CHAPTER – 1

INTRODUCTION

Ever since the structure of biomolecules was unraveled in 1953, molecular biology has observed remarkable advances. With the increase in our ability to manipulate biomolecular sequences, a huge amount of data has been and is being generated as shown in Fig. 1.1. There is a need to process the information that is pouring from laboratories all over the world, so that it can be of use to further scientific developments. This need has also created entirely new problems that are interdisciplinary in nature. Scientists from the biological sciences are the creators and ultimate users of this data. However, due to sheer size and complexity, between creation and use the help of many other disciplines is required, in particular those from the mathematical and computing sciences. As a result this need has created a new field that is popularly known as computational molecular biology.

![Fig. 1.1 Growth in Medline over the past few years](image)

Fig. 1.1 Growth in Medline over the past few years
In a very broad sense computational molecular biology consists of the development and use of mathematical and computer science techniques to help solve problems in molecular biology e.g. databases are needed to store all the information that is being generated and to retrieve the desired information. Several international sequence databases already exist, but scientists have recognized the need for new database models, given the specific requirements of molecular biology. These databases should be able to record changes in the understanding of molecular sequences as these are studied. Current models are not suitable for this purpose.

The understanding of molecular sequences in turn requires new sophisticated techniques of pattern recognition and others, which are being developed by researchers in artificial intelligence. Complex statistical issues have arisen in connection with database searches, and it has required the creation of new and specific tools.

Computational biology is concerned with utilizing the capacities of computers to address problems of biological interest. It spans several classical areas such as biology, chemistry, physics, statistics and computer science, and the activities in the area are numerous. From a computational point of view the activities are ranging from algorithmic theory focusing on problems with biological relevance, construction of computational tools for specific biological problems, for experimental work where a laboratory with test tubes and microscopes is substituted with a fast computer and a hard disk full of computational tools written to analyze huge amounts of biological data to prove or disprove a certain hypothesis.
The area of computational biology is also referred to as bioinformatics (Cohen, 2004). The work presented in this thesis is concerned with an interdisciplinary area called bioinformatics.

1.1 Bioinformatics

Bioinformatics is often defined as the application of computational techniques to understand and organize the information associated with biomolecules. It is the study and analysis of biological information using computers and statistical techniques. It is the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research. Bioinformatics is taken as more of a tool than a discipline for the analysis of biological data. From information technology (IT) point of view, the bioinformatics is the use of IT in biotechnology for the data storage, data warehousing and analyzing the biomolecules sequences.

The knowledge from other branches of science and engineering are required like biology, mathematics, computer science, laws of physics & chemistry, computer science and engineering and IT to analyze biotech data. Bioinformatics is not only limited to the analysis of biological data, but in reality it is being used to solve many biological problems like various disorders in human beings and to find out how living things works.

Bioinformatics has emerged out of the need to understand the code of life, DNA (deoxyribonucleic acid). Enormous DNA sequencing projects have been evolved and added in the growth of the science of bioinformatics. The
fundamental molecule of life (DNA) directly controls the fundamental biology of life. It codes for genes (gene represents all the information that every cell in the body needs so it can grow and carry out various functionalities) which code for proteins which decide the biological makeup of humans or any living organism. The variations and errors in the genomic DNA define the possibility of developing diseases or resistance to disorders. The ultimate goal of bioinformatics is to reveal the assets of biological information hidden in the crowd of sequence, structure, literature, and other biological data and to use this information to enhance the standard of life for mankind.

1.1.1 Tasks of Bioinformatics

Different biological problems considered within the scope of bioinformatics involve the study of genes, proteins, nucleic acid structure prediction, and molecular design. A broad classification of the various bioinformatics tasks is as below:

- Alignment and comparison of DNA, RNA (ribonucleic acid), and protein sequences.
- Gene mapping on chromosomes.
- Gene finding and promoter identification from DNA sequences.
- Interpretation of gene expression and micro-array data.
- Gene regulatory network identification.
- Construction of phylogenetic trees for studying evolutionary relationship.
- DNA structure prediction.
• Structure prediction of RNA and its classes.
• Protein structure prediction and classification.
• Molecular design and molecular docking.
• Motif identification.

1.1.2 Applications of Bioinformatics

Bioinformatics has found its applications in many areas and the applications are limitless. At the same time it is not only having collaborative efforts of different people from different background but also widely using web based tools and Internet. There are constantly being new research projects and studies being done on this amazing new line of DNA analysis. Scientists are now using bioinformatics to detect genetic abnormalities in different species. This is also creating breakthroughs in the medical community. Applications of bioinformatics have allowed doctors to conduct genetic testing in unborn babies to predict and find any signs of certain genetic disorders and conditions. Some of the applications of it are as below:

• To organize data in a way that allows researchers to access existing information and to submit new entries as they are produced.
• To develop tools and resources that aid in the analysis and management of data.
• To use this data to analyze and interpret the results in a biologically meaningful manner.
• To help researchers in the pharmaceutical industry in understanding the protein structures to make the drug design easy.
• To develop efficient algorithms for the searching of sequences.

• To develop a framework for collaboration of researchers.

• Another important application of bioinformatics is the direct prediction of biomolecules three-dimensional structure from the linear sequences. It also simplifies the problem of understanding complex genomes (a complete set of DNA in a living organism) by analyzing simple organisms and then applying the same principles to more complicated ones.

• The identification of DNA profile of a person can help the investigators in identifying criminals, ascertaining family associations, protecting rare species, matching organ donors, and for security.

1.1.3 Emerging Applications

While some of the work that bioinformatics has created is under scrutiny, other areas are taking off in the world of science and discovery. The doors are opening for scientists to be able to find new applications at all times. The emerging applications of bioinformatics in the field of network security are as below:

• Recognition of Network Threats: Bioinformatics has many interesting applications outside of biology not only including automatic voice and handwriting recognition, but also in computing systems security. The tools and techniques of bioinformatics are now being applied to the problem of recognition and characterization of
computer network threats. A lot of experiments have been done on cryptography based on DNA computing.

- **Network Intrusion Detection:** There have been several researches utilizing bioinformatics techniques for host based intrusion detection systems that detect anomalous behavior on each host by monitoring sequences of user commands or sequences of system calls invoked by applications.

### 1.1.4 Challenges in Bioinformatics

In the beginning, bioinformatics was applied in the creation and maintenance of a database to store biological information. Development of this type of database involved not only design issues but the development of complex interfaces whereby researchers could both access existing data as well as submit new or revised data. Bioinformatics is a promising and innovative research field in 21st century. The key challenges to bioinformatics essentially all relate to the current flood of raw data, aggregate information, prediction of biomolecules structure and evolving knowledge arising from the study of the genome and its manifestation. Some of these challenges are described as below:

- One of the key challenges in computational biology is the prediction of three-dimensional structures from sequences and the first problem in this is that the search space of the problem is too huge because of the vast range of possible conformations and second is the primary sequence may not fully specify the tertiary structure.
• Data from biological research is proliferating rapidly and there is a need of advanced data storage and analysis methods to manage it.

• The other challenge is that the massive amounts of sequence data are produced by modern large-scale DNA sequencing efforts such as the Human Genome Project. Despite community-wide efforts in structural genomics, the output of experimentally determined structures, typically by time-consuming and relatively expensive X-ray crystallography or NMR spectroscopy, is lagging far behind the output of biomolecules sequences. So there is need for a method that can predict the secondary structure in less time than in the X-ray crystallography or NMR spectroscopy.

• The other key challenges are that a number of factors exists that make structure prediction a very difficult task. The two main problems are that the number of possible structures is extremely large, and that the physical basis of structural stability is not fully understood.

• The fundamental challenge for genomics is to determine how gene variations are linked to a certain disease and, on a broader perspective, to determine how the interactions of genes vary with environment and lifestyle.

1.2 Computational Biology

The most critical tasks in bioinformatics involve the analysis of sequence information. Computational Biology is the name specified to this process, and it involves the following:
- To find the genes in the DNA sequences of various organisms
- To develop various methods to predict its structure
- To group protein sequences into classes
- To align similar proteins and generating phylogenetic trees

Computational biology are related to molecular or evolutionary biology and focus on analyzing and comparing the composition of the key biomolecules DNA, RNA and proteins that together constitute the fundamental building blocks of organisms. Most prominently DNA sequencing method for extracting the genetic material from DNA molecules has resulted in a food of available biological data to compare and analyze (Crochemore, 1994; Wang, 1996; Waterman et al., 1997). This was one of the important successes of ongoing efforts to develop models and use techniques for getting data about the composition of these biomolecules.

1.3 Basic Concepts of Molecular Biology

In nature both living and nonliving things are found. Living things can move, reproduce, grow, and eat etc. They have an active participation in their environment as opposed to nonliving things. Yet research in the past centuries reveals that both kinds of matter are composed by the same atoms and conform to the same physical and chemical rules but there are some differences. For a long time in human history people thought that some sort of extra matter bestowed upon living beings with their active characteristics that they were "animated" by such a thing.
But nothing of the kind has ever been found. Instead, our current understanding is that living beings act the way they do due to a complex array of chemical reactions that occur inside them. These reactions never cease. It is often the case that the products of one reaction are being constantly consumed by another reaction, keeping the system going. A living organism is also constantly exchanging matter and energy with its surroundings. In contrast, anything that is in equilibrium with its surrounding can generally be considered dead. (Some notable exceptions are vegetative forms, like seeds, and viruses, which may be completely inactive for long periods of time, and are not dead.)

Modern science has shown that life started some 3.5 billions of years ago after the earth itself was formed. The first life forms were very simple, but over billions of years a continuously acting process called evolution made them evolve and diversify, so that today very complex organisms as well as very simple ones are found. Both complex and simple organisms have a similar molecular chemistry. The main actors in the chemistry of life are molecules called proteins and nucleic acids. Roughly speaking, proteins are responsible for what a living being is and does in a physical sense. (The distinguished scientist Russell Doolittle once wrote that "we are our proteins.") Nucleic acids, on the other hand, encode the information necessary to produce proteins and are responsible for passing along this "recipe" to subsequent generations.

Molecular biology research is basically devoted to the understanding of the structure and function of biomolecules. These biomolecules are therefore the fundamental objects of this study. In the following section a basic and brief description of these biomolecules is given.
1.4 Biomolecules

A biomolecule is a molecule that naturally occurs in living organisms. Biomolecules consist primarily of carbon and hydrogen, along with nitrogen, oxygen, phosphorus and sulfur. Other elements sometimes are incorporated but are much less common. All known forms of life are composed solely of biomolecules. For example, humans possess skin and hair. The main component of hair is keratin, an agglomeration of proteins which are themselves polymers built from amino acids.

Amino acids are some of the most important building blocks used in nature to construct larger biomolecules. Another type of building blocks is the nucleotides, each of which consists of three components: a purine or pyrimidine base, a pentose sugar and a phosphate group. These nucleotides mainly form the nucleic acids. Besides the polymeric biomolecules, numerous organic molecules are absorbed by living systems.

1.5 Types of Biomolecules

A diverse range of biomolecules exist e.g. Lipid, Vitamin, Hormone, Neurotransmitter, Carbohydrate, Sugar, Amino acid, Nucleotide, Phosphate, Peptide, Protein, Nucleic acids like DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). Nucleosides are molecules formed by attaching a nucleobase to a ribose ring. Examples of these include cytidine, uridine, adenosine, guanosine, thymidine and inosine. Nucleosides can be phosphorylated by specific kinases in the cell producing nucleotides which are the molecular building blocks of DNA and RNA. Living organisms contain two kinds of nucleic acids:
ribonucleic acid, abbreviated by RNA, and deoxyribonucleic acid, or DNA. The DNA is described first just to understand RNA as RNA (siRNA) is the fundamental biomolecule in this study.

1.5.1 DNA

DNA is a chain of simpler molecules. Actually it is a double chain, but let us first understand the structure of one simple chain, called strand. It has a backbone consisting of repetitions of the same basic unit. This unit is formed by a sugar molecule called 2'-deoxyribose attached to a phosphate residue. The sugar molecule contains five carbon atoms, and they are labeled 1’ through 5’ as shown in Fig. 1.2. The bond that creates the backbone is between the 3’ carbon of one unit, the phosphate residue, and the 5’ carbon of the next unit. For this reason, DNA molecules also have an orientation, which by convention, starts at the 5’ end and finishes at the 3’ end. When one finds a single stranded DNA sequence in a technical paper, book, or a sequence database file, it is always written in this canonical, 5’ → 3’ direction, unless otherwise stated.

![Fig. 1.2 Sugars present in nucleic acids.](image-url)
Attached to each 1' carbon in the backbone are other molecules called bases. There are four kinds of bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The schematic molecular structure of each base and the single DNA strand described so far are shown in Fig. 1.3 and Fig. 1.4. Bases A and G belong to a larger group of substances called purines, whereas C and T belong to the pyrimidines. If the basic unit of a DNA molecule consists of the sugar, the phosphate, and its base, then it is called a nucleotide. Thus, although bases and nucleotides are not the same thing, a DNA molecule can have 200 bases or 200 nucleotides. A DNA molecule having a few (tens of) nucleotides is referred to as an oligonucleotide. DNA molecules in nature are very long, much longer than proteins. In a human cell, DNA molecules have hundreds of millions of nucleotides.

Fig. 1.3 Nitrogenated bases present in DNA.
As already mentioned, DNA molecules are double strands. The two strands are tied together in a helical structure, the famous double helix discovered by James Watson and Francis Crick in 1953. The two strands can hold together because each base in one strand is paired with (or bonds to) a base in the other strand. Base A is always paired with base T, and C is always paired with G, as shown in Fig. 1.3 and 1.5. Bases A and T are said to be the complement of each other, or a pair of complementary bases. Similarly, C and G are complementary bases. These pairs are known as *Watson-Crick base pairs*. When referring to DNA molecules, base pairs provide the unit of length that is abbreviated as \( bp \) e.g. a certain piece of DNA is 100,000 bp long or 100 kbp.

![Fig. 1.4 A schematic molecular structure view of one DNA strand.](image1)

![Fig. 1.5 A schematic molecular structure view of a double strand of DNA.](image2)
In this thesis a biomolecule is generally considered as string of letters, each letter representing a base. Fig. 1.6 presents this "string-view" of biomolecule, showing the double strand by placing one of the strings on top of the other. Even though the strands are linked, each one preserves its own orientation, and the two orientations are opposite. Fig. 1.6 illustrates this fact. Notice that the 3' end of one strand corresponds to the 5' end of the other strand. This property is sometimes expressed by saying that the two strands are antiparallel.

![String View of DNA](image)

**Fig. 1.6** A double-stranded DNA sequence represented by strings of letters.

The fundamental consequence of this structure is that it is possible to infer the sequence of one strand given the other. The operation that enables us to do that is called reverse complementation. For example, given strand $s = AGACGT$ in the canonical direction, following is done to obtain its reverse complement: First $s$ is reversed to obtain $s' = TGCAGA$, and then each base is replaced by its complement, obtaining $s = ACGTCT$. It is precisely this mechanism that allows DNA in a cell to replicate, therefore allowing an organism that starts its life as one cell to grow into billions of other cells, each one carrying copies of the DNA molecules from the original cell.
1.5.2 RNA

RNA is, like DNA, a heteropolymer, whose chemical building blocks are very similar to the building blocks of DNA. RNA is built of four different kinds of ribonucleotides. A ribonucleotide consists of a phosphate group, a sugar, and one out of four nitrogenous bases, just like DNA. However, the chemical structure of RNA differs from DNA in two ways. RNA has a different five-carbon sugar than DNA, ribose instead of deoxyribose and the base Thymine (T) in DNA is in RNA replaced by Uracil (U) as shown in Fig. 1.2 and 1.7. The RNA molecule has a broad repertoire of functions. The most well known is as a messenger. The messenger RNA (mRNA) carries a working copy of the genetic code and is translated by the ribosome into protein. However, there are several other kinds of RNAs, the so called functional RNAs, like the transfer RNA (tRNA) and the ribosomal RNAs (rRNA) etc.

![Fig. 1.7](image)

The nitrogenous bases Adenine, Cytosine, Guanine, and Uracil of RNA.

1.5.3 Classes of RNA

RNA is not just RNA. There are many different types of RNA - some of which are listed below. Although they are similar at the molecular level, each of them has a different function.
a) **rRNA**: The ribosomal RNAs (rRNAs) is found in the ribosomes, which are the “protein factories” of an organism. The rRNAs are usually the most abundant RNAs in the cell.

b) **mRNA**: Messenger or mRNA is a copy of the information carried by a gene on the DNA. The role of mRNA is to move the information contained in DNA to the translation machinery. mRNA is heterogeneous in size and sequence.

c) **snRNA**: The small nuclear RNAs (snRNAs) are small RNA molecules located in the eukaryotic nucleus. The snRNAs are involved in various processes, such as RNA splicing and maintenance of the telomeres. They are always found associated with specific proteins and the complexes are referred to as small nuclear ribonucleoproteins (SNRNP) or sometimes as snurps.

d) **siRNA**: Small interfering RNA (siRNA) are double stranded RNAs (dsRNAs) of 20-25 nucleotides length and involved in the process termed as RNA interference (RNAi). The discovery of which was awarded the Nobel prize in Physiology and Medicine in 2006. The siRNA is included in the RNA-induced silencing complex (RISC) as single-stranded RNA and programs the RISC through complex complementarity to bind an mRNA, inducing cleavage of the mRNA at that site, which is then degraded.

The siRNAs are produced by the enzyme dicer either from exogenous or cellular double stranded RNA or from a hairpin structured RNA. Essentially any gene of which the sequence is known can thus be targeted based on sequence complementarity with an appropriately tailored siRNA. This has
made siRNAs an important tool for gene function and drug target validation studies in the post genomic era.

e) **miRNA**: The microRNAs (miRNAs) have similar functions as siRNAs, but are produced from a single stranded genomic ncRNA. The pre-miRNA is a ~70 nucleotides long hairpin structure that is processed by dicer to produce a usually 21 nucleotides long mature miRNA. The mature miRNA can in complex with RISC inhibit the translation of many mRNAs. In contrast to siRNA, the miRNA usually does not bind perfectly to its target mRNA.

f) **tRNA**: tRNA (transfer RNA) is usually very short sequences of around 100 bases in length. The transfer RNA (tRNA) is a well-studied RNA responsible for bringing the correct amino acid to the ribosome in translation.

1.6 **Roles of RNA**

RNA has many roles, some of which were known for over 50 years, but the meaning and importance of some new classes of RNA have become visible just recently. The central dogma of molecular biology describes the two-step process, transcription and translation, by which the information in genes flows into proteins as DNA → RNA → protein (shown in Fig. 1.8).

![DNA → RNA → Protein](image)

**Fig. 1.8** The central dogma describes the flow of genetic information in the cell.
The genome encodes the sequence information for all the proteins synthesized by the cell. The segment of DNA that holds the information of how to construct a certain protein is called a protein-coding gene. However, DNA is not directly the template in the protein synthesis. DNA is transcribed by an enzyme, RNA polymerase, into a messenger RNA (mRNA) carrying the same information as the transcribed gene. The mRNA is then translated into protein by the ribosome.

A gene can also be transcribed into an RNA that is never translated into a protein. Such genes are called non-coding genes and the RNAs are called functional or non-coding RNAs (messenger RNA), as they are not translated into protein, but they have a function themselves. In the meanwhile, the importance of RNAs has remained rather obscure, and RNA was mainly viewed as a passive intermediary that bridges the gap between DNA and protein.

1.7 Sequences and Structures of Biomolecules

Nucleic acid sequence may be looked upon as words over of alphabets of nucleotides. Naturally occurring DNAs and RNAs form subsets of the set of all possible words. The sequence of a biological molecule is the specification of its atomic composition represented by the alphabets e.g. A, C, T, G, U, etc. Biomolecular structure is the structure of biomolecules, mainly proteins and the nucleic acids DNA and RNA. The structure of these molecules is frequently decomposed into primary structure (that is also known as sequence), secondary structure, tertiary structure, and quaternary structure. The scaffold for this
structure is provided by secondary structural elements which are hydrogen bonds within the molecule. It also leads to several recognizable "domains" of protein structure (like alpha helices, beta sheets for proteins) and nucleic acid structure (like hairpin loops, bulges and internal loops).

The secondary structure of a nucleic acid molecule refers to the basepairing interactions within a single molecule or set of interacting molecules. The secondary structure of biological RNA's can be uniquely decomposed into stems and loops. Frequently these elements, or combinations of them, can be further classified, for example, tetraloops, pseudoknots and stem-loops. There is a minor industry of researchers attempting to determine the secondary structure of RNA molecules. Approaches include both experimental and computational methods.

Biomolecular structure prediction is the prediction of the three-dimensional structure of a nucleic acid from its base sequence. In other words, it is the prediction of secondary and tertiary structure from its primary structure. There has also been a significant amount of bioinformatics research directed at the RNA structure prediction problem. A common problem for researchers working with RNA is to determine the three-dimensional structure of the molecule given just the nucleic acid sequence. However, in the case of RNA much of the final structure is determined by the secondary structure or intramolecular base-pairing interactions of the molecule. In this thesis RNA and siRNA are generally considered as string of letters, each letter representing a base.
1.8 Modeling in Computational Molecular Biology

The use of models in biology is at once both familiar and arcane (Douglas B. Kell and Joshua D. Knowles, 1999). It is familiar because biologists presently and regularly use models as abstractions of reality. Diagrams, laws, graphs, plots and relationships, chemical formulae and so on are all essentially models of some external reality that one is trying to describe and understand. Indeed, the theories and hypotheses about biological objects and systems are in one sense also just models (Vayttaden et al., 2004). Yet the use of models is for most biologists arcane because familiarity with a subset of model types, especially quantitative mathematical models, has lain outside the mainstream during the last 50 years of the purposely reductionist and qualitative era of molecular biology.

It is largely these types of model that are an integral part of the new systems biology area. Since all such models are developed for some kind of a purpose and likely become part of the standard armory of successful biologists.

1.9 The Purpose and Implications of Modeling

The purpose of academic biological research is to allow one to understand more than one presently knows about the behavior and workings of biological systems (Klipp et al., 2005) and in due time to exploit that knowledge for agricultural, medical, commercial, or other purposes. It is considered that there are several main reasons why one would wish to make models of biological systems and processes, and one can consider each in turn. In summary, these can all be characterized as variations of simulation and prediction.
In simulation a mathematical or computational model of a system or subsystem is produced that seeks to represent or reproduce some properties that system displays. Prediction involves the production of a similar type of mathematical or computational model that predicts the behavior of a system. Simulation and prediction are thus related to each other, and the important concept of generalization describes the ability of a model derived for one purpose to predict the properties of a related system under a separate set of conditions. Thus some of the main reasons why one would wish to model a (biological) system include:

- Testing whether the model is accurate in the sense that it reflects or can be made to reflect known experimental facts.
- Analyzing the model to understand which parts of the system contribute most to some desired properties of interest.
- Hypothesis generation and testing allow one rapidly to analyze the effects of manipulating experimental conditions in the model without having to perform complex and costly experiments.
- Testing what changes in the model would improve the consistency of its behavior with experimental observations.

### 1.10 Modeling Process

The Fig. 1.9 outlines the steps involved in the development, evaluation and refinement of a model. The first, essential, and most frequently overlooked step in modeling is to decide exactly what the model is for. Anybody cannot ask
models to be literally true, but one can insist that they should be useful, and usefulness is measured against one’s objectives and the value of those objectives. One important aspect of setting objectives is to decide where they fall on the continuum between theoretical and practical modeling i.e. whether one uses the model to understand the system and interpret observations of its behavior, or to predict the system, running either its own or with outside interventions.

![Modeling Process Diagram](image)

Fig. 1.9 Modeling Process
Another important decision is how much numerical accuracy one needs. Accurate prediction is often the primary goal in practical applications. But if theoretical understanding is the major goal, it may be good enough if the model gets the sign right or in some other way gives a reasonable qualitative match. The next step is to assess the feasibility of one’s goals. The most common constraints are time and data. Some pessimism about time requirement is usually a good idea, especially for beginners. It is usually a good idea to start with a small project that can later be expanded to a more complete model, or a simple model to which detail can be added later.

In contrast, assessment of whether the available data will meet one’s needs should be optimistic. Beginners frequently decide that a project cannot be done because some ‘crucial’ piece of information is missing. But models often have several parameters or assumptions that have little or no impact on relevant aspects of model behavior. The only way to find out if the data one is missing are actually needed is to build the model, and then do a sensitivity analysis to find out which parts really matter. If one seems to have most of the data that one need, the odds are good that one or an experienced advisor can find some way of working around the gaps.

1.11 Importance of Natural Language Processing (NLP) in Bioinformatics

Natural language processing (NLP) is the processing, or treatment by computer, of natural language, i.e., human languages, as opposed to programming languages. The steps in NLP are morphological analysis, syntactic analysis, semantic analysis, discourse integration, and pragmatic analysis. These
are the similar steps that one uses to process the computer languages. There are strong parallels between language and biological data affording the development and use of common tools as shown in Fig 1.10. Some language technology tools have been highly successful. Other language technology tools are promising. New field of computational sequence analysis e.g. computational bio-linguistics, sequence-to-structure technology are also emerging. Applications of NLP tools in biology are the constructions of grammars for biomolecule’s sequences, to predict function from sequences, and information extraction from scientific literature.

**Fig. 1.10** Comparison between Biology and NLP

The biological field has long been a target of interest for NLP community. However, the biological community has yet to adopt NLP techniques in a wide spread way. Biologists are gaining computational tools and NLP researchers are taking advantage of vast and well-curated data from biologists. NLP can offer
techniques for handling large set of literature available in biological field and also biomolecules sequence can be used for training and evaluation of NLP systems.

One of the significant NLP applications in bioinformatics is to predict structure and functionalities of biomolecules. A key theoretical principle in NLP for understanding an unknown language is the recognition of its syntactic patterns. Biologists wishes to find the biological meaning of some portions of biomolecular sequences. For biomolecules, these patterns might have similarities either in sequence or in structure or both (Betty Yee Man Cheng et al., 2004).

A particular biological application can be implemented using a wide range of different NLP techniques. Similarly a single NLP technique may be used for a variety of application in biological field. NLP applications in bioinformatics consist of automated searching of literature, defining grammar for sequences and structure of biomolecules, and enhancement of homology search. There are some tasks where NLP tools are good like for detecting terms such as names of proteins, drug and disease.

The statistical tools and techniques have also been integrated with NLP to explain the relationship between these terms in a probabilistic way that provides great flexibility to the system. In the recent past it was observed that a growing interest of computer professionals in the field of molecular biology has resulted into development of computational techniques and models for experimental data generated in laboratories. Finite automata, formal grammar and molecular computing can be viewed as a hope for reducing the huge gap between laboratory experiments and computational models.
1.12 Need of the Study

The purpose of biological sequence analysis is to analyze genetic information from sequence data by using methods in informatics and statistics, including gene finding, homology searches and structure prediction. In particular, analyzing the structure of a biomolecule leads to the elucidation of its function since it is empirically known that if the structure of one molecule is similar to that of another, both functions will be similar. The prediction of RNA secondary structure is a long standing challenge in the field of bioinformatics.

The problem of RNA secondary structure prediction is to derive a two-dimensional structure from the information given by the one-dimensional RNA sequence. The secondary structure is a summary of the most important aspects in the full three-dimensional structure - namely a list of base pairings. Diverse computational methods have been developed and used with the aim of solving this problem. The problem becomes even more difficult when the classes of RNA like mRNA, tRNA and siRNA are considered for structure prediction and are therefore important to model.

It is always fascinating when a problem has been analysed for a long period of time and no definite method for solving the problem has proven to be the best. It makes it even more interesting when methods very different in nature have evolved to try to tackle the same problem. This motivates us to do this study to predict the structures from biological sequences.


1.13  **Research Objectives**

The present study attempts to investigate the significance of NLP techniques specially CFGs (context free grammars and its variations) in biological domain and the development of the model using this technique for sequences and structure of biomolecules. The study also aims for predicting the sequences and structure of biomolecules using the developed model. Efforts will also be made to identify and find the important characteristics of biomolecules. Suitable qualitative and quantitative techniques will be employed to evaluate and compare the developed model.

1.14  **Scope of the Present Study**

The results derived from this study will provide a better understanding of the significance of NLP techniques in the development of model for the prediction of biomolecules. The findings will enrich the existing approaches and assist the scientists and researchers in the analysis of sequences and structures of biomolecules.

1.15  **Overview of Dissertation**

The dissertation has been organized into six chapters. A brief description of the content of these chapters is given below:

**Chapter - 1: Introduction**

This chapter gives an overview of computational biology and bioinformatics. The basic concepts of molecular biology are also highlighted. This chapter also outlines the biomolecules and its types e.g. DNA, RNA along
with the difference between DNA and RNA, classes of RNA (like mRNA, siRNA), sequences and structure of biomolecules, and modeling in computational molecular biology, and the importance of NLP in bioinformatics. In the end, the objectives and scope of the research work are presented.

Chapter - 2: Literature Review

The chapter reports the work carried out by different researchers regarding the use of NLP techniques in finding and prediction of biomolecules. The chapter describes the various methods and models used for the predictions of structure of biomolecules. The sequence and structures predicted by the methods and models have been analyzed in detail. It is intended to give an overview of the literature in order to see the SCFG approach in the context of other methods for siRNA secondary structure prediction. On the basis of literature presented in the chapter, the existing gaps in the literature regarding the use of grammar models for prediction of biomolecules have been identified. The chapter ends with formulation of the problem.

Chapter - 3: Grammar Based Model for Biomolecules

This chapter focuses on grammar based models for secondary structure - both a biological model and a computational model. The chapter also defines key aspects and notation for the computational model, which will be used throughout the rest of the thesis. The chapter describes the methodology adopted to evolve the computational model for prediction and identification of biomolecules. The chapter also includes sections pertaining to the discussions of siRNA sequences
and its structural elements, the context free grammars, and stochastic context free grammars. The chapter also describes the step by step approach followed during development of model. It also introduces the theory of formal languages, context-free grammars, SCFG and the probabilistic aspects related to the later.

Chapter - 4: Development of Model for siRNA

This chapter links the theory of SCFG described in the previous chapter with the problem of siRNA secondary structure prediction as described in chapter 3 to develop the model for structure prediction. The chapter also presents some of the algorithms that operate on SCFGs along with the discussion of the training of SCFGs.

Chapter - 5: Refinement of Developed Model for siRNA

This chapter discusses various techniques to refine the developed model. It also highlights some specific grammars that are made for the purpose of secondary structure prediction. The chapter also describes the implementation of a CYK algorithm using BioPerl programming language. The chapter also examines practical issues such as handling small probabilities on a computer. It also focuses on the results obtained by the model on siRNA sequences. Furthermore, the chapter discusses various measures for comparing two secondary structures, making it possible to test the accuracy of predictions.
Chapter - 6: Conclusions and Scope for Future Work

The salient conclusions drawn from the current study are presented in this chapter. In addition, the chapter also highlights the recommendations for future research work.