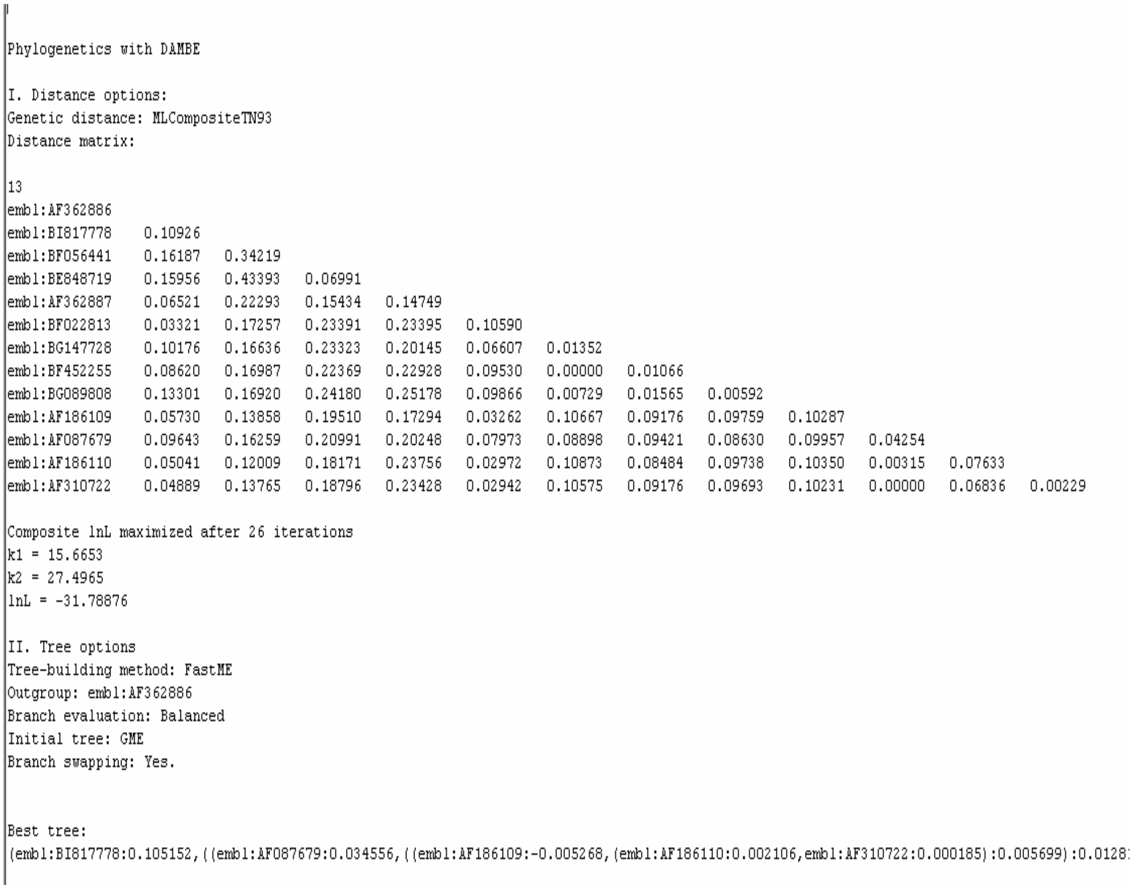


Figure 6.75: Creation of Phylogenetics Tree

Phylogenetics with DAMBE



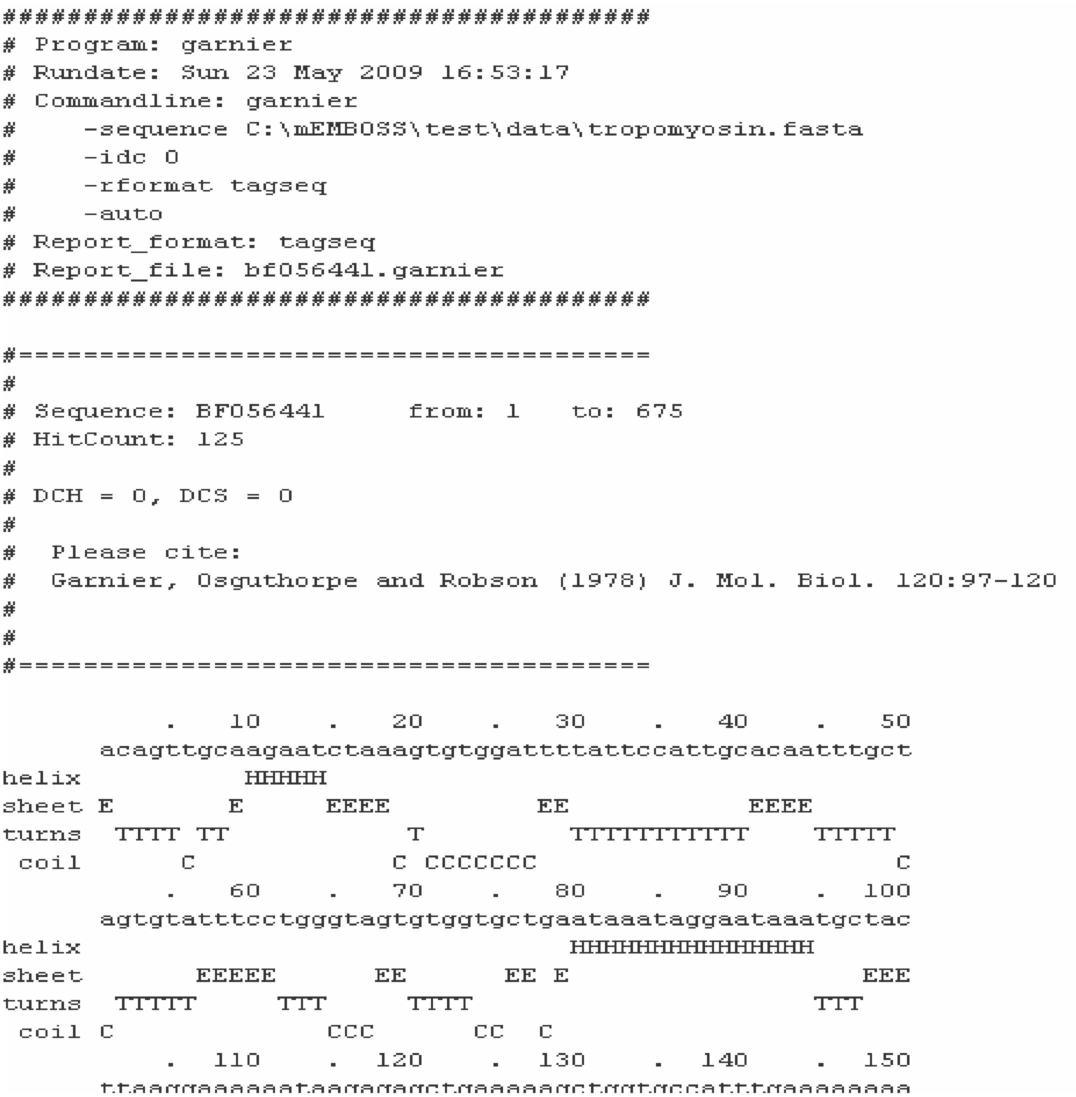
```
Phylogenetics with DAMBE

I. Distance options:
Genetic distance: MLCompositeTN93
Distance matrix:

13
embl:AF362886
embl:BI817778    0.10926
embl:BF056441    0.16187   0.34219
embl:BE848719    0.15956   0.43393   0.06991
embl:AF362887    0.06521   0.22293   0.15434   0.14749
embl:BF022813    0.03321   0.17257   0.23391   0.23395   0.10590
embl:BG147728    0.10176   0.16636   0.23323   0.20145   0.06607   0.01352
embl:BF452255    0.08620   0.16987   0.22369   0.22928   0.09530   0.00000   0.01066
embl:BG089808    0.13301   0.16920   0.24180   0.25178   0.09866   0.00729   0.01565   0.00592
embl:AF186109    0.05730   0.13858   0.19510   0.17294   0.03262   0.10667   0.09176   0.09759   0.10287
embl:AF087679    0.09643   0.16259   0.20991   0.20248   0.07973   0.08898   0.09421   0.08630   0.09957   0.04254
embl:AF186110    0.05041   0.12009   0.18171   0.23756   0.02972   0.10873   0.08484   0.09738   0.10350   0.00315   0.07633
embl:AF310722    0.04889   0.13765   0.18796   0.23428   0.02942   0.10575   0.09176   0.09693   0.10231   0.00000   0.06836   0.00229

Composite lnL maximized after 26 iterations
k1 = 15.6653
k2 = 27.4965
lnL = -31.78876

II. Tree options
Tree-building method: FastME
Outgroup: embl:AF362886
Branch evaluation: Balanced
Initial tree: GME
Branch swapping: Yes.


Best tree:
(embl:BI817778:0.105152,((embl:AF087679:0.034556,((embl:AF186109:-0.005268,(embl:AF186110:0.002106,embl:AF310722:0.000185):0.005699):0.0128
```

Figure 6.76: Phylogenetics with DAMBE

6.2.3.3 Protein Sequence Using Jemboss

In this phase of research work protein structure has been analyzed using *tropomyosin.fasta* file through Jemboss software.

Prediction of Protein Secondary structure using GOR Method



```
########################################
# Program: garnier
# Rundate: Sun 23 May 2009 16:53:17
# Commandline: garnier
#    -sequence C:\mEMBOSS\test\data\tropomyosin.fasta
#    -idc 0
#    -rformat tagseq
#    -auto
# Report_format: tagseq
# Report_file: bf056441.garnier
########################################

#=======================================
#
# Sequence: BF056441     from: 1   to: 675
# HitCount: 125
#
# DCH = 0, DCS = 0
#
#  Please cite:
#  Garnier, Osguthorpe and Robson (1978) J. Mol. Biol. 120:97-120
#
#
#=======================================

          .   10    .   20    .   30    .   40    .   50
      acagttgcaagaatctaaagtgtggattttattccattgcacaatttgct
helix          HHHHH
sheet E       E     EEEE         EE           EEEE
turns  TTTT TT           T         TTTTTTTTTTT    TTTTT
 coil      C            C CCCCCCC                      C
          .   60    .   70    .   80    .   90    .  100
      agtgtatttcctgggtagtgtggtgctgaataaataggaataaatgctac
helix                                  HHHHHHHHHHHHHHH
sheet       EEEEE      EE      EE E                  EEE
turns  TTTTT     TTT     TTTT                     TTT
 coil C             CCC      CC  C
          .  110    .  120    .  130    .  140    .  150
      ttaaggaaaaaataagagagctgaaaaagctggtgccatttgaaaaaaaa
```

Figure 6.77: Prediction of Protein Secondary structure using GOR Method

Plot of Hydrophobic Moment for Protein Sequences

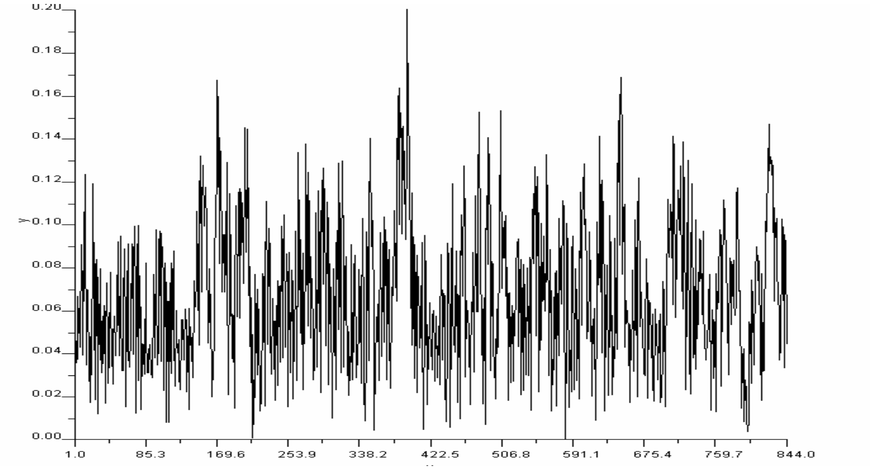


Figure 6.78: Plot of Hydrophobic Moment for Protein Sequences

Drawing a helical Net for Protein Sequence

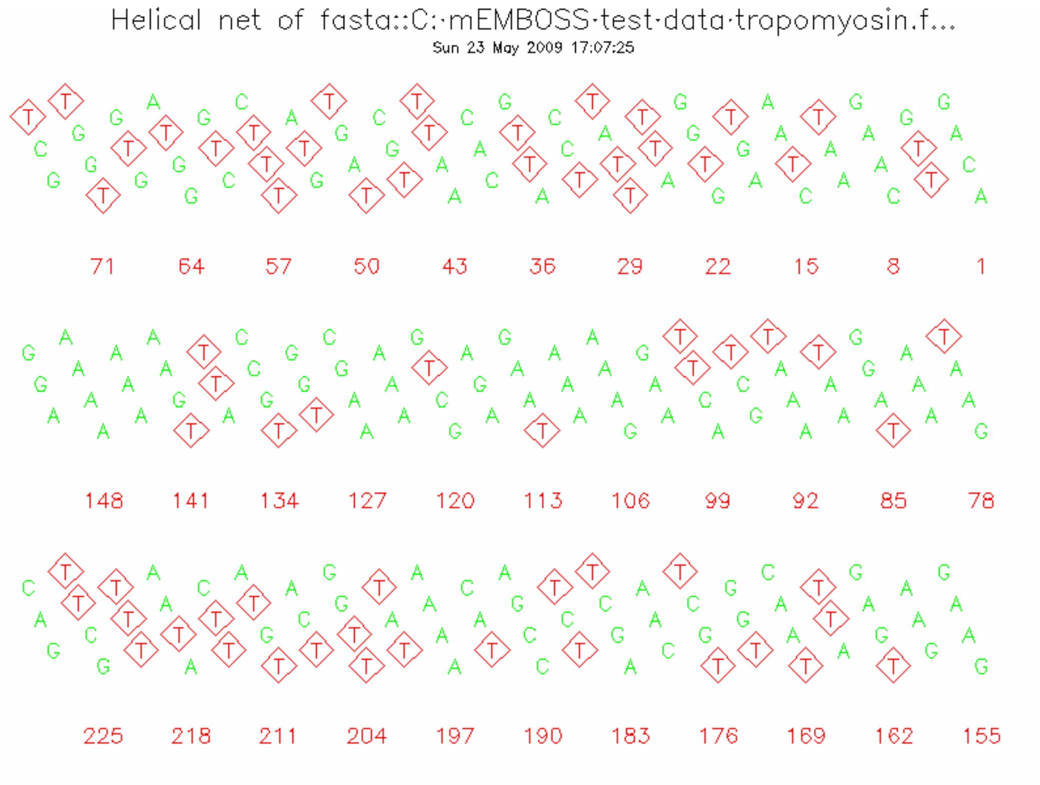
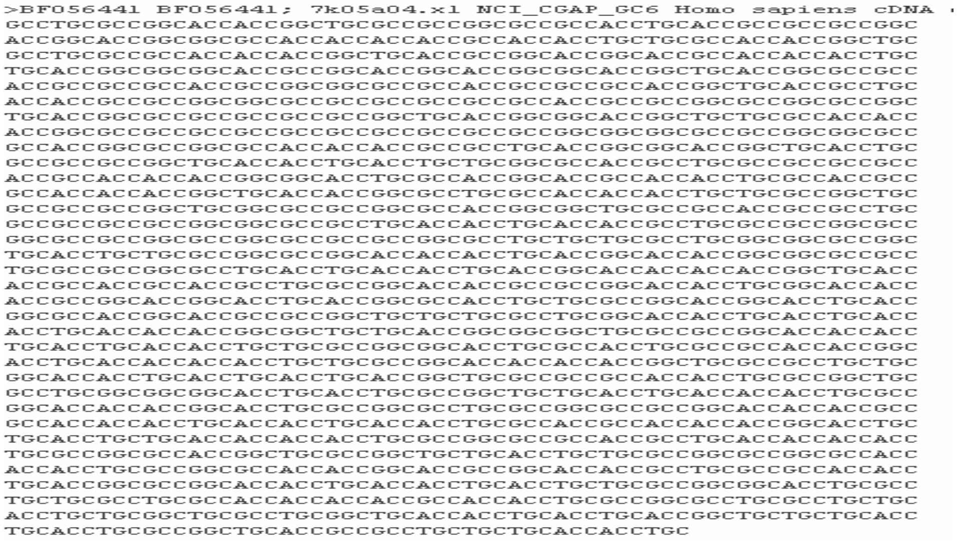


Figure 6.79: Drawing a helical Net for Protein Sequence

Back -translate Protein sequence to Nucleotide Sequence

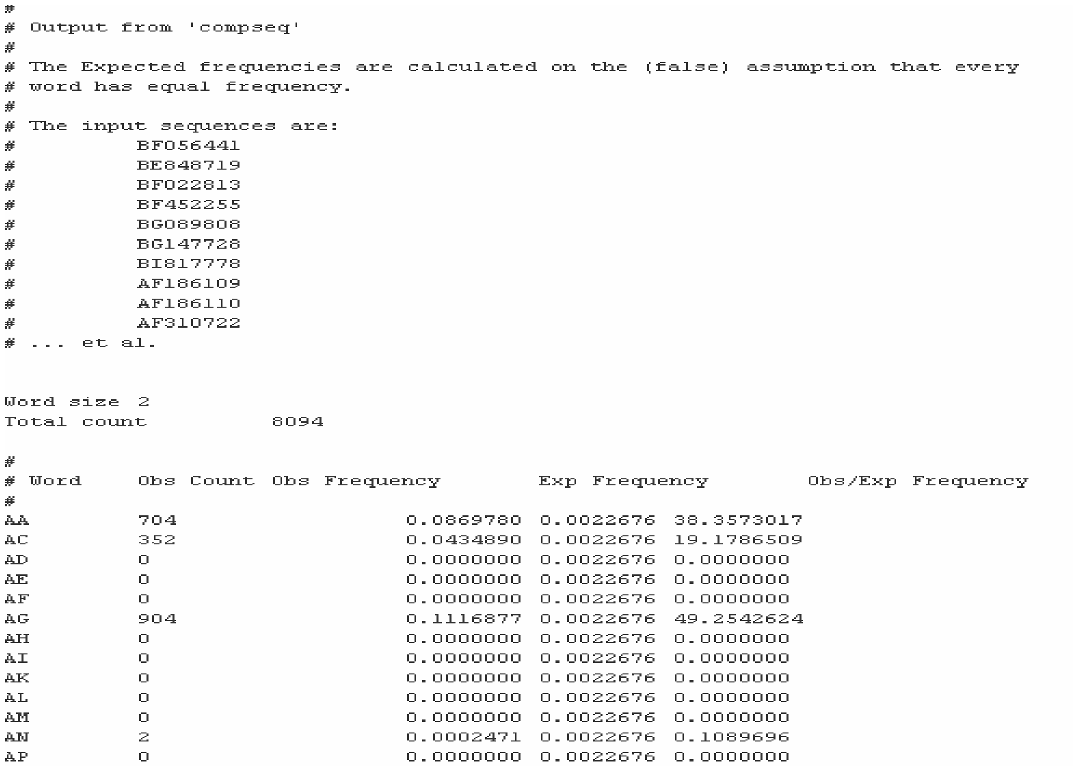


```
>BF056441 BF056441; 7k05a04.x1 NCI_CGAP_GC6 Homo sapiens cDNA
GCCTGCGCCGGCACCACCGGCTGCGCCGCCGGCGCCGCCACCTGCACCGCCGCCGCCGGC
ACCGGCACCGGCGGCGCCACCACCACCACCGCCACCACCTGCTGCGCCACCACCGGCTGC
GCCTGCGCCGCCACCACCACCGGCTGCACCGCCGGCACCGGCACCGCCACCACCACCTGC
TGCACCGGCGGCGGCACCGCCGGCACCGGCACCGGCGGCACCGGCTGCACCGGCGCCGCC
ACCGCCGCCGCCACCGCCGGCGGCGCCGCCACCGCCGCCGCCACCGGCTGCACCGCCTGC
ACCACCGCCGCCGGCGGCGCCGCCGCCGCCGCCGCCACCGCCGCCGGCGCCGGCGCCGGC
TGCACCGGCGCCGCCGCCGCCGCCGGCTGCACCGGCGGCACCGGCTGCTGCGCCACCACC
ACCGGCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGGCGGCGGCGCCGCCGGCGGCGCC
GCCACCGGCGCCGGCGCCACCACCACCGCCGCCTGCACCGGCGGCACCGGCTGCACCTGC
GCCGCCGCCGGCTGCACCACCTGCACCTGCTGCGGCGCCACCGCCTGCGCCGCCGCCGCC
ACCGCCACCACCACCGGCGGCACCTGCGCCACCGGCACCGCCACCACCTGCGCCACCGCC
GCCACCACCACCGGCTGCACCACCGGCGCCTGCGCCACCACCACCTGCTGCGCCGGCTGC
GCCGCCGCCGGCTGCGGCGCCGCCGGCGCCACCGGCGGCTGCGCCGCCACCGCCGCCTGC
GCCGCCGCCGCCGGCGGCGCCGCCTGCACCACCTGCACCACCGCCTGCGCCGCCGGCGCC
GGCGCCGCCGGCGCCGGCGCCGCCGCCGGCGCCTGCTGCTGCGCCTGCGGCGGCGCCGGC
TGCACCTGCTGCGCCGGCGCCGGCACCACCACCTGCACCGGCACCACCGGCGGCGCCGCC
TGCGCCGCCGGCGCCTGCACCTGCACCACCTGCACCGGCACCACCACCACCGGCTGCACC
ACCGCCACCGCCACCGCCTGCGCCGGCACCACCGCCGCCGGCACCACCTGCGGCACCACC
ACCGCCGGCACCGGCACCTGCACCGGCGCCACCTGCTGCGCCGGCACCGGCACCTGCACC
GGCGCCACCGGCACCGCCGCCGGCTGCTGCTGCGCCTGCGGCACCACCTGCACCTGCACC
ACCTGCACCACCACCGGCGGCTGCTGCACCGGCGGCGGCTGCGCCGCCGGCACCACCACC
TGCACCTGCACCACCTGCTGCGCCGGCGGCACCTGCGCCACCTGCGCCGCCACCACCGGC
ACCTGCACCACCACCACCTGCTGCGCCGGCACCACCACCACCGGCTGCGCCGCCTGCTGC
GGCACCACCTGCACCTGCACCTGCACCGGCTGCGCCGCCGCCACCACCTGCGCCGGCTGC
GCCTGCGGCGGCGGCACCTGCACCTGCGCCGGCTGCTGCACCTGCACCACCACCTGCGCC
GGCACCACCACCGGCACCTGCGCCGGCGCCTGCGCCGGCGCCGCCGGCACCACCACCGCC
GCCACCACCACCTGCACCACCTGCACCACCTGCGCCACCGCCACCACCACCGGCACCTGC
TGCACCTGCTGCACCACCACCACCTGCGCCGGCGCCGCCACCGCCTGCACCACCACCACC
TGCGCCGGCGCCACCGGCTGCGCCGGCTGCTGCACCTGCTGCGCCGGCGCCGGCGCCACC
ACCACCTGCGCCGGCGCCACCACCGGCACCGCCGGCACCACCGCCTGCGCCGCCACCACC
TGCACCGGCGCCGGCACCACCTGCACCACCTGCACCTGCTGCGCCGGCGGCACCTGCGCC
TGCTGCGCCTGCGCCACCACCACCACCGCCACCACCTGCGCCGGCGCCTGCGCCTGCTGC
ACCTGCTGCGGCTGCGCCTGCGGCTGCACCACCTGCACCTGCACCGGCTGCTGCTGCACC
TGCACCTGCGCCGGCTGCACCGCCGCCTGCTGCTGCACCACCTGC
```

Figure 6.80: Back -translate Protein sequence to Nucleotide Sequence

Calculation of Composition of Unique words in Sequences

Table 6.36: Calculation of Composition of Unique words in Sequences



```
#
# Output from 'compseq'
#
# The Expected frequencies are calculated on the (false) assumption that every
# word has equal frequency.
#
# The input sequences are:
#         BF056441
#         BE848719
#         BF022813
#         BF452255
#         BG089808
#         BG147728
#         BI817778
#         AF186109
#         AF186110
#         AF310722
# ... et al.


Word size  2
Total count          8094

#
# Word     Obs Count  Obs Frequency        Exp Frequency        Obs/Exp Frequency
#
AA         704                  0.0869780  0.0022676  38.3573017
AC         352                  0.0434890  0.0022676  19.1786509
AD         0                    0.0000000  0.0022676  0.0000000
AE         0                    0.0000000  0.0022676  0.0000000
AF         0                    0.0000000  0.0022676  0.0000000
AG         904                  0.1116877  0.0022676  49.2542624
AH         0                    0.0000000  0.0022676  0.0000000
AI         0                    0.0000000  0.0022676  0.0000000
AK         0                    0.0000000  0.0022676  0.0000000
AL         0                    0.0000000  0.0022676  0.0000000
AM         0                    0.0000000  0.0022676  0.0000000
AN         2                    0.0002471  0.0022676  0.1089696
AP         0                    0.0000000  0.0022676  0.0000000
```

Generation of residue / base frequency plot

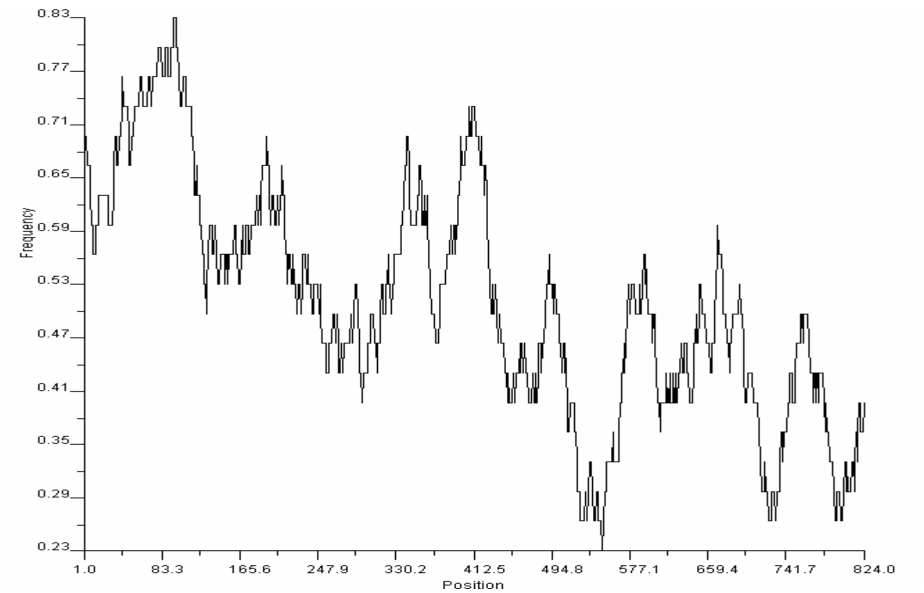
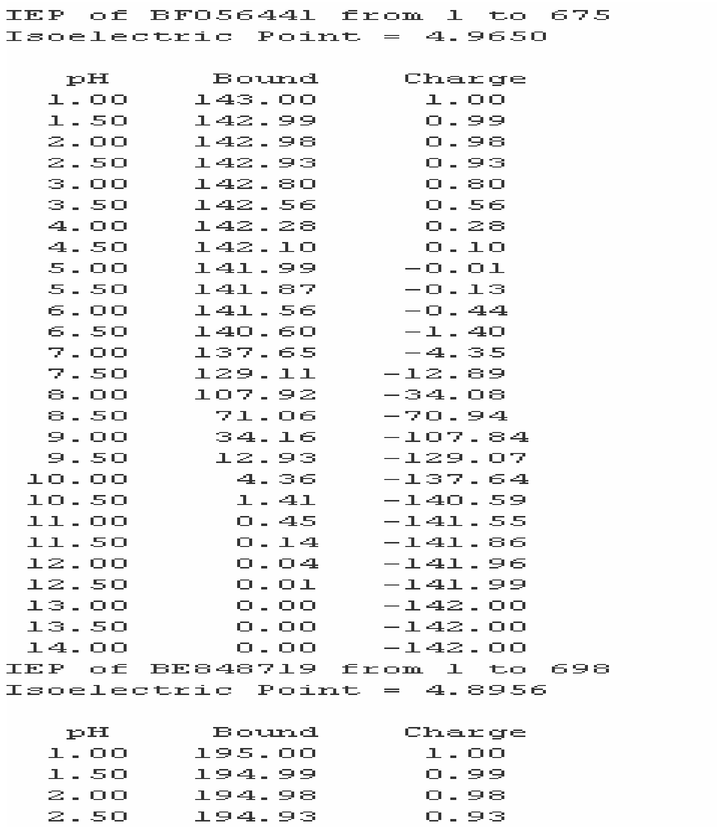


Figure 6.81 : Generation of residue / base frequency plot

Calculation of isoelectric points of protein

Table 6.37: Calculation of isoelectric points of protein



```
IEP of BF056441 from 1 to 675
Isoelectric Point = 4.9650

    pH       Bound      Charge
   1.00    143.00       1.00
   1.50    142.99       0.99
   2.00    142.98       0.98
   2.50    142.93       0.93
   3.00    142.80       0.80
   3.50    142.56       0.56
   4.00    142.28       0.28
   4.50    142.10       0.10
   5.00    141.99      -0.01
   5.50    141.87      -0.13
   6.00    141.56      -0.44
   6.50    140.60      -1.40
   7.00    137.65      -4.35
   7.50    129.11     -12.89
   8.00    107.92     -34.08
   8.50     71.06     -70.94
   9.00     34.16    -107.84
   9.50     12.93    -129.07
  10.00      4.36    -137.64
  10.50      1.41    -140.59
  11.00      0.45    -141.55
  11.50      0.14    -141.86
  12.00      0.04    -141.96
  12.50      0.01    -141.99
  13.00      0.00    -142.00
  13.50      0.00    -142.00
  14.00      0.00    -142.00
IEP of BE848719 from 1 to 698
Isoelectric Point = 4.8956

    pH       Bound      Charge
   1.00    195.00       1.00
   1.50    194.99       0.99
   2.00    194.98       0.98
   2.50    194.93       0.93
```

Isoelectric Point Plot Charge vs PH

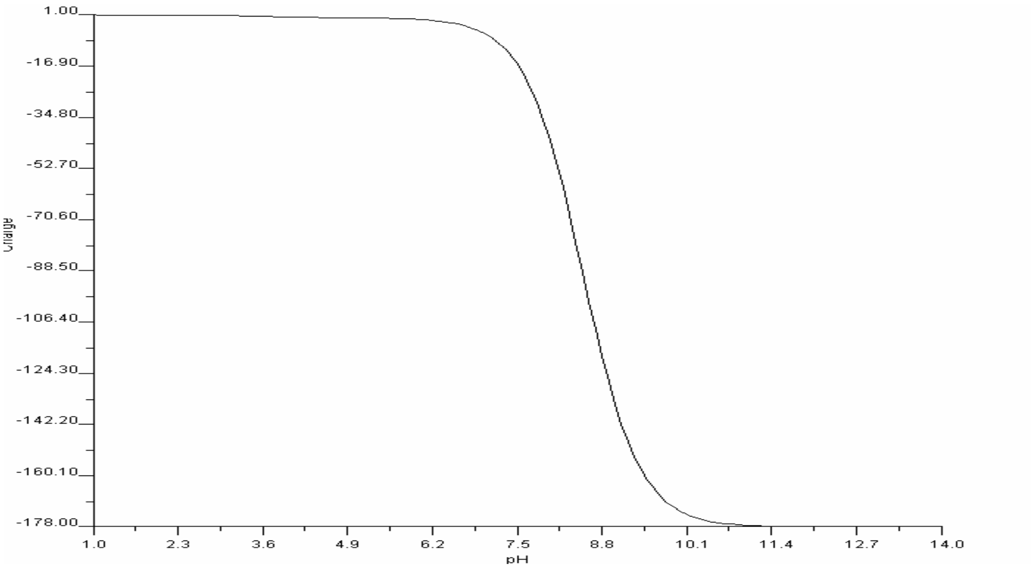


Figure 6.82: Isoelectric Point Plot Charge vs PH

Calculation of Statistics of protein properties

Table 6.38: Calculation of Statistics of protein properties

```
PEPSTATS of BF056441 from 1 to 675

Molecular weight = 57252.78          Residues = 675
Average Residue Weight  = 84.819     Charge   = 0.0
Isoelectric Point = 4.9650
A280 Molar Extinction Coefficient  = 0
A280 Extinction Coefficient 1mg/ml = 0.00
Improbability of expression in inclusion bodies = 0.586

Residue        Number         Mole%          DayhoffStat

A = Ala        187            27.704         3.221
B = Asx        0              0.000          0.000
C = Cys        141            20.889         7.203
D = Asp        0              0.000          0.000
E = Glu        0              0.000          0.000
F = Phe        0              0.000          0.000
G = Gly        129            19.111         2.275
H = His        0              0.000          0.000
I = Ile        0              0.000          0.000
J = ---        0              0.000          0.000
K = Lys        0              0.000          0.000
L = Leu        0              0.000          0.000
M = Met        0              0.000          0.000
N = Asn        0              0.000          0.000
O = ---        0              0.000          0.000
P = Pro        0              0.000          0.000
Q = Gln        0              0.000          0.000
R = Arg        0              0.000          0.000
S = Ser        0              0.000          0.000
T = Thr        218            32.296         5.294
U = ---        0              0.000          0.000
V = Val        0              0.000          0.000
```

```
W = Trp        0              0.000          0.000
X = Xaa        0              0.000          0.000
Y = Tyr        0              0.000          0.000
Z = Glx        0              0.000          0.000

Property       Residues              Number         Mole%

Tiny           (A+C+G+S+T)           675            100.000
Small          (A+B+C+D+G+N+P+S+T+V) 675            100.000
Aliphatic      (A+I+L+V)             187             27.704
Aromatic       (F+H+W+Y)             0                0.000
Non-polar      (A+C+F+G+I+L+M+P+V+W+Y)457            67.704
Polar          (D+E+H+K+N+Q+R+S+T+Z) 218             32.296
Charged        (B+D+E+H+K+R+Z)       0                0.000
Basic          (H+K+R)               0                0.000
Acidic         (B+D+E+Z)             0                0.000
```

Plot of amino acid properties of proteins in parallel

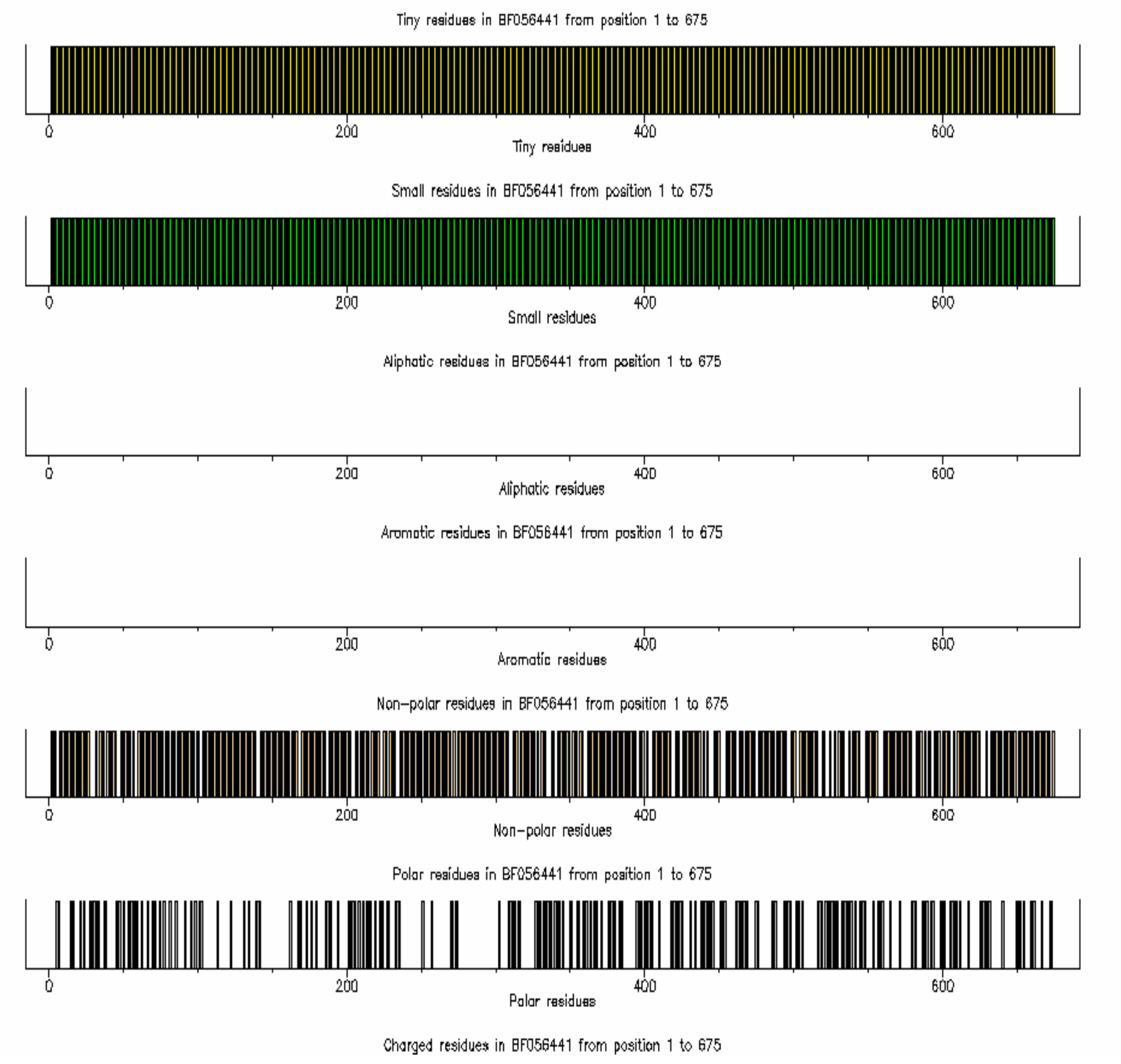


Figure 6.83 : Plot of Histogram of general properties

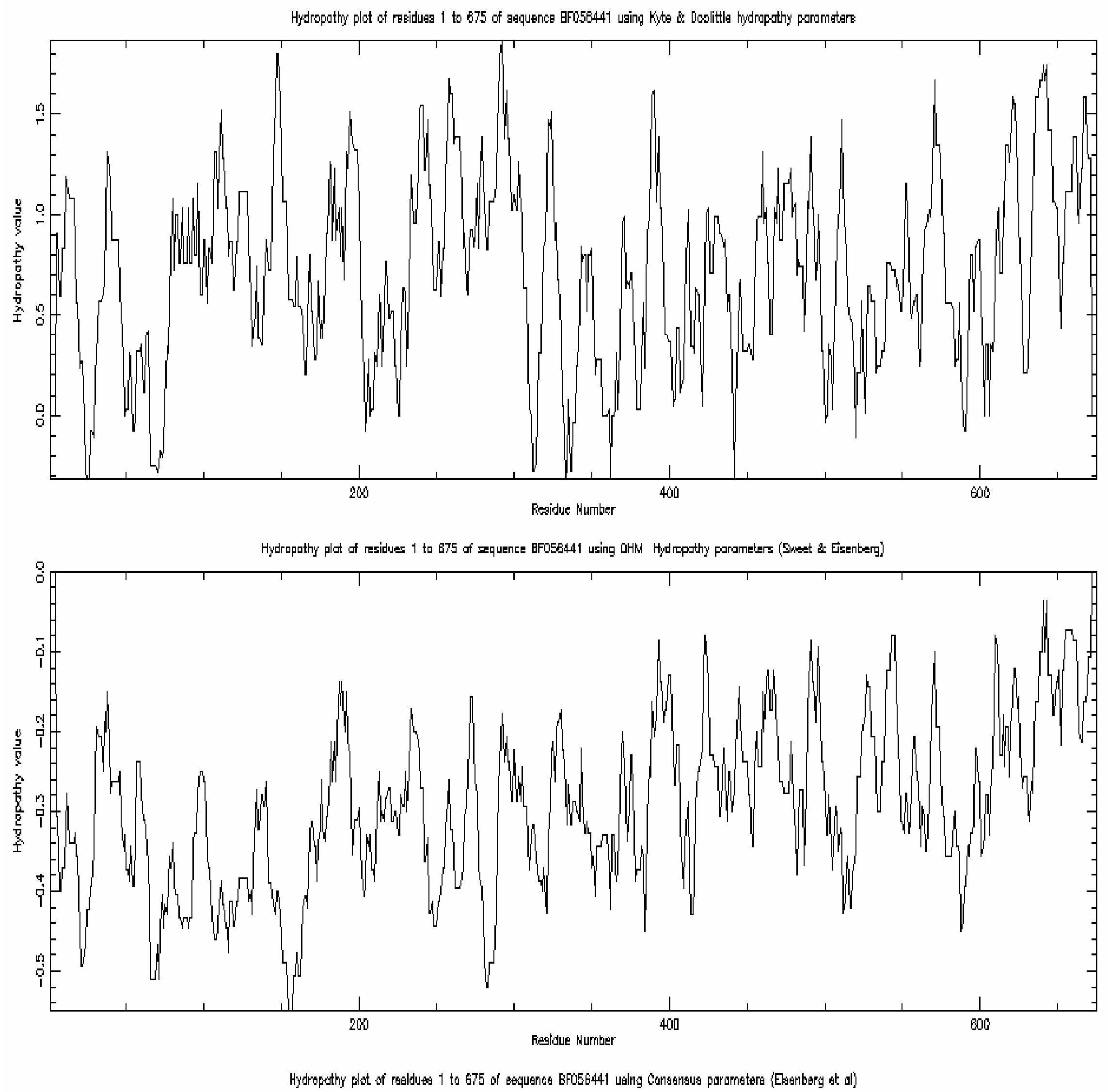


Figure 6.84: Plot graph of hydropathy

WaterMan- Eggert Local Alignment of two Protein Sequences

```
#######################################
# Program: matcher
# Rundate: Sun 23 May 2009 20:40:49
# Commandline: matcher
#    -asequence C:\mEMBOSS\test\data\tropomyosin.fasta
#    -sprotein1
#    -bsequence C:\mEMBOSS\test\data\opsd.fasta
#    -sprotein2
#    -alternatives 1
#    -gapopen 0
#    -gapextend 0
#    -aformat markx0
#    -auto
# Align_format: markx0
# Report_file: bf056441.matcher
########################################
```

```
# Aligned_sequences: 2
# 1: BF056441
# 2: OPSD_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 0
# Extend_penalty: 0
#
# Length: 624
# Identity:      88/624 (14.1%)
# Similarity:   105/624 (16.8%)
# Gaps:         519/624 (83.2%)
# Score: 487
#
#
#=======================================

             10                20
BF0564 GT-TGCAAG-------A-ATCTAAAG---T---------GTGGA-----T
       ::      :       . ::     :   .             :
OPSD_H GTE-----GPNFYVPFSNAT-----GVVRSPFEYPQYYL----AEPWQF-
                  10             20        30

        30        40                        50
BF0564 TTTATTCCATTGCA--CAA--------TTTG------CT---AGT-----
                    .   ::           :       :     :
OPSD_H -------------SML-AAYMFLLIVL---GFPINFL-TLYV--TVQHKK
                     40         50            60

                        60              70         80
BF0564 --GT---------A-TTTCCTGGGTA------GTG-TGGTGCTGAAT--A
          :         :           :      : : :  .      :
OPSD_H LR-TPLNYILLNLAV----------ADLFMVLG-GFT--S------TLY-
         70        80                   90

                    90       100            110
BF0564 AATA--G----G-AATAAATGCTAC---TTAAG--GAAAAAAT-AAGAG-
         :.  :    :       ::   :       :       : :   : :
OPSD_H --TSLHGYFVFGP------TG---CNLE----GFF-----A-TL--G-GE
          100                110                   120

         120                  130          140       150
BF0564 -A--GCT----GAAA-------AAGCTGGTGC---CATTTGAAAAAAAA-
        :    .       :                :                .
OPSD_H IALW--SLVVL---AIERYVVV---------CKPM-------------SN
              130                   140

                       160            170       180
BF0564 ---AAG---GGA--AG-GA-AT---GA-GATTTAACTGGTGCTCAA---A
            :     :   :  :  :    :       :        :::   :
OPSD_H FRF--GENH--AIM-GV-AF-TWVM-AL------A--------CAAPPLA
           150            160                      170

                 190         200       210
BF0564 G-CTTCT-----CCG--ATACAAAATATTTGGTCATG----T--------
       :     .       :     :           .:  :    :
OPSD_H GW----SRYIPE--GLQ---C-----------SC--GIDYYTLKPEVNNE
               180                         190       200

                              220              230
BF0564 A----------T--------TC-ATAATTTG----CT---TGACATTTCC
       .          :         :        :     :     : :
OPSD_H SFVIYMFVVHFTIPMIIIFF-CY-------GQLVF-TVKE--A-A-----
             210       220                230

           240       250         260           270
BF0564 A----GCAAAGCGAAGATGGCAAT--AACAAAAGGA----ACTTCT----
       :      .:        :     :           :         :
OPSD_H AQQQE--SA--------T-----TQK---------AEKEV-----TRMVI
            240                             250

                      280       290            300
BF0564 ----TA---C----AAGAGAAGAGAAAGACCCA-CGGA----GCT--CCA
            :   :                  :   .    :      :
OPSD_H IMVI-AFLICWVPY--------------A---SV---AFYIF--THQ---
          260                        270

               310           320
BF0564 GA--G----TTTCTGT--TGGA--A-CAAGA----------------C--
       :.  :          :     :  :  .: :                :
OPSD_H GSNFGPIFM------TIP---AFFAK-SA-AIYNPVIYIMMNKQFRNCML
     280             290            300       310
```

```
        330        340       350              360
BF0564 TCTTCTGTT-TTGCTTATATACAGTTAAG----TTCG---TTTAGTGTCT
              ::    :       :      :       :      :   .
OPSD_H -------TTI---C-------C------GKNPL---GDDE---A---S--
             320                          330

      370          380
BF0564 GAT-CCA-GT-GTCT--GA-TGTA
        ::   .  :  : .   :    :
OPSD_H -ATV--SK-TE-T-SQV-AP---A
              340
```

Figure 6.85: WaterMan- Eggert Local Alignment of two Protein Sequences

Needleman-Wunsch Global Alignment of two sequences

```
########################################
# Program: needle
# Rundate: Sun 23 May 2009 20:46:00
# Commandline: needle
#    -asequence C:\mEMBOSS\test\data\tropomyosin.fasta
#    -sprotein1
#    -bsequence C:\mEMBOSS\test\data\opsd.fasta
#    -sprotein2
#    -gapopen 0.0
#    -gapextend 0.0
#    -brief
#    -aformat srspair
#    -auto
# Align_format: srspair
# Report_file: bf056441.needle
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: BF056441
# 2: OPSD_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 0.0
# Extend_penalty: 0.0
#
# Length: 894
# Identity:      88/894 ( 9.8%)
# Similarity:   105/894 (11.7%)
# Gaps:         765/894 (85.6%)
# Score: 486.0
#
#
#=======================================

BF056441           1 acagttgcaagaatctaaagtgtggattttattccattgcacaatttgct     50

OPSD_HUMAN         0 --------------------------------------------------      0

BF056441          51 agtgtatttcctgggtagtgtggtgctgaataaataggaataaatgctac    100

OPSD_HUMAN         0 --------------------------------------------------      0

BF056441         101 ttaaggaaaaaataagagagctgaaaaagctggtgccatttgaaaaaaaa    150

OPSD_HUMAN         0 --------------------------------------------------      0

BF056441         151 aagggaaggaatgagatttaactggtgctcaaagcttctccgatacaaaa    200

OPSD_HUMAN         0 --------------------------------------------------      0

BF056441         201 tatttggtcatgtattcataatttgcttgacatttccagcaaagcgaaga    250

OPSD_HUMAN         0 --------------------------------------------------      0

BF056441         251 tggcaataacaaaaggaacttcttacaagagaagagaaagacccacgga-    299
                                  ..|       |
OPSD_HUMAN         1 -------------MNG-------T-------------------------E      5
```

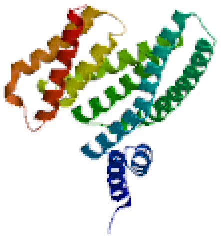
```
BF056441         300 gctcc-------agagtttctgttgg---a---------ac-----a--a    323
                     |           :.|        | |   :         |      :  |
OPSD_HUMAN         6 G----PNFYVPFSNA--------T-GVVRSPFEYPQYYLA-EPWQFSMLA     41

BF056441         324 gactcttctgttt------t-gc---t--t--ata------tac------    347
                      |                 . |    .  |  .|.      |
OPSD_HUMAN        42 -A-----------YMFLLIVLG-FPINFLTLYVTVQHKKLRT--PLNYIL     76

BF056441         348 ---agtta----a-gttcgtt-tagtg--tct--gatcc--a-g-tgtct    380
                        |  .|    .  |   |   |: |   | :  |      . | || |
OPSD_HUMAN        77 LNLA--VADLFMVLG---G--FTS-T-LYT-SLHG----YFVFGPTG-C-    110

BF056441         381 gatgtaa---gccc--acgttctcttctttggcctg-ggca--agtttct    422
                               |     |                 |  ||    |
OPSD_HUMAN       111 -------NLEG---FFA-----------------T-LGG--EIA------    124

BF056441         423 ct--tcc----aggtc-----atcaa---tt---gtcttttcc---agtt    452
                         :      |         ..|     :.   |           |
OPSD_HUMAN       125 --LWS--LVVLA----IERYVVVC--KPMSNFRFG--------ENHA---    153

BF056441         453 tt--gcaaccgttctctc-tgca-a-attc-agcacgggtctcagcctct    496
                         | .|           |    . |    | ||        |
OPSD_HUMAN       154 --IMG-VA----------FT---WVMA---LA-CA--------A------    169

BF056441         497 ttc---agtt-t-----gt--cagacagaagtttaatt----tc------    525
                           ||   :     |   |  :|    |           |
OPSD_HUMAN       170 ---PPLAG--WSRYIPEG-LQC--SC----G-------IDYYT-LKPEVN    199

BF056441         526 t-tc-------t--t--------catattt-gtcctcctt--t-tc---a    550
                     . :        .  |        |       |          . |    |
OPSD_HUMAN       200 NES-FVIYMFVVHFTIPMIIIFFC------YG--------QLVFT-VKEA    233

BF056441         551 gaa----tacttttc--ag---atgc--------agcctc---c----ag    576
                      ||    :|   ||   |    .|          |        |    |
OPSD_HUMAN       234 -AAQQQESA---TT-QKA-EKEVT--RMVIIMVIA-----FLICWVPYA-    269

BF056441         577 ag-att----tca--gatt-g----t--agtt--ac-aattctga--g--    605
                     :  |      |    | :. |    |  |     |  :|     |  .
OPSD_HUMAN       270 S-VA--FYIFT--HQG-SNFGPIFMTIPA---FFA-KSA-----AIYNPV    304

BF056441         606 -----t----tct--tct-ccaggtcaccacatttattca----gac--    637
                          .    .|   | | ||  |                     |
OPSD_HUMAN       305 IYIMMNKQFRNC-MLT-TICC--G-----------------KNPLG--DD    331

BF056441         638 -acctccgcacgcttctc-tgccc-tc-tcagc-ta-acccttc    675
                      |  :    |      |  :     |  | :   .| |
OPSD_HUMAN       332 EA--S----A------T-VS----KT-ET-S--QVAPA------    348


#=======================================
#
# Aligned_sequences: 2
# 1: BF056441
# 2: OPSD_XENLA
# Matrix: EBLOSUM62
# Gap_penalty: 0.0
# Extend_penalty: 0.0
#
# Length: 897
# Identity:      84/897 ( 9.4%)
# Similarity:   106/897 (11.8%)
# Gaps:         765/897 (85.3%)
# Score: 484.0
#
#
#=======================================

BF056441           1 acagttgcaagaatctaaagtgtggattttattccattgcacaatttgct     50

OPSD_XENLA         0 --------------------------------------------------      0

BF056441          51 agtgtatttcctgggtagtgtggtgctgaataaataggaataaatgctac    100

OPSD_XENLA         0 --------------------------------------------------      0

BF056441         101 ttaaggaaaaaataagagagctgaaaaagctggtgccatttgaaaaaaaa    150

OPSD_XENLA         0 --------------------------------------------------      0

BF056441         151 aagggaaggaatgagatttaactggtgctcaaagcttctccgatacaaaa    200
```

```
OPSD_XENLA         0 --------------------------------------------------      0

BF056441         201 tatttggtcatgtattcataatttgcttgacatttccagcaaagcgaaga    250

OPSD_XENLA         0 --------------------------------------------------      0

BF056441         251 tggcaataacaaaag-gaac----t--tct-tacaagagaagagaa-ag-    290
                     ..|   |         |       .  : . |           |.. :
OPSD_XENLA         1 MNG---T--------EG---PNFYVPMS-NKT-----------GVVRS-P     23

BF056441         291 --------accc-----acggagctcc-aga--------gtttc------    312
                             |        :   |      | |        |
OPSD_XENLA        24 FDYPQYYLA---EPWQYS---A-----LA-AYMFLLILLG----LPINFM     57

BF056441         313 tg---t-------tggaacaa-------g---actct---t-ctgttttg    338
                     |    |       |              .   |   .   . |      |
OPSD_XENLA        58 T-LFVTIQHKKLRT-------PLNYILLNLVFA---NHFMVLC------G     90

BF056441         339 cttata-tacagttaag--ttc--gttta----g-tgtct---gatcc--    373
                            |    .|     |:   |        | || |    |
OPSD_XENLA        91 ------FT----VT---MYTS-MHG----YFIFGPTG-C-YIEG----FF    116

BF056441         374 agt-gtctg--a--tg-ta-agcc-------c---acgttctcttctttg    406
                     | | |   |  |  :  .. |          |   |
OPSD_XENLA       117 A-TLG---GEVALWS-LVVLA---VERYIVVCKPMA--------------    144

BF056441         407 gcctgg---gca---a--gtttctcttcc-agg-tc---a-tcaattgtc    442
                          .   |     |  |           |   |    | :| |
OPSD_XENLA       145 -----NFRFG--ENHAIMG----------VA--FT-WIMALSC-A-----    168

BF056441         443 ttttccagtttt----gca-acc-----gttct--ctctgcaaat---tc    477
                           |         |   :       |      |:| |    .   |
OPSD_XENLA       169 ------A-----PPLFG--WS--RYIPEG----MQCSC-G----VDYYT-    193

BF056441         478 agc--------acggg----------tc----t---ca-gcctcttt---    498
                                :              |     .   |  |
OPSD_XENLA       194 ---LKPEVNNES----FVIYMFIVHFT-IPLIVIFFC-YG-------RLL    227

BF056441         499 cagtttgtcagac---agaag----t-tta--atttc---ttc------t    529
                     |      |        | ||     : ||   |       .|       .
OPSD_XENLA       228 C------T-----VKEA-AA-QQQESLTT-QKA----EKEVT-RMVVIMV    258

BF056441         530 t----c----a-tatt----t--gtcctccttttca--gaatactt-t--    559
                     .    |    | .|      |  |           :  |        .
OPSD_XENLA       259 VFFLICWVPYAYVA--FYIFTHQG-----------SNFG-------PVFM    288

BF056441         560 tcag--atgc--agcc-tccaga--g-a---t-t----tcagattgtag-    592
                     |     |     |    :  : |  . .   . .    .|
OPSD_XENLA       289 T---VPA---FFA---KS--S-AIYNPVIYIVLNKQFRNC---------L    317

BF056441         593 -tta-caattctgagttcttctcca----g---gtcaccacattttattc    633
                      ||  |    |   |              |   |  :  : |    | |
OPSD_XENLA       318 ITT-LC----C---G----------KNPFGDEDG--S--S-A----A-T-    338

BF056441         634 agacacc-tccgc-acgcttctctgccc-tctcagc-ta-acccttc       675
                         :   |     |      : :     : : :   .: |
OPSD_XENLA       339 ----S--KT----EA------S-S----VS-S-S--QVSPA------       354
```

Figure 6.86: Needleman-Wunsch Global Alignment of two sequences

SWISS-MODEL Repository Model

Model 3D Structure



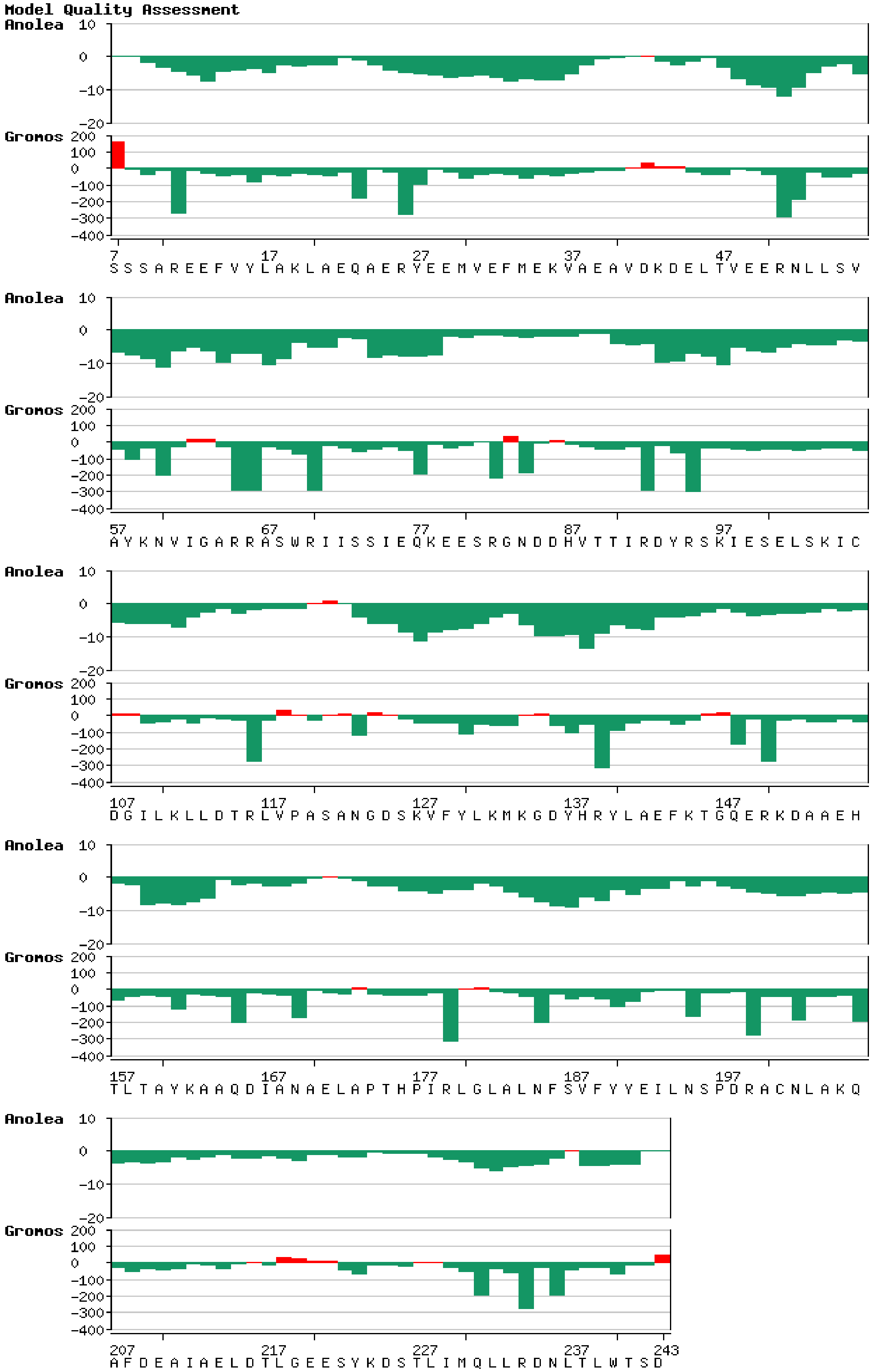
Based on template: 2o98 [SMTL] [PDB] [SCOP] [CATH]

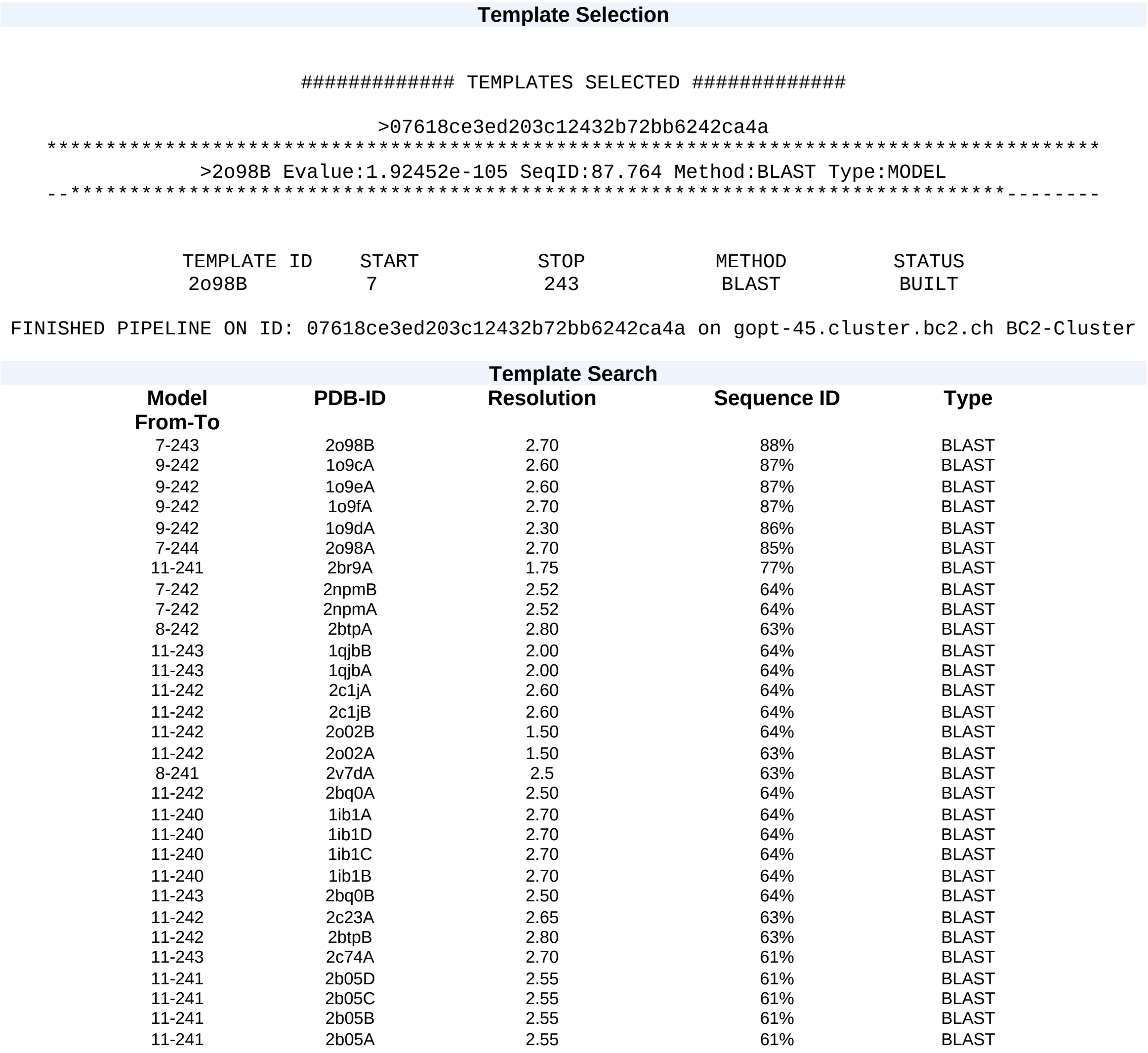
Sequence identity: 87%

Residue range: 7 to 243

Alignment

```
TARGET    7       SSSAREEF VYLAKLAEQA ERYEEMVEFM EKVAEAVDKD ELTVEERNLL
2o98B     4       aptareen vymaklaeqa eryeemvefm ekvsnslgse eltveernll

TARGET                 hhhhh hhhhhhhhh    hhhhhhhh hhhhhh        hhhhhhh
2o98B                  hhhhh hhhhhhhh    hhhhhhhhh hh            hhhhhhh


TARGET    55      SVAYKNVIGA RRASWRIISS IEQKEESRGN DDHVTTIRDY RSKIESELSK
2o98B     52      svayknviga rraswriiss ieqkeesrgn eehvnsirey rskienelsk

TARGET            hhhhhhhhhh hhhhhhhhhh hhhhhh     hhhhhhhhhh hhhhhhhhhh
2o98B             hhhhhhhhhh hhhhhhhhhh hhhhhh     hhhhhhhhhh hhhhhhhhhh


TARGET    105     ICDGILKLLD TRLVPASANG DSKVFYLKMK GDYHRYLAEF KTGQERKDAA
2o98B     102     icdgilklld aklipsaasg dskvfylkmk gdyhrylaef ktgaerkeaa

TARGET            hhhhhhhhhh hhhhhh      hhhhhhhhh hhhhhhhh      hhhhhhh
2o98B             hhhhhhhhhh hhhhhh      hhhhhhhhh hhhhhhhh      hhhhhhh


TARGET    155     EHTLTAYKAA QDIANAELAP THPIRLGLAL NFSVFYYEIL NSPDRACNLA
2o98B     152     estltaykaa qdiattelap thpirlglal nfsvfyyeil nspdracnla

TARGET            hhhhhhhhhh hhhhhh       hhhhhhhh hhhhhhhhh    hhhhhhhh
2o98B             hhhhhhhhhh hhhhhh       hhhhhhhh hhhhhhhhh    hhhhhhhh


TARGET    205     KQAFDEAIAE LDTLGEESYK DSTLIMQLLR DNLTLWTSD
2o98B     202     kqafdeaiae ldtlgeesyk dstlimqllr dnltlwtsd-

TARGET            hhhhhhhhh       hhhhh hhhhhhhhhh hhhhhh
2o98B             hhhhhhhhh       hhhhh hhhhhhhhhh hhhhhh
```


Model Quality Assessment
Anolea
10
0
-10
-20
Gromos
200
100
0
-100
-200
-300
-400
7 17 27 37 47
S S S A R E E F V Y L A K L A E Q A E R Y E E M V E F M E K V A E A V D K D E L T V E E R N L L S V
Anolea
Gromos
57 67 77 87 97
A Y K N V I G A R R A S W R I I S S I E Q K E E S R G N D D H V T T I R D Y R S K I E S E L S K I C
Anolea
Gromos
107 117 127 137 147
D G I L K L L D T R L V P A S A N G D S K V F Y L K M K G D Y H R Y L A E F K T G Q E R K D A A E H
Anolea
Gromos
157 167 177 187 197
T L T A Y K A A Q D I A N A E L A P T H P I R L G L A L N F S V F Y Y E I L N S P D R A C N L A K Q
Anolea
Gromos
207 217 227 237 243
A F D E A I A E L D T L G E E S Y K D S T L I M Q L L R D N L T L W T S D



Template Selection

```
                 ############# TEMPLATES SELECTED #############

                        >07618ce3ed203c12432b72bb6242ca4a
******************************************************************************************
                 >2o98B Evalue:1.92452e-105 SeqID:87.764 Method:BLAST Type:MODEL
--********************************************************************************--------


           TEMPLATE ID    START          STOP           METHOD         STATUS
            2o98B          7              243            BLAST          BUILT

FINISHED PIPELINE ON ID: 07618ce3ed203c12432b72bb6242ca4a on gopt-45.cluster.bc2.ch BC2-Cluster
```

Template Search

Model From-To	PDB-ID	Resolution	Sequence ID	Type
7-243	2o98B	2.70	88%	BLAST
9-242	1o9cA	2.60	87%	BLAST
9-242	1o9eA	2.60	87%	BLAST
9-242	1o9fA	2.70	87%	BLAST
9-242	1o9dA	2.30	86%	BLAST
7-244	2o98A	2.70	85%	BLAST
11-241	2br9A	1.75	77%	BLAST
7-242	2npmB	2.52	64%	BLAST
7-242	2npmA	2.52	64%	BLAST
8-242	2btpA	2.80	63%	BLAST
11-243	1qjbB	2.00	64%	BLAST
11-243	1qjbA	2.00	64%	BLAST
11-242	2c1jA	2.60	64%	BLAST
11-242	2c1jB	2.60	64%	BLAST
11-242	2o02B	1.50	64%	BLAST
11-242	2o02A	1.50	63%	BLAST
8-241	2v7dA	2.5	63%	BLAST
11-242	2bq0A	2.50	64%	BLAST
11-240	1ib1A	2.70	64%	BLAST
11-240	1ib1D	2.70	64%	BLAST
11-240	1ib1C	2.70	64%	BLAST
11-240	1ib1B	2.70	64%	BLAST
11-243	2bq0B	2.50	64%	BLAST
11-242	2c23A	2.65	63%	BLAST
11-242	2btpB	2.80	63%	BLAST
11-243	2c74A	2.70	61%	BLAST
11-241	2b05D	2.55	61%	BLAST
11-241	2b05C	2.55	61%	BLAST
11-241	2b05B	2.55	61%	BLAST
11-241	2b05A	2.55	61%	BLAST

Figure 6. 87: SWISS-MODEL Repository Model

6.3 Data Analysis and Experimental Outcome

In the first phase of experimental work of research (**Activity No-1**) analyses on compounds like Alanine, Amino butyric Acid, Asparagine and Glutamine have been performed using NMRPrediction and ACD/ChemSketch.

This research work describes the characteristics of a free web-based spectral database for chemical research community, containing ^{13}C NMR spectra data from natural compounds. This database allows flexible searching via chemical structure , substructure, name and family of compounds as well as spectral features as chemical shifts , allowing the structural elucidation of known and unknown compounds by comparison of ^{13}C NMR data.

In this experiment script calculates and represents the ^{13}C NMR spectra of the compound. The chemical shifts value obtained in different NMR experiments can be entered with the carbon's hybridization type. This experiment permits to carry out the enquiry with the required number of carbons, from one carbon to the totality of the compound's carbon. It is possible to specify the required deviation (+/-), to the limit in a detailed way. So it limits the search distinctly and therefore a reasonable and manageable compound can be obtained.

If the skeleton of the studied substance is known, and if some distinctive chemical shifts of most important signals are also available, a search by shifts in each particular position of the molecule can be carried out. So it can be obtained the compounds of the family whose shifts, in those position match with those with problem compound.

In this research using ACD/ChemSketch compounds are stored in databases and SMILE codes (Simplified Molecular Input Line Specification) have been generated. A SMILE defines the molecules in the form of alphanumeric chains. This format of structural specification has been used for sharing chemical structure information.

Under this research CML codes of molecules have been developed and that codes have been used for molecular information like symmetry, and atom and bond attributes. Here multiple observations of the same molecule (e.g. conformational analysis and NMR prediction) have been performed.

After that web based structure search queries have been performed on these compounds using web based Pubchem/NCBI. Here activities like bioactivity analysis by structure & activity similarity of molecule , bioactivity analysis by structure & activity similarity of molecule from normalized score percentile , bioactivity analysis by activity & protein target similarity of molecule from normalized score percentile , bioactivity analysis by addition of similar compounds of molecule and revised compound selection after addition of similar compounds of molecule have been performed.

In the second phase of experimental work of research (**Activity No-2**) analyses on same compounds like Alanine, Amino butyric Acid, Asparagine and Glutamine have been performed using ArgusLab tool. Under these experimental work calculations like single entry point calculation, geometry optimization, quick plot HOMO, Quick plot LUMO and quick plot ESP mapped density have been performed.

The outcomes of Activity No-2 have been derived in form of like heat of formation, atom-atom bond orders, atomic spin densities, ground state dipole, SCF plot between energy vs. difference per cycle, final SCF energy, geometric search, comparison between exact and calculated of s2 operator, calculation of ground-state density on grid, calculation of ground-state electrostatic potential on grid and total elapsed time etc. The above all outcomes of compounds like Alanine, Amino butyric Acid, Asparagine and Glutamine have been compared.

It is possible to carry out a combined and simultaneous use of outcomes of Activity No-1 and Activity No-2 like SMILE, chemical shifts, CML, bioactivity analysis of structures, atom-atom bond orders, atomic spin densities, SCF energy and geometric search etc. that undoubtedly amplifies the search capacity and increases the possibilities of finding compounds and to predict their molecular structure.

PROPOSED MODEL FOR MOLECULAR STRUCTURE PREDICTION

In this research a Model for Molecular Structure Prediction has been developed. This model has been used for prediction of molecular structure. The basic components of this model have been shown in this **Figure 6.88**.

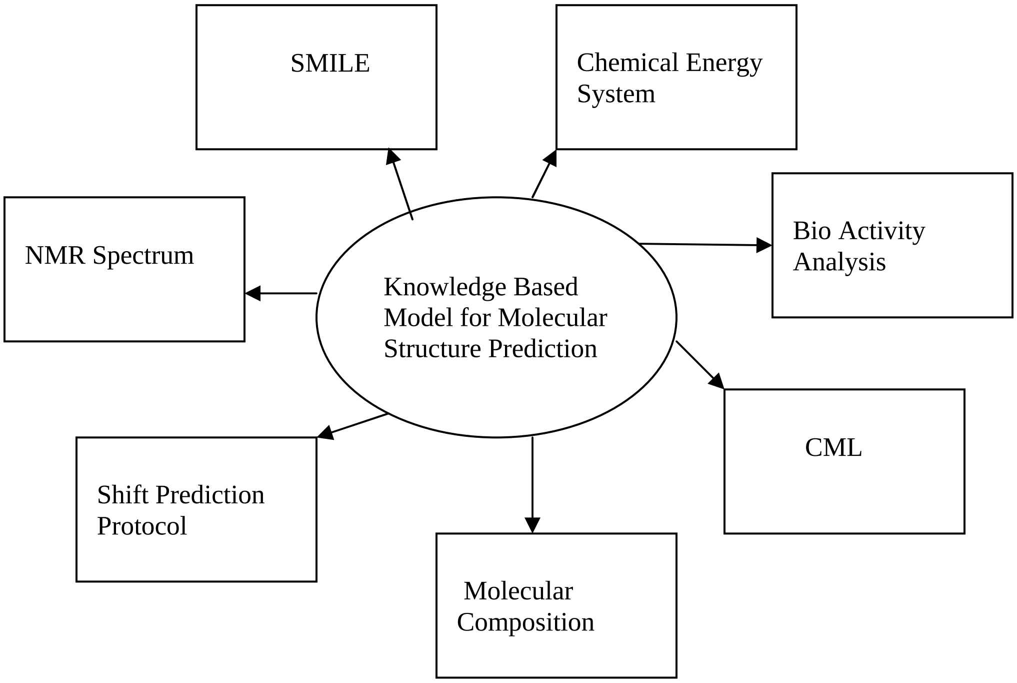
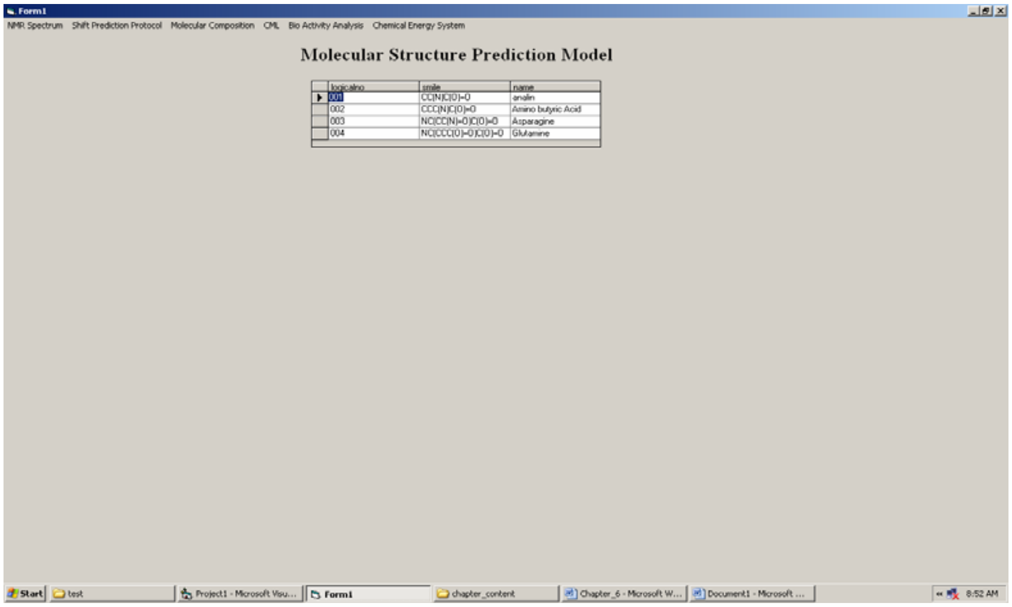


Figure 6.88: Basic Components of Knowledge Based Model for Molecular Structure Prediction

The combinations of all above basic components enhance the capability of this model in predicting molecular structure.

Figure 6.89 is basic view of the model that has been developed. In this model all basic components of compounds like Alanine, Amino butyric Acid, Asparagine and Glutamine have been analyzed and these have been shown in the form of Menu. Here SMILE structure of all above molecules has been generated.

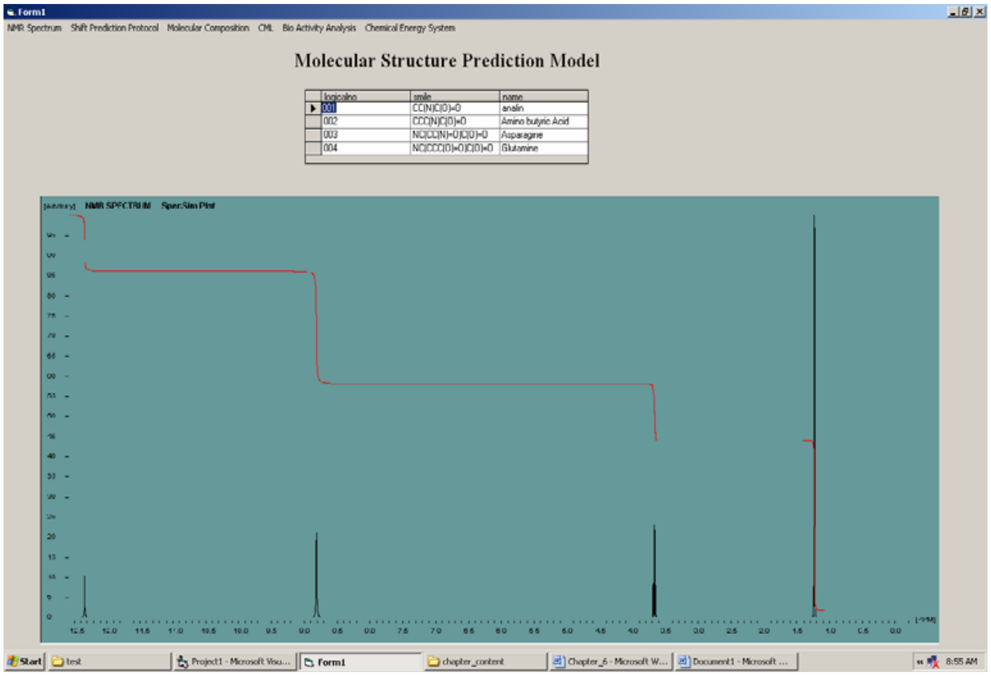


Molecular Structure Prediction Model

	logicalno	smile	name
▶	001	CC(N)C(O)=O	analin
	002	CCC(N)C(O)=O	Amino butyric Acid
	003	NC(CC(N)=O)C(O)=O	Asparagine
	004	NC(CCC(O)=O)C(O)=O	Glutamine

Figure 6.89: Basic view of Knowledge Based Model for Molecular Structure Prediction

In **Figure 6.90** different nodes of Analine like –C, CH and -CH_3 have been generated in the model in form of ^{1}H-NMR curve.

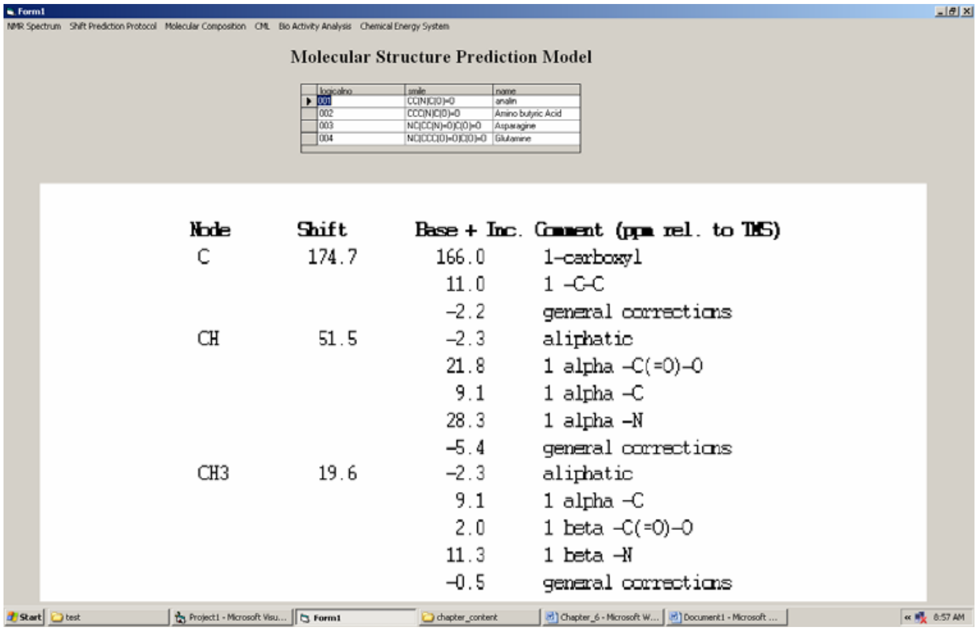


Molecular Structure Prediction Model

	logicalno	smile	name
▶	001	CC(N)C(O)=O	analin
	002	CCC(N)C(O)=O	Amino butyric Acid
	003	NC(CC(N)=O)C(O)=O	Asparagine
	004	NC(CCC(O)=O)C(O)=O	Glutamine

Figure 6.90: Nodes of Analine in ^{1}H-NMR in model

In **Figure 6.91** chemical shift values of different nodes of analine has been generated in model.



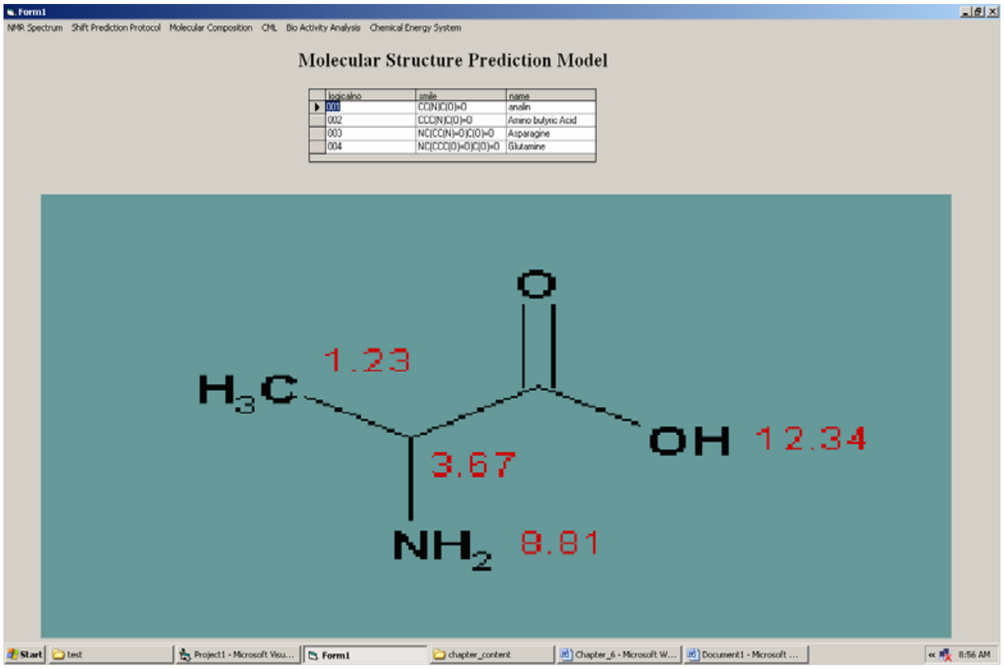
Molecular Structure Prediction Model

logicalno	smile	name
001	CC(N)C(O)=O	analin
002	CCC(N)C(O)=O	Amino butyric Acid
003	NC(CC(N)=O)C(O)=O	Asparagine
004	NC(CCC(O)=O)C(O)=O	Glutamine

```
Node     Shift      Base + Inc.  Comment (ppm rel. to TMS)
 C       174.7       166.0       1-carboxyl
                      11.0       1 -C-C
                      -2.2       general corrections
 CH       51.5        -2.3       aliphatic
                      21.8       1 alpha -C(=O)-O
                       9.1       1 alpha -C
                      28.3       1 alpha -N
                      -5.4       general corrections
 CH3      19.6        -2.3       aliphatic
                       9.1       1 alpha -C
                       2.0       1 beta -C(=O)-O
                      11.3       1 beta -N
                      -0.5       general corrections
```

Figure 6.91: Chemical shift values of different nodes of analine in model

In **Figure 6.92** bond orders of the analine has been generated in model using ^{13}C- NMR.

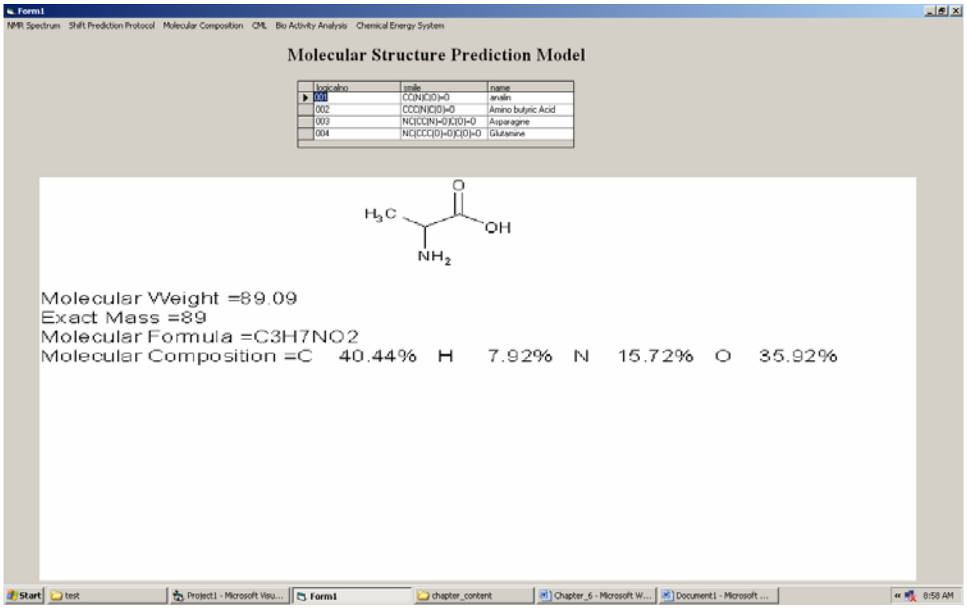


Molecular Structure Prediction Model

logicalno	smile	name
001	CC(N)C(O)=O	analin
002	CCC(N)C(O)=O	Amino butyric Acid
003	NC(CC(N)=O)C(O)=O	Asparagine
004	NC(CCC(O)=O)C(O)=O	Glutamine

Figure 6.92: Bond order of the analine in model using ^{13}C- NMR

In **Figure 6.93** molecular structure, molecular weight, molecular formula and composition of analine has been generated in model.



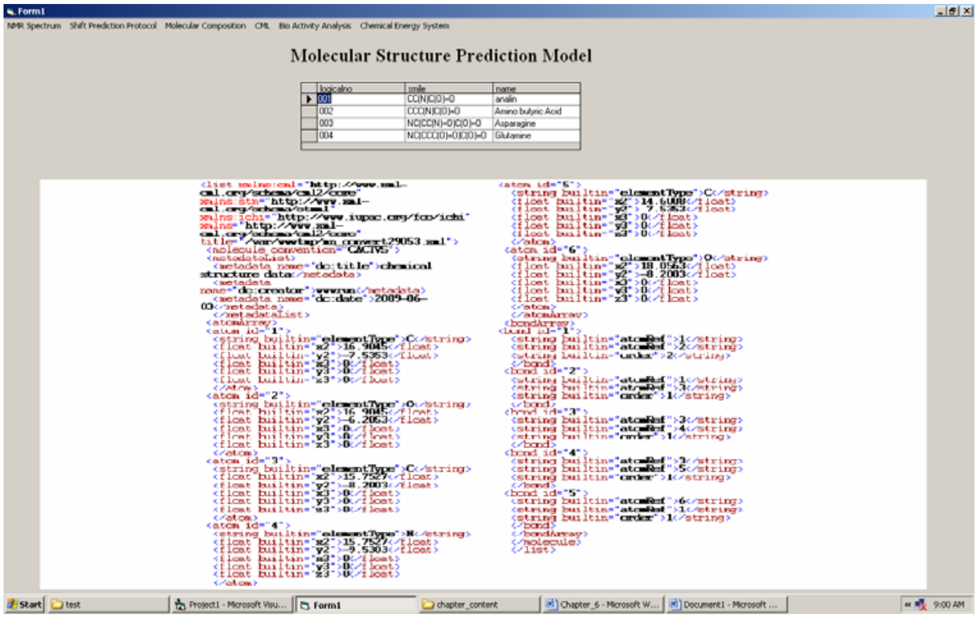
Molecular Structure Prediction Model

logicalno	smile	name
001	CC(N)C(O)=O	analin
002	CCC(N)C(O)=O	Amino butyric Acid
003	NC(CC(N)=O)C(O)=O	Asparagine
004	NC(CCC(O)=O)C(O)=O	Glutamine

Molecular Weight =89.09
Exact Mass =89
Molecular Formula =C3H7NO2
Molecular Composition =C 40.44% H 7.92% N 15.72% O 35.92%

Figure 6.93: Molecular Composition of Analine in the model

In **Figure 6.94** CML structure of analine has been generated in model.



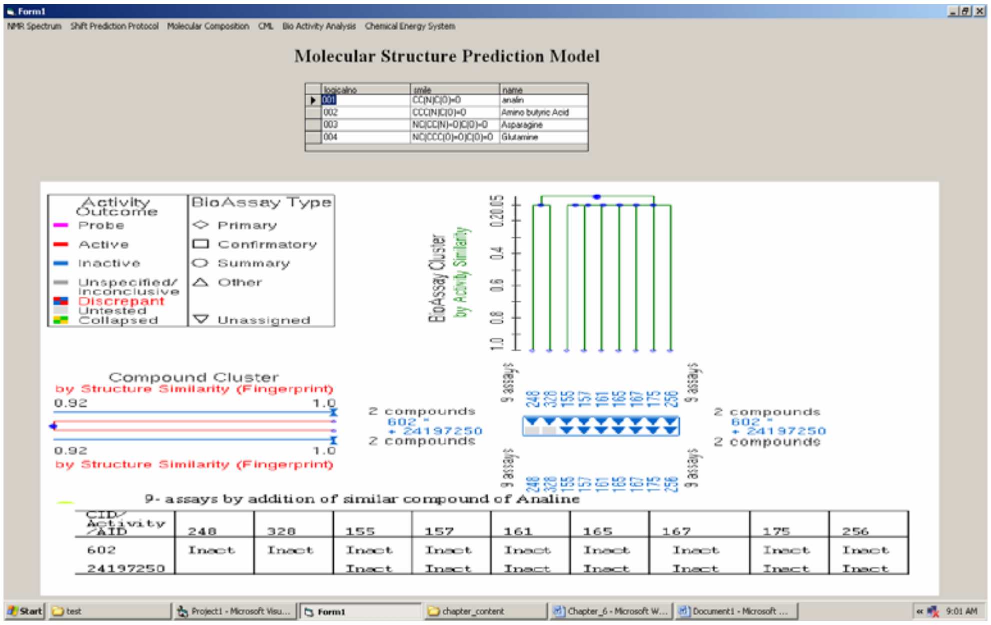
Molecular Structure Prediction Model

logicalno	smile	name
001	CC(N)C(O)=O	analin
002	CCC(N)C(O)=O	Amino butyric Acid
003	NC(CC(N)=O)C(O)=O	Asparagine
004	NC(CCC(O)=O)C(O)=O	Glutamine

```
<list xmlns:cml="http://www.xml-
cml.org/schema/cml2/core"
xmlns:stm="http://www.xml-
cml.org/schema/stmml"
xmlns:ichi="http://www.iupac.org/foo/ichi"
xmlns="http://www.xml-
cml.org/schema/cml2/core"
title="/var/wwwtmp/mn_convert29053.xml">
 <molecule convention="CACTVS">
 <metadataList>
  <metadata name="dc:title">chemical
structure data</metadata>
  <metadata
name="dc:creator">wwwrun</metadata>
  <metadata name="dc:date">2009-06-
03</metadata>
  </metadataList>
 <atomArray>
 <atom id="1">
  <string builtin="elementType">C</string>
  <float builtin="x2">16.9045</float>
  <float builtin="y2">-7.5353</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="2">
  <string builtin="elementType">O</string>
  <float builtin="x2">16.9045</float>
  <float builtin="y2">-6.2053</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="3">
  <string builtin="elementType">C</string>
  <float builtin="x2">15.7527</float>
  <float builtin="y2">-8.2003</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="4">
  <string builtin="elementType">N</string>
  <float builtin="x2">15.7527</float>
  <float builtin="y2">-9.5303</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
<atom id="5">
  <string builtin="elementType">C</string>
  <float builtin="x2">14.6008</float>
  <float builtin="y2">-7.5353</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
 <atom id="6">
  <string builtin="elementType">O</string>
  <float builtin="x2">18.0563</float>
  <float builtin="y2">-8.2003</float>
  <float builtin="x3">0</float>
  <float builtin="y3">0</float>
  <float builtin="z3">0</float>
  </atom>
  </atomArray>
 <bondArray>
<bond id="1">
  <string builtin="atomRef">1</string>
  <string builtin="atomRef">2</string>
  <string builtin="order">2</string>
  </bond>
 <bond id="2">
  <string builtin="atomRef">1</string>
  <string builtin="atomRef">3</string>
  <string builtin="order">1</string>
  </bond>
 <bond id="3">
  <string builtin="atomRef">3</string>
  <string builtin="atomRef">4</string>
  <string builtin="order">1</string>
  </bond>
 <bond id="4">
  <string builtin="atomRef">3</string>
  <string builtin="atomRef">5</string>
  <string builtin="order">1</string>
  </bond>
 <bond id="5">
  <string builtin="atomRef">6</string>
  <string builtin="atomRef">1</string>
  <string builtin="order">1</string>
  </bond>
  </bondArray>
  </molecule>
  </list>
```

Figure 6.94: CML structure of analine in the model

In **Figure 6.95** different curves like activity outcome, compound cluster have been generated in the model after addition of similar compounds of analine.



Molecular Structure Prediction Model

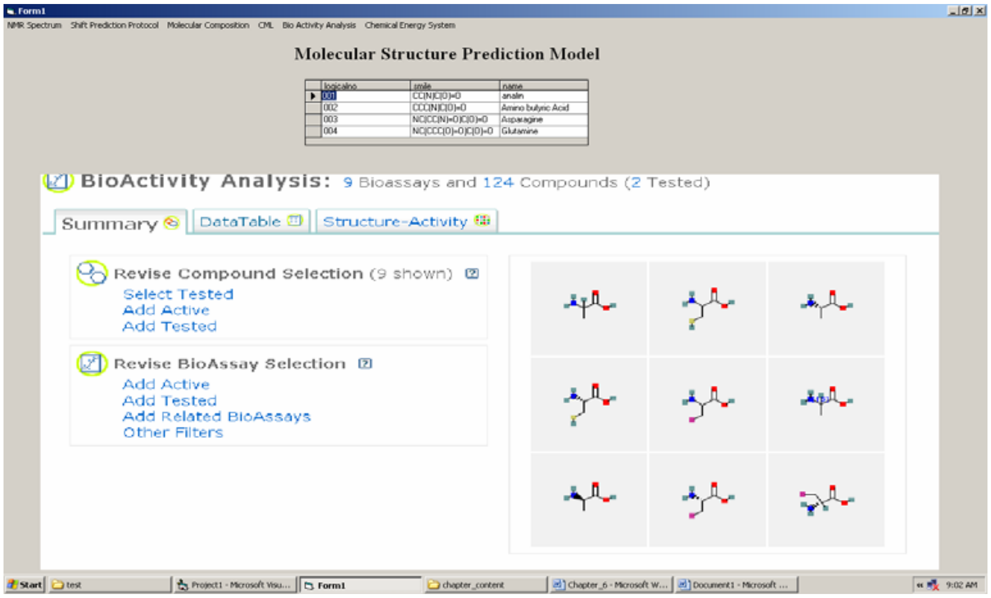
	logicalno	smile	name
▶	001	CC(N)C(O)=O	analin
	002	CCC(N)C(O)=O	Amino butyric Acid
	003	NC(CC(N)=O)C(O)=O	Asparagine
	004	NC(CCC(O)=O)C(O)=O	Glutamine

9- assays by addition of similar compound of Analine

CID/ Activity /AID	248	328	155	157	161	165	167	175	256
602	Inact	Inact	Inact	Inact	Inact	Inact	Inact	Inact	Inact
24197250			Inact	Inact	Inact	Inact	Inact	Inact	Inact

Figure 6.95: Bio activity analysis of analine after addition of similar compounds in the model

In **Figure 6.96** revised compound selection of analine has been generated in the model.



Molecular Structure Prediction Model

	logicalno	smile	name
▶	001	CC(N)C(O)=O	analin
	002	CCC(N)C(O)=O	Amino butyric Acid
	003	NC(CC(N)=O)C(O)=O	Asparagine
	004	NC(CCC(O)=O)C(O)=O	Glutamine

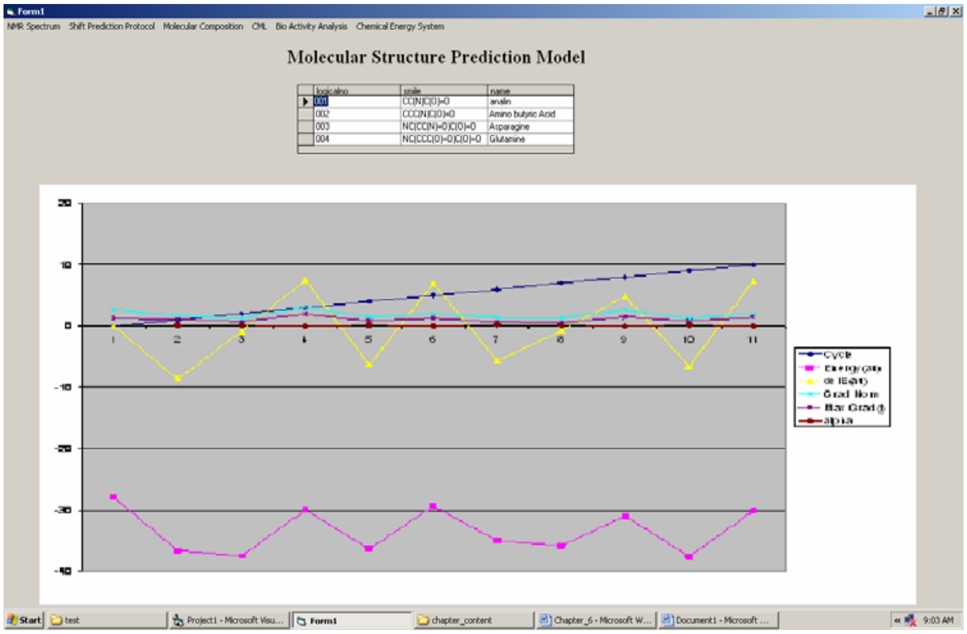
BioActivity Analysis: 9 Bioassays and 124 Compounds (2 Tested)

Revise Compound Selection (9 shown)
Select Tested
Add Active
Add Tested

Revise BioAssay Selection
Add Active
Add Tested
Add Related BioAssays
Other Filters

Figure 6.96: Revised compound selection of analine in the model

In **Figure 6.97** geometric optimization of analine for different components have generated and it ahs been shown in the model.



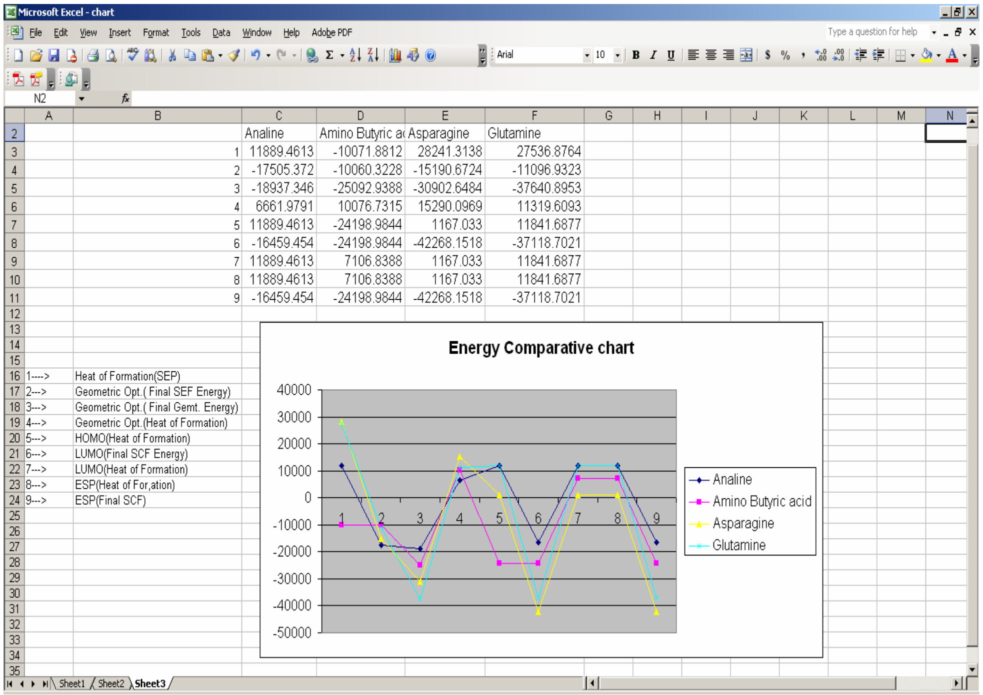
logicalno	smile	name
001	CC(N)C(O)=O	analin
002	CCC(N)C(O)=O	Amino butyric Acid
003	NC(CC(N)=O)C(O)=O	Asparagine
004	NC(CCC(O)=O)C(O)=O	Glutamine

Figure 6.97: Geometric Optimization of analine for different components in the model

In this experimental part of research work different types of energy calculations like Heat of Formation (SEP), Geometric Opt.(Final SEF Energy), Geometric Opt.(Final Gemt. Energy), Geometric Opt.(Heat of Formation), HOMO(Heat of Formation), LUMO(Final SCF Energy), LUMO(Heat of Formation), ESP(Heat of Formation) and ESP(Final SCF) on Alanine, Amino butyric Acid, Asparagine and Glutamine have performed and their result have been presented in form of graph in **Figure 6.98**. All energies values have been shown in **Table 6.39**.

Table 6.39 : Energy Comparative Table

	Analine	Amino Butyric acid	Asparagine	Glutamine
Heat of Formation(SEP)	11889.4613	-10071.8812	28241.3138	27536.8764
Geometric Opt. (Final SEF Energy)	-17505.372	-10060.3228	-15190.672	-11096.9323
Geometric Opt. (Final Gemt.Energy)	-18937.346	-25092.9388	-30902.648	-37640.8953
Geometric Opt. (Heat of Formation)	6661.9791	10076.7315	15290.0969	11319.6093
HOMO (Heat of Formation)	11889.4613	-24198.9844	1167.033	11841.6877
LUMO (Final SCF Energy)	-16459.454	-24198.9844	-42268.152	-37118.7021
LUMO (Heat of Formation)	11889.4613	7106.8388	1167.033	11841.6877
ESP (Heat of Formation)	11889.4613	7106.8388	1167.033	11841.6877
ESP(Final SCF)	-16459.454	-24198.9844	-42268.152	-37118.7021



	A	B	C	D	E	F
2			Analine	Amino Butyric ac	Asparagine	Glutamine
3		1	11889.4613	-10071.8812	28241.3138	27536.8764
4		2	-17505.372	-10060.3228	-15190.6724	-11096.9323
5		3	-18937.346	-25092.9388	-30902.6484	-37640.8953
6		4	6661.9791	10076.7315	15290.0969	11319.6093
7		5	11889.4613	-24198.9844	1167.033	11841.6877
8		6	-16459.454	-24198.9844	-42268.1518	-37118.7021
9		7	11889.4613	7106.8388	1167.033	11841.6877
10		8	11889.4613	7106.8388	1167.033	11841.6877
11		9	-16459.454	-24198.9844	-42268.1518	-37118.7021
16	1--->	Heat of Formation(SEP)				
17	2--->	Geometric Opt.(Final SEF Energy)				
18	3--->	Geometric Opt.(Final Gemt. Energy)				
19	4--->	Geometric Opt.(Heat of Formation)				
20	5--->	HOMO(Heat of Formation)				
21	6--->	LUMO(Final SCF Energy)				
22	7--->	LUMO(Heat of Formation)				
23	8--->	ESP(Heat of For,ation)				
24	9--->	ESP(Final SCF)				

Figure 6.98: Energy Comparative Chart

From this chart it has been interpreted that there has been minor increase in all types of energies due to increase in complexity of structure as complexity in structure is getting increased from aniline to glutamine.

In the last phase of experimental work of research (**Activity No-3**) various analyses on nucleotide and protein sequences have been performed by using DAMBE and Jemboss tools. The outcomes of this experimental part of research have been analyzed and evaluated. Multiple sequence alignment is an extension of pair-wise alignments. Phylogenetic trees are useful representation and method for multiple alignments. The pattern-matching approaches using scores for gaps and inexact matching are statistically valid for assessing the degree of string similarity. It is easy to rationalize the need for gaps because of computational infeasibility of solving long string comparisons without the provision for gaps. However, even a short gap in polypeptide sequence can disturb secondary and tertiary structures of protein and probably alter its function as well. Heuristics approaches, such as match matrices, attempt to add some sense of biological relevance to the mathematical equations that define the relative similarity of nucleotide and polypeptide sequences.

6.4 Conclusions and Future Scope of Research

The outcomes of this work provides an innovative platform to solve complex chemical problems such as structure elucidation, required the joint efforts of computer science and chemistry specialists. The results obtained are a good reason to expect success with novel approaches for current research challenges, as the field of chemoinformatics matures and a closer collaboration with bioinformatics is developed.

By increasing the number of stored compounds, by adding more searches and use information visualization techniques could give more insight in the analysis process. Supervised and unsupervised machine learning methods that could lead to interesting predictions for the different substructures. Use of genetic software techniques could lead to automatically design molecules.

In the extension of this research work, computational neural networks could be used to predict the mapping between protein sequence and secondary structure. By adding neural network units that detect periodicities in the input sequence, secondary structure and tertiary structure prediction accuracy could be increased.