1. INTRODUCTION

Computers serve four interdependent functions in Bioinformatics which are communications, computation, control and storage. Embedded computer controllers in sequencing machines, fermentation tanks and bireactor direct the programmable robotics arms that automatic intricate processes and markedly decrease the need of human operators. Computers are used for tasks that range from searching for nucleotide or protein sequences and visualizing protein folding patterns to simulating complex 3D protein-protein interactions, for applications ranging from drug discovery to biomaterials research and development. All these activities revolve around database technology.

1.1 Database

A database is a collection of persistent data that is used by the application systems of a given enterprise (Date 2003). Data base is an updatable storage of information of an application's world and managing software that conceals from the user the physical aspects of information storage and information representation (Naphtali Rishe 1992). The collection of data usually referred to as database contains information relevant to enterprise (Silberschatz et al. 2002) A database is an integrated collection of logically related records or files which consolidates records into a common pool of data records that provides data for many applications (Kroenke et al. 2007) . A database is a collection of information that is organized so that it can easily be accessed, managed, and updated (O'Brien et al. 1999). A data base is a shared integrated
computer storage structure that houses collection of data which are used to provide information to the users (Connolly et al. 2002). These are all the various definitions of database given in heavily referenced books. Data base management system is a collection of interrelated data and set of programs to access those data (Silberschatz et al. 2002). A Database Management System (DBMS) is a set of computer programs that controls the creation, maintenance, and the use of the database of an organization and its end users (O’Brien et al. 1999). So data base and database management systems are the two important words used in this world for getting the information. Information is the backbone for the researchers who want to do achievement.

**Biological databases**

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high throughput experiment technology, and computational analyses. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics (Altman 2004). Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures. These biological databases give lot of ideas to the researchers. Researchers collect the current and past information from these databases using internet. These biological databases organize the information on a large scale in world wide. Some of the biological databases are SWISSPROT, PROSITE, Protein Data Bank(PDB), Nucleic Acids Database(NDB), GenBank, EMBL, Entrez Genomes, Entrez etc. These databases classified based on its
contents into protein databases, macromolecular structure databases, nucleotide sequence databases, genome sequence databases and integrated databases (Bryan Bergeron 2003 and Bourne 2005). Researchers create their own databases from these web databases using database management systems like MS Access, Oracle etc. They are doing their analysis using the front end tool like Visual basic. Visual basic is a simple and efficient graphical user interface (GUI) tool used to develop software applications. These applications provide the lot of complicated ideas to the researchers.

A protein is a complex, high-molecular-mass, organic compound that consists of amino acids joined by peptide bonds. Proteins are essential to the structure and function of all living cells and viruses. Different proteins perform a wide variety of biological functions. Some proteins are enzymes, which catalyze chemical reactions. Other proteins play structural or mechanical roles, such as those that form the struts and joints of the cytoskeleton, which is like a system of scaffolding within a cell. Still more functions filled by proteins include immune response and the storage and transport of various ligands. Analyzing the protein and nucleotide sequences of various species of an organism will give the many ideas (Nelson et al. 2005 and Baxevanis et al.).

1.2 Phospholipase A2

Phospholipase A2 is an enzyme available in many living organisms. In this research work the protein and nucleotide sequences of PLA2, which are related to various snakes are used. Phospholipases A2 (PLA2s) EC 3.1.1.4 are enzymes that release fatty acids from the second carbon group of glycerol. This particular phospholipase specifically recognizes the sn-2 acyl bond of phospholipids and
catalytically hydrolyzes the bond releasing arachidonic acid and lysophospholipids. Upon downstream modification by cyclooxygenases, arachidonic acid is modified into active compounds called eicosanoids. Eicosanoids include prostaglandins and leukotrienes which are categorized as inflammatory mediators (Dennis 1983).

Phospholipase A2 represents a class of heat-stable, calcium-dependent enzymes catalyzing the hydrolysis of the 2-acyl bond of 3-n-phosphoglycerides. [This enzyme is named Phospholipase A2 to denote its 2-acyl specificity (Uthe 1971).] The following Figure 1 shows the PLA2 equation

![Figure 1 Chemical equation of PLA2](image)

Phospholipase A2 has been isolated from pancreas, snake and bee venoms. Forst et al. 1986, Chan et al. 1982, Coulard et al. 1987 and Horigome et al. 1987 report on platelet phospholipase A2. The primary structure of human pancreatic PLA2 has been reported by Verheij et al. 1983. Phospholipases are involved in lipid metabolism and are important probes of structure-function relationships in biological membranes. (Dawson 1973).

PLA2 are commonly found in mammalian tissues as well as insect and snake venom. Venom from both snakes and insects is largely composed of melittin which is a
stimulant of PLA2. Due to the increased presence and activity of PLA2 resulting from a snake or insect bite, arachidonic acid is released from the phospholipid membrane disproportionately. As a result, inflammation and pain occur at the site. There are also prokaryotic A2 phospholipases. Additional types of phospholipases include phospholipase A1, phospholipase B, phospholipase C, and phospholipase D (Lehninger 2004)

Families

Phospholipases A2 include several unrelated protein families with common enzymatic activity. Two most notable families are secreted and cytosolic phospholipases A2. Other families include Ca2+ independent PLA2 (iPLA2) and lipoprotein-associated PLA2s (lp-PLA2), also known as platelet activating factor acetylhydrolase (PAF-AH).

Secreted phospholipases A2 (sPLA2)

The extracellular forms of phospholipases A2 have been isolated from different venoms (snake, bee, and wasp), from virtually every studied mammalian tissue (including pancreas and kidney) as well as from bacteria. They require Ca2+ for activity. Pancreatic PLA2 serve for the initial digestion of phospholipid compounds in dietary fat. Venom phospholipases help to immobilize prey by promoting cell lysis.

Cytosolic phospholipases A2 (cPLA2)

The intracellular PLA2 are also Ca-dependent, but they have completely different 3D structure and significantly larger than secreted PLA2 (more than 700
residues). They include C2 domain and large catalytic domain. These phospholipases are involved in cell signaling processes, such as inflammatory response. The produced Arachidonic acid is both a signaling molecule and the precursor for other signalling molecules termed eicosanoids. These include leukotrienes and prostaglandins. Some eicosanoids are synthesized from diacylglycerol, released from the lipid bilayer by phospholipase C (see below). Phospholipases A2 can be classified based on sequence homology(Six DA et al 2000).

**Lipoprotein-associated PLA2s (lp-PLA2)**

Increased levels of lp-PLA2 are associated with cardiac disease, and may contribute to atherosclerosis (www.nature.com).

**Mechanism**

The catalytic mechanism of pancreatic sPLA2 is initiated by a His-48/Asp-99/calcium complex within the active site. The calcium ion polarizes the sn-2 carbonyl oxygen while also coordinating with a catalytic water molecule, w5. His-48 improves the nucleophilicity of the catalytic water via a bridging second water molecule, w6. It has been suggested that two water molecules are necessary to traverse the distance between the catalytic histidine and the ester. The basicity of His-48 is thought to be enhanced through hydrogen bonding with Asp-99. An asparagine substitution for His-48 maintains wild-type activity, as the amide functional group on asparagine can also function to lower the pKa, or acid dissociation constant, of the bridging water molecule. The rate limiting state is characterized as the degradation of the tetrahedral intermediate composed of a calcium coordinated oxyanion. The role of calcium can also be
duplicated by other relatively small cations like cobalt and nickel. PLA2 can also be characterized as having a channel featuring a hydrophobic wall in which hydrophobic amino acid residues such as Phe, Leu, and Tyr serve to bind the substrate. Another component of PLA2 is the seven disulfide bridges which are influential in regulation and stable protein folding (Berg et al). Due to the importance of PLA2 in inflammatory responses, regulation of the enzyme is essential. PLA2 is regulated by phosphorylation and calcium concentrations. PLA2 is phosphorylated by a MAPK at Serine-505. When phosphorylation is coupled with an influx of calcium ions, PLA2 becomes stimulated and can translocate to the membrane to begin catalysis. Phosphorylation of PLA2 may be a result of ligand binding to receptors, including: 5-HT2 receptors, mGLUR1, bFGF receptor, IFN-α receptor, IFN-γ receptor (Walter F et al. 2003)

Relevance in Neurological Disorders

In normal brain cells, PLA2 regulation accounts for a balance between arachidonic acid conversion into proinflammatory mediators and arachidonic acid reincorporation into the membrane. In the absence of strict regulation of PLA2 activity, a disproportionate amount of proinflammatory mediators are produced. The resulting induced oxidative stress and neuroinflammation is analogous to neurological diseases such as Alzheimer’s disease, epilepsy, multiple sclerosis, ischemia. Lysophospholipids are another class of molecules released from the membrane that are upstream predecessors of platelet activating factors (PAF). Abnormal levels of potent PAF are also associated with neurological damage. An optimal enzyme inhibitor would specifically target PLA2 activity on neural cell membranes already under oxidative
stress and potent inflammation. Thus, specific inhibitors of brain PLA2 could be a pharmaceutical approach to treatment of several disorders associated with neural trauma. Increase in phospholipase A2 activity is an acute phase reaction that rises during inflammation, which is also seen to be exponentially higher in low back disc herniations compared to rheumatoid arthritis. It is a mixture of inflammation and substance P that are responsible for pain. Increased phospholipase A2 has also been associated with neuropsychiatric disorders such as schizophrenia and pervasive developmental disorders (such as autism), though the mechanisms involved are not known (Farooqui et al. 2004)

1.3 Snake Database tool for Sequence Analysis

It is a unique software tool contain a database gives the details about 250 snakes in India (Malcolm Smith and Romulus Whitaker). It is also used to analyse protein and nucleotide sequences for phylogenetic analysis. Proteins are assembled from 20 amino acids using information encoded in genes. Each protein has its own unique amino acid sequence that is specified by the nucleotide sequence of the gene encoding this protein. The genetic code is a set of three-nucleotide sets called codons and each three-nucleotide combination stands for an amino acid, for example AUG stands for methionine. Because DNA contains four nucleotides, the total number of possible codons is 64; hence, there is some redundancy in the genetic code and some amino acids are specified by more than one codon. Genes encoded in DNA are first transcribed into pre-messenger RNA (mRNA) by proteins such as RNA polymerase. Most organisms then process the pre-mRNA (also known as a primary transcript) using
various forms of post-transcriptional modification to form the mature mRNA, which is then used as a template for protein synthesis by the ribosome. Analysing the primary sequences of DNA, RNA or protein is used to identify the functional, structural, or evolutionary relationships between the sequences (Nelson et al. 2005, Dobson 2000).

The protein and DNA sequences of hundreds of organisms have been decoded and stored in databases. The information is analyzed to determine genes that encode polypeptides, as well as regulatory sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematics to construct phylogenetic trees). With the growing amount of data, it long ago became impractical to analyze DNA sequences manually. Today, computer programs are used to search the genome of thousands of organisms, containing billions of nucleotides. These programs would compensate for mutations (exchanged, deleted or inserted bases) in the DNA sequence, in order to identify sequences that are related, but not identical (Aluru et al. 2006; Baxevanis et al. 2005; Durbin 1998).

Nowadays lot of software programs are available to align the sequences to find out the function and evolution. Software can analyse the sequences by pairwise alignment or multible sequence alignment. The three primary methods of producing pairwise alignments are dot-matrix methods, dynamic programming, and word methods (Mount 2004). Similarly many techniques are used for multiple sequence alignment in most of the softwares (Lipman et al. 1989, Higgins 1988, Chenna et al.)
This software has some special features compared to other softwares.

1.4 Prediction of Venomous and Nonvenomous PLA2

Phospholipase A2 is an important component found in almost all snake whether they are venomous or non-venomous. Comparison of amino acid sequences gives some ideas, how the venomous sequences varying from non-venomous sequences. Through our software anyone can identify the venomous and non venomous sequences.

Snake Venom is a complex substance. It was once believed that snake venom is either hemotoxic or neurotoxic. It is now known that venom is not simple. Most snake venom is composed of hyaluronidase, proteolytic enzymes, phospholipases, proteases, thrombin like enzymes, peptide bradykinin potentiators, polypeptide toxins, and nerve growth factor that affects the body in different ways. Thus, even a snake typically known as having only hemotoxic venom, most likely has some neurotoxic compounds as well (http://www.explore.biodiversity.com; http://en.wikipedia.org). Phospholipase A2 calcium-dependent enzyme catalyses the selective hydrolysis of the 2-acyl groups in 3-sn-phosphatidyl derivatives playing a central role in lipid metabolism and has been applied in important research in several fields of investigation. This enzyme is named as Phospholipase A2 to denote its 2-acyl specificity (Uthe et al.1971; http://www.worthington-biochem.com). Phospholipase A2 (PLA2; EC 3.1.1.4) is one of the important enzyme extracted from numerous snake venoms (Heintrikson et al.1977 and Dennis 1983). Phospholipase A2 has been found in all snake venoms thus far examined and varies from approximately 10000 to about 36000 in molecular weight (Tu 1977). Phospholipase A2 has been extracted from the venoms of various snakes.
including those belonging to the three families Elapidae, Colubridae and Viperidae (Shigeru Nishida et al. 1982).

The Cobra venom (*Naja naja naja*) has two kinds of functional sites: an activator site with minimum specificity for phosphorycholine-containing lipid and a catalytic site with little specificity for the polar group phospholipids (Adamich et al. 1979 and Roberts *et al.* 1979). The enzyme first binds a phospholipid in one of the sites, thereby binding to the lipid water interface. This enhances catalysis of the substrate (Dominique Lombardo *et al.* 1984). Cobra venom cause fatal physiological disorders, such as neuro-toxicity and haemolysis, as well as cardiac and muscular failure, many of which can be explained by the inhibition of synaptic ion channels, by the action of snake phospholipase A$_2$ and by changes in the membrane structure (Smithies 1955 and Mao *et al.* 1982). The toxins may be divided into two distinct classes according to their size: one group of 6.8-7 kDa and the other of ~8 kDa (Jingyu Shao *et al.* 1993). The apparent ubiquitous presence of phospholipase A$_2$ in snake venom has interesting implication. Although its pathological effects appear to be indirect from the evidence thus far obtained, its role in the overall toxicity of these venoms is certain. Since the protein is ubiquitous in virtually all snake venoms its primary structure proves valuable information about the evolutionary origin and mutational history of these venomous reptiles (Lourival *et al.* 1979).

The elucidation of the amino acid sequences is related to its structure and function (Shigeru Nishida *et al.* 1982). The amino acid sequence of a protein determines
the structural elements and the functional properties of the molecule. Differences in the amino acid sequences reflect one or more of the following parameters: 1) phylogenetic 2) structural 3) functional variations (Manjunatha et al. 1987). Analyzing phospholipase A2 is used to predict its functions (Pearson 1988). Such prediction makes the experimental determination of function simpler as it is in silico clearly more efficient to test an accurate prediction than to randomly test for possible functions (Ross et al. 2001).

This work describes the variations in the amino acid groups of PLA2 of various venomous species like snakes, bees and scorpions and non-venomous species like human, rabbit etc. The software tool ‘mcompu’ is used for this work.

1.5 Phylogenetic tree Analysis using Protein Sequences

In biology, evolution is the process in which some inherited traits in a population become more common relative to others through successive generations. This includes both pre-existing traits as well as new traits introduced by mutations. Over time, the processes of evolution can lead to speciation: the development of a new species from existing ones. All life is a result of such speciation events and thus all organisms are related by common descent from a single ancestor (FutuymaD.J 2005; Gould et al. 2002).

Natural selection is a key part of this process. Since some traits or collections of traits allow an organism to survive and produce more offspring than an organism
lacking them, and genes are passed on by reproduction, those that increase survival and reproductive success are more likely to be passed on in comparison to those genes that do not. Therefore, the number of organisms with these traits will tend to increase with each passing generation. Given enough time, this passive process can result in varied adaptations to changing environmental conditions. Other mechanisms of evolutionary change include genetic drift, or random changes in frequency of traits (most important when the traits are, at that time, reproductively neutral), and, at the population level, immigration from other populations can bring in new traits ("gene flow") and the founder effect, in which a small group of organisms isolated from the main population will have more of the traits of the founders for many generations after isolation, even when some of the traits are detrimental (Futuyma 2005; Haldane 1953; Landae et al. 1983).

Phylogenetic tree also called an **evolutionary tree** or a tree of life is used to (i) understand evolutionary relationships (ii) map the pathogen strain diversity for vaccines (iii) assist in epidemiology of infectious diseases or genetic defects (iv) aid in prediction of function novel genes (v) study biodiversity and (vi) understand microbial ecologies. Phylogenetic relationships among genes can help to predict which one might have similar functions (e.g. ortholog detection). It also used to follow changes occurring in rapidly changing species (e.g. HIV virus). Phylogenetic tree is a labeled tree where the internal node represents ancestral species and the leaves represent modern day species. It may be represented either (i) **Dendrogram**-broad term for the diagrammatic representation of a phylogenetic tree or (ii) **Cladogram**-represents a branching pattern,
i.e., its branch lengths do not represent time or (iii) **Phylogram**-represents number of character changes through its branch lengths or (iv) **Chronogram**-represents evolutionary time through its branch lengths (Atteson 1999; Dan Graur 2000; Daniel Huson 1999).

Traditionally morphological features are used as data to construct these trees. The basic problem in morphological phylogenetics is the assembly of a matrix representing a mapping from each of the taxa being compared to representative measurements for each of the phenotypic characteristics being used as a classifier. The types of phenotypic data used to construct this matrix depend on the taxa being compared; for individual species, they may involve measurements of average body size, lengths or sizes of particular bones or other physical features, or even behavioral manifestations. Of course, since not every possible phenotypic characteristic could be measured and encoded for analysis, the selection of which features to measure is a major inherent obstacle to the method. The decision of which traits to use as a basis for the matrix necessarily represents a hypothesis about which traits of a species or higher taxon are evolutionarily relevant (Studier et al. 1988). The inclusion of extinct taxa in morphological analysis is often difficult due to absence of or incomplete fossil records. Because morphological data is extremely labor-intensive to collect, whether from literature sources or from field observations, reuse of previously compiled data matrices is not uncommon, although this may propagate flaws in the original matrix into multiple derivative analyses (Jenner 2001).
The above mentioned problems can be solved when molecular data (DNA, RNA and protein sequences) are used as data to construct these trees (Page et al. 1998). Phylogenetic tree can be constructed using isoelectric point (PI) and molecular weight of proteins (Milner et al. 2003). Various methods are followed to construct these trees. Main methods are distance matrix methods [Neighbor joining & UPGMA], Character based methods [Parsimony method, Maximum likelihood method] and Validation methods [Boots trapping, Jack knife] (Doolittle 1996; Mount 2000).

This research work describes how the phylogenetic trees can be constructed using matrix methods [Neighbor joining & UPGMA(Unweighted Pair Group Method with Arithmetic mean)] (Nei et al. 2000, Saitou et al. 1987; Sokal et al. 1958) with the Isoelectric point (PI) and molecular weight (MW) of a protein. This study used to predict the evolution of the snakes using the developed software.

1.6 Phylogenetic tree Analysis using Nucleotide Sequences

Computational phylogenetics is the application of computational algorithms, methods and programs to phylogenetic analyses. The goal is to assemble a phylogenetic tree representing a hypothesis about the evolutionary ancestry of a set of genes, species, or other taxa. For example, these techniques have been used to explore the family tree of hominid species (Strait et al. 2004) and the relationships between specific genes shared by many types of organisms (Hodge et al. 2000). Traditional phylogenetics relies on morphological data obtained by measuring and quantifying the phenotypic properties of representative organisms, while the more recent field of molecular phylogenetics
uses nucleotide sequences encoding genes or amino acid sequences encoding proteins as the basis for classification. Many forms of molecular phylogenetics are closely related to and make extensive use of sequence alignment in constructing and refining phylogenetic trees, which are used to classify the evolutionary relationships between homologous genes represented in the genomes of divergent species. The phylogenetic trees constructed by computational methods are unlikely to perfectly reproduce the evolutionary tree that represents the historical relationships between the species being analyzed. The historical species tree may also differ from the historical tree of an individual homologous gene shared by those species

Phylogenetic trees generated by computational phylogenetics can be either rooted or unrooted depending on the input data and the algorithm used. A rooted tree is a directed graph that explicitly identifies a most recent common ancestor (MRCA), usually an imputed sequence that is not represented in the input. Genetic distance measures can be used to plot a tree with the input sequences as leaf nodes and their distances from the root proportional to their genetic distance from the hypothesized MRCA. Identification of a root usually requires the inclusion in the input data of at least one "outgroup" known to be only distantly related to the sequences of interest. By contrast, unrooted trees plot the distances and relationships between input sequences without making assumptions regarding their descent. An unrooted tree can always be produced from a rooted tree, but a root cannot usually be placed on an unrooted tree without additional data on divergence rates, such as the assumption of the molecular clock hypothesis (Mount 2004).
The set of all possible phylogenetic trees for a given group of input sequences can be conceptualized as a discretely defined multidimensional "tree space" through which search paths can be traced by optimization algorithms. Although counting the total number of trees for a nontrivial number of input sequences can be complicated by variations in the definition of a tree topology, it is always true that there are more rooted than unrooted trees for a given number of inputs and choice of parameters (Felsenstein 2004).

The data for phylogenetic tree may be either morphological or molecular. Morphological data is extremely labor-intensive to collect, whether from literature sources or from field observations, reuse of previously compiled data matrices is not uncommon, although this may propagate flaws in the original matrix into multiple derivative analyses. But these problems are solved when using nucleotides in DNA or RNA as data (Jenner 2001).

Distance-matrix methods of phylogenetic analysis explicitly rely on a measure of "genetic distance" between the sequences being classified, and therefore they require an MSA as an input. Distance is often defined as the fraction of mismatches at aligned positions, with gaps either ignored or counted as mismatches (Mount 2004). Distance methods attempt to construct an all-to-all matrix from the sequence query set describing the distance between each sequence pair. From this is constructed a phylogenetic tree that places closely related sequences under the same interior node and whose branch
lengths closely reproduce the observed distances between sequences. Distance-matrix methods may produce either rooted or unrooted trees, depending on the algorithm used to calculate them (Felsenstein 2004).

Neighbor-joining methods apply general data clustering techniques to sequence analysis using genetic distance as a clustering metric. The simple neighbor-joining method produces unrooted trees, but it does not assume a constant rate of evolution (i.e., a molecular clock) across lineages (Saitou et al. 1987; Nei et al. 1988). Its relative, UPGMA (Unweighted Pair Group Method with Arithmetic mean) produces rooted trees and requires a constant-rate assumption - that is, it assumes an ultrametric tree in which the distances from the root to every branch tip are equal (Michener 1957; Sokal et al. 1958).

This study describes how the phylogenetic trees can be constructed using matrix methods [Neighbor joining & UPGMA] with nucleotide sequences using the codon impact parameter (Smarajit Das et al. 2005). This work has done with the software tool 'mCompu'.