APPLICATION OF SIGNAL PROCESSING AND SOFT COMPUTING TO GENOMICS

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Bachelor of technology in 'Electronics and Communication'

BY

Chinmaya Mahapatra

SUPERVISOR

Professor Ganapati Panda, FNAE, FNASc

ELECTRONICS AND COMMUNICATION ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
ROURKELA
INDIA

CERTIFICATE

*This is to certify that the thesis entitled "**Application Of Signal Processing And Soft Computing To Genomics**" by Mr. Chinmaya Mahapatra, submitted to the National Institute of Technology, Rourkela for the award of Bachelor of Technology in Electronics and Communication Engineering, is a record of bona fide research work carried out by him in the department of Electronics and Communication Engineering, National Institute of Technology, Rourkela under my supervision. I believe that this thesis fulfills part of the requirements for the award of degree of Bachelor of Technology. The results embodied in the thesis have not been submitted for the award of any other degree.*

Prof. G. Panda, FNAE, FNASc.
Department of ECE
National Institute of Technology
Rourkela- 769008

ACKNOWLEDGEMENT

On the submission of my Thesis report of "**APPLICATION OF SIGNAL PROCESSING AND SOFT COMPUTING TO GENOMICS**", I would like to extend my gratitude & my sincere thanks to my supervisor Prof. G. Panda, Professor, Department of Electronics and communication Engineering for his constant motivation and support during the course of my work in the last one year. I truly appreciate and value his esteemed guidance and encouragement from the beginning to the end of this thesis. His knowledge and company at the time of crisis would be remembered lifelong. I want to thank all my teachers Prof. G.S Rath, Prof. S.K. Patra , Dr S.Meher, Prof.S.k.Behera, Prof.D.P.Acharya for providing a solid background for my studies and research thereafter. They have been great sources of inspiration to me and I thank them from the bottom of my heart. I would also like to thank Shri. Sitanshu Sekhar Sahoo and Shri Jagannath Nanda for their guidance and support. I will be failing in my duty if I do not mention the laboratory staff and administrative staff of this department for their timely help.

I would like to thank all whose direct and indirect support helped me completing my thesis in time. I would like to thank all those who made my stay in Rourkela an unforgettable and rewarding experience. Last but not least I would like to thank my parents, who taught me the value of hard work by their own example. I would like to share this moment of happiness with my father and mother. They rendered me enormous support during the whole tenure of my stay in NIT Rourkela.

Chinmaya Mahapatra

CONTENTS

Abstract

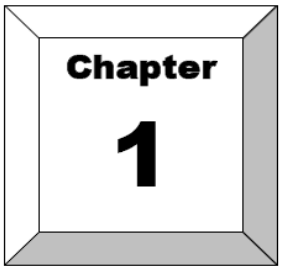
A major challenge for genomic research is to establish a relationship among sequences, structures and function of genes. In addition processing and analyzing this information are of prime importance. Basically genes are repositories for protein coding information and proteins in turn are responsible for most of the important biological functions in all cells. These in turn gives rise to analysis of DNA sequences in proteins, designing of various drugs for genetic diseases.

This thesis deals with the applications of signal processing and soft computing algorithms to the field of genomics and proteinomics. Diseases like SARS and Migraine have been modeled using these tools and potential druggable compounds have been proposed which are better than the previous available drugs. Protein structural classes have been identified more accurately based on Genetic Algorithm and Particle Swarm Optimization.

Better and efficient methods like Sliding-DFT and Adaptive AR Modeling were proposed to identify Protein coding regions in genes. The proposed methods showed better results as compared to existing methods.

LIST OF FIGURES

LIST OF TABLES



INTRODUCTION

1.1BACKGROUND

Genomic signal processing (GSP) is the engineering discipline that studies the processing of genomic signals, by which we mean the measurable events, principally the production of mRNA and protein that are carried out by the genome. Based upon current technology, GSP primarily deals with extracting information from gene expression measurements. The analysis, processing, and use of genomic signals for gaining biological knowledge constitute the domain of GSP. The aim of GSP is to integrate the theory and methods of signal processing with the global understanding of functional genomics, with special emphasis on genomic regulation. Hence, GSP encompasses various methodologies concerning expression profiles: detection, prediction, classification, control, and statistical and dynamical modeling of gene networks. GSP is a fundamental discipline that brings to genomics the structural model-based analysis and synthesis that form the basis of mathematically rigorous engineering. Signals generated by the genome must be processed to characterize their regulatory effects and their relationship to changes at both the genotypic and phenotypic levels. Application is generally directed towards tissue classification and the discovery of signaling pathways. Because transcriptional control is accomplished by a complex method that interprets a variety of inputs, the development of analytical tools that detect multivariate influences on decision making present in complex genetic networks is essential. To carry out such an analysis, one needs appropriate analytical methodologies. Perhaps the most salient aspect of GSP is that it is an engineering discipline, having strong roots in signals and systems theory. In GSP, the point of departure is that the living cell is a system in which many interacting components work together to give rise to execution of normal cellular functions, complex behavior, and interaction with the environment, including other cells. In such systems, the “whole” is often more than the “sum of its parts,” frequently referred to as emergent or complex behavior. The collective behavior of all relevant components

in a cell, such as genes and their products, follows a similar paradigm, but gives rise to much richer behavior, that is characteristic of living systems. To gain insight into the behavior of such systems, a systems-wide approach must be taken. This requires us to produce a model of the components and their interactions and apply mathematical, statistical, or simulation tools to understand its behavior, especially as it relates to experimental data.

1.2RESEARCH OBJECTIVE

The fundamental research objectives of this thesis are:

- *Time And Cost Effectiveness Of In-Silico Drug Discovery Process*

The following flow chart shows the cost and time involved in a particular drug discovery process

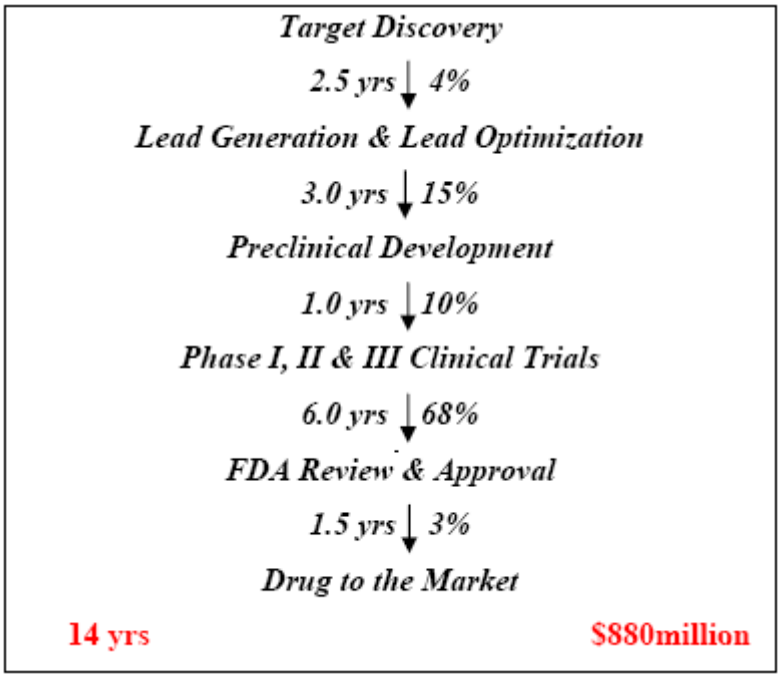


Fig 1.1 Cost and Time involved in Drug Discovery

As structures of more and more protein targets become available through crystallography, NMR and bioinformatics methods, there is an increasing demand for computational tools that can identify and analyze active sites and suggest potential drug molecules that can bind to these sites specifically. Also to combat life-threatening diseases such as AIDS, Tuberculosis, Malaria etc., a global push is essential. Millions for Viagra and pennies for the diseases of the

poor is the current situation of investment in Pharma R&D. Time and cost required for designing a new drug are immense and at an unacceptable level. According to some estimates it costs about $880 million and 14 years of research to develop a new drug before it is introduced in the market intervention of computers at some plausible steps is imperative to bring down the cost and time required in the drug discovery process

- *Increasing Importance Of Protein Structure Prediction*

It is a critical challenge to develop automated methods for fast and accurately determining the structures of proteins Because, of the increasingly widening gap between the number of sequence known proteins and that of structure known proteins in the post-genomic age. The knowledge of protein structural class can provide useful information towards the determination of protein structure. Thus, it is highly desirable to develop computational methods for identifying the structural classes of newly found proteins based on their primary sequence. The structural class has become one of the most important features for characterizing the overall folding type of a protein and it has played as important role in molecular biology, pharmacology, rational drug design and many other applications.

- *Effective Method of study of genetic material present in DNA*

The genomic information is present in the strands of DNA and represented by nucleotide symbols (A, T, C and G).The segments of DNA molecule called gene is responsible for protein synthesis and contains code for protein in exon regions within it. When a particular instruction becomes active in a cell, the corresponding gene is turned on and the DNA is converted to RNA and then to protein by slicing up to exons (protein coding regions of gene). Therefore finding coding regions in a DNA strand involves searching of many nucleotides which constitute the DNA strand. Thus the DNA transcription and replication process can be decoded in various organisms including human beings and virus which will help in study of these organisms and will also assert in making effective medicine for different diseases.

1.3 THESIS CONTRIBUTION

This section outlines the major contribution of the thesis. The work done in this thesis has been divided into three chapters.

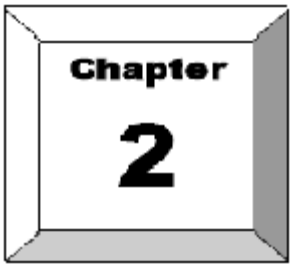
The first chapter includes the drug designing applications of Soft Computing and Signal processing. It consists of two parts of work. In the first work, a novel low complexity functional link neural network (FLANN) method is proposed to design inhibitors against the SARS virus (SARS 3CL protease) through the quantitative structure activity relationship (QSAR) study of the known twenty eight druggable compounds. The binding preferences as well as the hydrogen bonds contributing to the interaction between ligand and the SARS protein receptor are studied. Based on this, eight compounds are found promising and the main structural features shared by these are used to design analogues. Among the analogues the best performing eleven ligands are retrieved and put forward for QSAR analysis. Simulation study reveals that the proposed FLANN based QSAR model outperforms the existing artificial neural network (ANN) based approaches. The new model provides results which are in excellent agreement with the experimentally obtained values (r^2=0.9900, q^2=0.9900, MSE=0.0001). Using the new model the biological activities of the eleven modified compounds are predicted and subsequently screened via docking techniques. It is observed that three predicted SARS 3CL protease inhibitors are highly active and possess better calculated binding energies compared to those provided by experimentally known drug candidates.

In the second work, we analyzed the performance of a real coded "steady-state" genetic algorithm (SSGA) using a grid-based methodology in docking 16 migraine protein-ligand complexes having known three-dimensional structures. The SSGA was tested for the rigid and flexible ligand docking cases. The final complexes were ranked according to their minimum binding energies and were better with respect to previously used drugs for migraine.

In the second chapter, a protein is expressed as a vector of 20-dimensional space, in which its components are defined by the composition of its 20 amino acids. From these values the structural class predictions for each protein are derived. Examples consisting of 204 previously classified proteins are taken and then each sequence is normalized with respect to the 20-amino

acids by dividing by their lengths. Then the structural classes were predicted using the Genetic algorithm (GA) and Particle Swarm Optimization (PSO) approach. This method was compared with the previously used Euclidean distance, Hamming distance, AAPCAb (amino acid Principal Component Analysis), Logit-boost and Support vector machine(SVM) method for prediction of structural class of proteins. In comparison with the existing methods, the new method yields a higher accuracy of prediction.

The third chapter presents two new efficient approaches as based on sliding DFT (SDFT), adaptive AR modeling for identification of coding region exploring the period-3 behavior. The performance of the new methods is shown to be identical to that obtained by the Fourier transform based method. In addition, the proposed method offers substantial computational advantage over the conventional methods. Thus in general the four novel methods proposed have distinct computational advantage without sacrificing the quality of gene and exon prediction and can be applied for rapid medical applications.



IN-SILICO DRUG DESIGN

2.1INTRODUCTION

Drug design and drug discovery are of critical importance in human life. Its success rate depends on in depth study and use of both biological and chemical research employing computational approaches. In this chapter a novel low complexity functional link neural network (FLANN) method is proposed to design inhibitors against the SARS virus (SARS 3CL protease) through the quantitative structure activity relationship (QSAR) study of the known twenty eight druggable compounds. The binding preferences as well as the hydrogen bonds contributing to the interaction between ligand and the SARS protein receptor are studied. Based on this, eight compounds are found promising and the main structural features shared by these are used to design analogues. Among the analogues the best performing eleven ligands are retrieved and put forward for QSAR analysis. Simulation study reveals that the proposed FLANN based QSAR model outperforms the existing artificial neural network (ANN) based approaches. The new model provides results which are in excellent agreement with the experimentally obtained values (r^2=0.9900, q^2=0.9900, MSE=0.0001). Using the new model the biological activities of the eleven modified compounds are predicted and subsequently screened via docking techniques. It is observed that three predicted SARS 3CL protease inhibitors are highly active and possess better calculated binding energies compared to those provided by experimentally known drug candidates.

Migraine is basically a neurological disorder. Migraine headache is a severe pain that is typically on one side of the head but sometimes on both sides and can occur at any time of day and can last a few hours or up to one or two days. Attacks can be very intense, forcing the sufferer to

abandon normal daily activities. This chapter gives a comprehensive pathway for lead molecule design for migraine focusing on the emerging in silico trends and techniques which includes generation of candidate molecules, checking them for their toxicity and human body likeliness, docking them with the target and ranking them based on their binding affinities. After literature survey of recent journals it has been found that elevated levels of human receptor activity-modifying protein-1 (hRAMPI) (2YX8.PDB) is one of the major factors contributing to migraine attacks. First this protein sequence is taken from PDB. Its missing side chain prediction was performed using sqwrl and the energy minimization was done in chimera using AMBER force field. Then blind docking was performed with some of the available drugs in autodock and Sumatriptan was found to be the best ligand as it had the lowest binding energy. Pharmacophore for Sumatritan is analyzed using the ligandscout and its 15 derivatives are drawn with minimized energy in PRODRG using the GROMOS 96.1 force field. The derivatives were analyzed for drug-likeliness using Lipinski filters and toxicity using ADME/Tox filter. Protein-ligand interactions (docking) for all the derivatives with the protein was found out using Autodock 4.0 and hex 4.0. By the comparative study of the binding energies of all the complexes thus formed, three of the best ligands were chosen and were analyzed for active amino acid groups mainly involved in ligand –protein interaction using ligandscout 2.0. The amino acid groups ALA70A and ASP90A are found to be involved in binding interaction.

2.2 IN-SILICO DRUG TARGET DESIGN AND ITS APPROACH

2.2.1 DRUG

A **drug**, is any chemical substance that, when absorbed into the body of a living organism, alters normal bodily function.

2.2.2 DRUG DESIGN BASED ON BIOINFORMATICS TOOLS

Drug design is the approach of finding drugs by design, based on their biological targets. Typically a drug target is a key molecule involved in a particular metabolic or signaling pathway

that is specific to a disease condition or pathology, or to the infectivity or survival of a microbial pathogen. Some approaches attempt to stop the functioning of the pathway in the diseased state by causing a key molecule to stop functioning. Drugs may be designed that bind to the active region and inhibit this key molecule. However these drugs would also have to be designed in such a way as not to affect any other important molecules that may be similar in appearance to the key molecules. Sequence homologies are often used to identify such risks. Other approaches may be to enhance the normal pathway by promoting specific molecules in the normal pathways that may have been affected in the diseased state.

The structure of the drug molecule that can specifically interact with the biomolecules can be modeled using computational tools. These tools can allow a drug molecule to be constructed within the biomolecule using knowledge of its structure and the nature of its active site. Construction of the drug molecule can be made inside out or outside in depending on whether the core or the R-groups are chosen first. However many of these approaches are plagued by the practical problems of chemical synthesis. Newer approaches have also suggested the use of drug molecules that are large and proteinaceous in nature rather than as small molecules. There have also been suggestions to make these using mRNA. Gene silencing may also have therapeutical applications. The processes of designing a new drug using bioinformatics tools have open a new area of research. However, computational techniques assist one in searching drug target and in designing drug in silico, but it takes long time and money. In order to design a new drug one need to follow the following path.

- *Identify Target Disease:*

One needs to know all about the disease and existing or traditional remedies. It is also important to look at very similar afflictions and their known treatments. Target identification alone is not sufficient in order to achieve a successful treatment of a disease. A real drug needs to be developed. This drug must influence the target protein in such a way that it does not interfere with normal metabolism. One way to achieve this is to block activity of the protein with a small

molecule. Bioinformatics methods have been developed to virtually screen the target for compounds that bind and inhibit the protein. Another possibility is to find other proteins that regulate the activity of the target by binding and forming a complex.

- *Study Interesting Compounds:*

One needs to identify and study the lead compounds that have some activity against a disease. These may be only marginally useful and may have severe side effects. These compounds provide a starting point for refinement of the chemical structures.

- *Detect the Molecular Bases for Disease:*

If it is known that a drug must bind to a particular spot on a particular protein or nucleotide then a drug can be tailor made to bind at that site. This is often modeled computationally using any of several different techniques. Traditionally, the primary way of determining what compounds would be tested computationally was provided by the researchers' understanding of molecular interactions. A second method is the brute force testing of large numbers of compounds from a database of available structures.

- *Rational drug design techniques:*

Unlike the historical method of drug discovery, by trial-and-error testing of chemical substances on cultured cells or animals, and matching the apparent effects to treatments, rational drug design begins with a knowledge of specific chemical responses in the body or target organism, and tailoring combinations of these to fit a treatment profile. Due to the complexity of the drug design process two terms of interest are still serendipity and bounded rationality. Those challenges are caused by the large chemical space describing potential new drugs without side-effects.

A particular example of rational drug design involves the use of three-dimensional information about biomolecules obtained from such techniques as x-ray crystallography and NMR spectroscopy. This approach to drug discovery is sometimes referred to as structure-based drug design. The first unequivocal example of the application of structure-based drug design leading to an approved drug is the carbonic anhydrase inhibitor dorzolamide which was approved in 1995.

Another important case study in rational drug design is imatinib, a tyrosine kinase inhibitor designed specifically for the bcr-abl fusion protein that is characteristic for Philadelphia chromosome-positive leukemias (chronic myelogenous leukemia and occasionally acute lymphocytic leukemia). Imatinib is substantially different from previous drugs for cancer, as most agents of chemotherapy simply target rapidly dividing cells, not differentiating between cancer cells and other tissues.

The activity of a drug at its binding site is one part of the design. Another to take into account is the molecule's drug likeness, which summarizes the necessary physical properties for effective absorption. One way of estimating drug likeness is Lipinski's Rule of Five.

- *Refinement of compounds:*

Once you got a number of lead compounds have been found, computational and laboratory techniques have been very successful in refining the molecular structures to give a greater drug activity and fewer side effects. This is done both in the laboratory and computationally by examining the molecular structures to determine which aspects are responsible for both the drug activity and the side effects.

- *Molecular Docking*

In the field of molecular modeling, docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex. Knowledge of

the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using scoring functions. The associations between biologically relevant molecules such as proteins, nucleic acids, carbohydrates, and lipids play a central role in signal transduction. Furthermore, the relative orientation of the two interacting partners may affect the type of signal produced (e.g., agonism vs. antagonism).Therefore docking is useful for predicting both the strength and type of signal produced. Most popular approach for structure based approach.

- *Solubility of Molecule:*

One need to check whether the target molecule is water soluble or readily soluble in fatty tissue will affect what part of the body it becomes concentrated in. The ability to get a drug to the correct part of the body is an important factor in its potency. Ideally there is a continual exchange of information between the researchers doing QSAR studies, synthesis and testing. These techniques are frequently used and often very successful since they do not rely on knowing the biological basis of the disease which can be very difficult to determine.

- *Drug Testing:*

Once a drug has been shown to be effective by an initial assay technique, much more testing must be done before it can be given to human patients. Animal testing is the primary type of testing at this stage. Eventually, the compounds, which are deemed suitable at this stage, are sent on to clinical trials. In the clinical trials, additional side effects may be found and human dosages are determined.

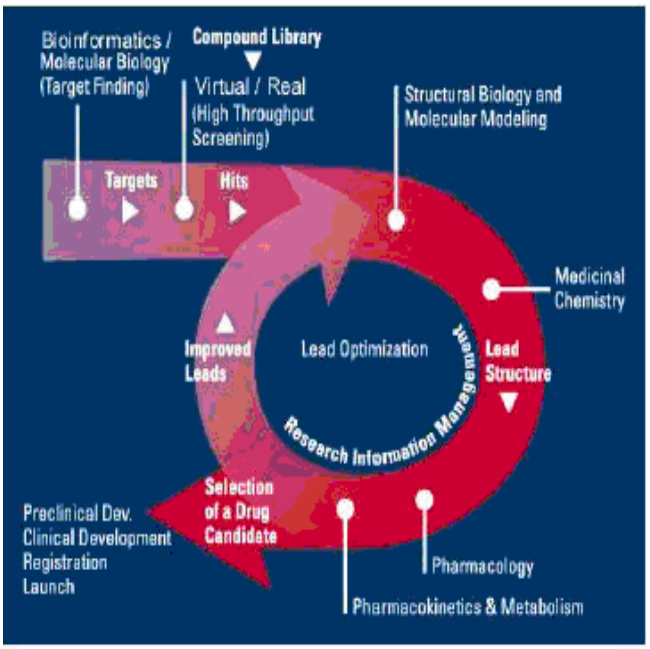


Fig 2.1 Steps in Drug Target Design

2.3 SEVERE ACUTE RESPIRATORY SYNDROME (SARS)

2.3.1 WHAT IS SARS?

Severe acute respiratory syndrome (SARS) is the first new infectious disease of this millennium which is a form of pneumonia .The symptoms are high fever, nonproductive cough, chills, myalgia, lymphopenia, and progressing infiltrates in chest radiography. WHO's report says, that it was one of the major epidemic between November 2002 and July 2003 with the mortality rate of 9.6%,. Coronaviruses are positive-strand, enveloped RNA viruses that are important pathogens of mammals and birds. These groups of viruses cause enteric or respiratory tract infections in a variety of animals including humans, livestock and pets.

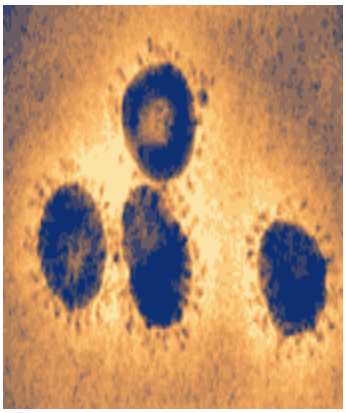


Fig 2.2 SARS coronavirus (SARS-CoV), the causative agent of the syndrome

2.3.2 HOW DID SARS-COV SUDDENLY APPEAR IN HUMANS?

Past analysis about the origin of the virus says that SARS-CoV genetic sequence with human and domestic animal coronaviruses showed a large difference in them. So, it was hypothesized that the new virus might have originated from wild animals, particularly Civet Cats. The Animal (Civet cat) coronavirus was found to have 99% sequence similarity with a few deletions and mutations in them. A review on the recent reports says that SARS-CoV is different from that of the Civet cat virus, and the question whether Civet cat would have been the origin of SARS Virus, else Civet cat were also infected from other species is unanswered. Therefore, there are no data available on the possibility of horizontal transmission between animals, which hinders our knowledge to know the relative speed of evolution of different proteins and further indicate the expected stability of therapy against development of resistant strains. This fact raises the question whether the jump of the virus from an animal to humans was a single accident or may frequently occur in future?, So far, Literatures show that SARS-CoV has the ability to not only infect humans but also macaque monkeys, domestic cats, and ferrets, But yet the transmission of the virus from the domestic cat to man has not been shown. In this respect, coronaviruses are known to easily jump to other species.

2.3.3 VIRAL REPLICATION

SARS-CoV 3C-like protease (SARS-CoV 3CLpro), is a receptor, which is a part of the replicase polyproteins, cleaves a functional polypeptide and, consequently, leads to the maturation of SARS-CoV. Because of its functional importance in the SARS-CoV replication cycle, SARS-CoV 3CLpro is considered a potential target to develop novel anti-SARS drugs. Although a number of non-peptide inhibitors of SARS-CoV 3CLpro, such as bifunctional aryl boronic acids, isatin derivatives, polyphenols, etacrynic acid analogues, cinanserin, and other chemically diverse small molecules have been identified, only a few of these show potent inhibitory activity.

2.3.4 SARS-COV-VARIOUS TARGETS FOR DRUGS /VACCINES-(WHAT FEATURES OF SARS-COV AND ITS REPLICATION ARE POTENTIAL TARGETS FOR DEVELOPMENT OF NEW ANTIVIRAL DRUGS AND VACCINES?)

Unfortunately, there are no approved antiviral drugs that are highly effective against coronaviruses. However, many steps unique to coronavirus replication could be targeted for development of antiviral drugs. Coronavirus infection begins with binding of the spike protein (S) on the viral envelope to a specific receptor on the cell membrane. Conformational changes are induced in S that probably lead to fusion of the viral envelope with the host cell membrane .Molecules that block binding to the receptor or inhibit the receptor-induced conformational change in S might block SARS-CoV infection. Inhibitors of HIV-1 entry and membrane fusion are good models for new drugs that target this first step in coronavirus infection.

2.4 MIGRAINE

2.4.1 WHAT IS MIGRAINE?

Migraine is basically a neurological disorder. The word *migraine* is French in origin and comes from the Greek hemi crania. Literally meaning "half (the) head". Migraine headache is a severe pain that is typically on one side of the head but sometimes on both sides and can occur at any time of day and can last a few hours or up to one or two days. Attacks can be very intense, forcing the sufferer to abandon normal daily activities. It is commonly experienced between the

ages of 15 and 55, most of these sufferers have a family history of Migraine and women are affected more than men.

2.4.2 TYPES OF MIGRAINE

1) Common migraine
2) Classic migraine
3) Ophthalmoplegic migraine
4) Hemiplegic migraine
5) Basilar type migraine
6) Abdominal migraine
7) Acephalgic migraine

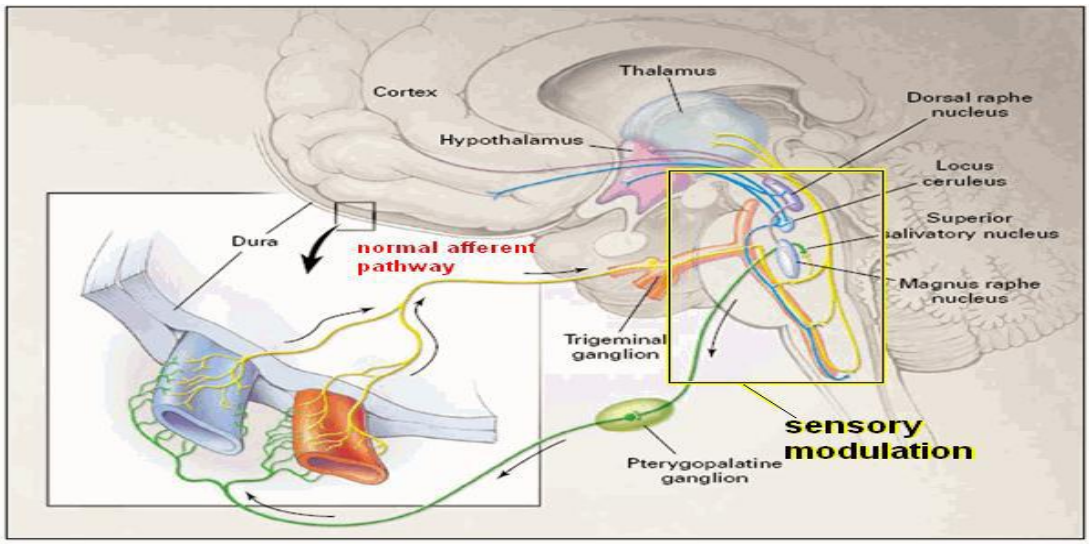


Fig 2.3 Path Physiology of Migraine (As Taken From Kegg Database).

2.4.3 MIGRAINE TRIGGERS

A migraine trigger is any factor that, on exposure or withdrawal, leads to the development of an acute migraine headache. Triggers may be categorized as behavioral, environmental, infectious, dietary, chemical, or hormonal. In the medical literature, these factors are known as 'precipitants.'

Migraine attacks may be triggered by: Allergic reactions Bright lights, loud noises, and certain odors or perfumes Physical or emotional stress ,Changes in sleep patterns ,Smoking or exposure to smoke **,**Skipping meals ,Dehydration ,Alcohol or caffeine ,Menstrual cycle fluctuations, birth control pills ,Tension headaches ,Foods containing tyramine (red wine, aged cheese, smoked fish, chicken livers, figs, and some beans), monosodium glutamate or nitrates (preserved meats) ,Other foods such as chocolate, nuts, peanut butter, avocado, banana, citrus, onions, dairy products, and fermented or pickled foods.

2.4.4 TARGETS FOR MIGRAINE

The neuropeptide ***calcitonin gene-related peptide (CGRP)*** from the trigeminal ganglion has been established in playing a key role in the pathogenesis of migraine. This study provides the evidence that the responsiveness of neuronal CGRP receptors is strongly enhanced in vitro and in vivo by expression of ***human receptor activity-modifying protein-1 (hRAMPI)***, an obligatory subunit of the CGRP receptor. It was demonstrated that activation of CGRP receptors on cultured trigeminal ganglion neurons increased endogenous CGRP mRNA levels and promoter activity. Here it was shown that ***RAMP1*** is functionally rate limiting for CGRP receptor activity in the trigeminal ganglion, which raises the possibility that elevated ***RAMP1*** might sensitize some individuals to CGRP actions in ***migraine***. This was taken here as the potential drug target for migraine.

2.5 QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIPS (QSAR)

Quantitative structure-activity relationship (QSAR) is the process by which chemical structure is quantitatively correlated with a well defined process, such as biological activity or chemical reactivity. Quantitative structure-activity relationship (QSAR) is a ligand based approach, which does not use any structural information of the target, but attempts to correlate the biological activity of the ligands with structural description. For example, biological activity can be expressed quantitatively as in the concentration of a substance required to give a certain

biological response. Additionally, when physiochemical properties or structures are expressed by numbers, one can form a mathematical relationship, or quantitative structure-activity relationship, between the two. The mathematical expression can then be used to predict the biological response of other chemical structures.

QSAR's most general mathematical form is:

Activity = f (physiochemical properties and/or structural properties)

This computational technique should be used to detect the functional group in your compound in order to refine your drug. This can be done using QSAR that consists of computing every possible number that can describe a molecule then doing an enormous curve fit to find out which aspects of the molecule correlate well with the drug activity or side effect severity. This information can then be used to suggest new chemical modifications for synthesis and testing.

2.6 FUNCTIONAL-LINK ARTIFICIAL NEURAL NETWORK (FLANN)

Pao originally proposed FLANN and it is a novel single layer ANN structure capable of forming arbitrarily complex decision regions by generating nonlinear decision boundaries. Here, the initial representation of a pattern is enhanced by using nonlinear function and thus the pattern dimension space is increased. The functional link acts on an element of a pattern or entire pattern itself by generating a set of linearly independent function and then evaluates these functions with the pattern as the argument. Hence separation of the patterns becomes possible in the enhanced space. The use of FLANN not only increases the learning rate but also has less computational complexity. Pao *et al* have investigated the learning and generalization characteristics of a random vector FLANN and compared with those attainable with MLP structure trained with back propagation algorithm by taking few functional approximation problems. A FLANN structure is shown in Fig. 2.4.

Let **X** is the input vector of size N×1 which represents N number of elements; the kth element is given by

$$\mathbf{X(k) = x(k), 1 \le k \le N}$$

Each element undergoes nonlinear expansion to form M elements such that the resultant matrix has the dimension of N×M.

The functional expansion of the element by power series expansion is carried out using the equation

$$s_i(k) = \begin{cases} 1 & \text{for } i = 0 \\ x(k) & \text{for } i = 1 \\ x^i(k) & \text{for } i = 2,3,4...M+1 \end{cases}$$

For trigonometric expansion, the expanded elements are

$$s_i(k) = \begin{cases} 1 & \text{for } i = 0 \\ x(k) & \text{for } i = 1 \\ \sin(i\pi x(k)) & \text{for } i = 2,4...M \\ \cos(i\pi x(k)) & \text{for } i = 3,5...M+1 \end{cases}$$

where. I=1,2,…,M/2. The bias input is unity. So total expanded values including the bias becomes Q=M+2.

Let the weight vector is represented as **W** having Q elements. The output y is given as

$$y(k) = \sum_{i=1}^{Q} s_i(k) w_i(k)$$

In matrix notation the output can be,

$$\mathbf{Y} = \mathbf{S} \cdot \mathbf{W}^{T}$$

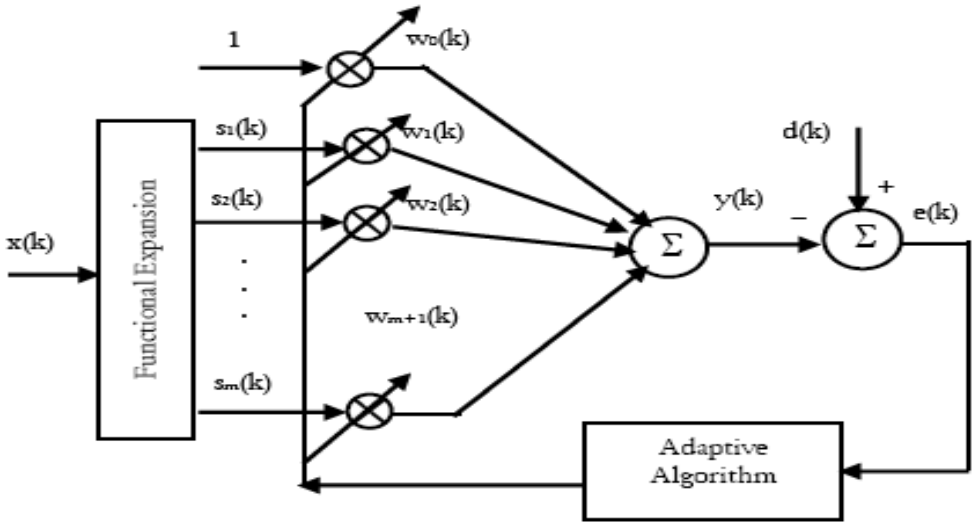


Fig 2.4 Functional Link Neural Network (FLANN) Structure

Referring to Fig. 2.4 the error signal e(k) at k^{th} iteration can be computed as

$$e(k) = d(k) - y(k)$$

Let ξ(k) denotes the cost function at iteration k and is given by

$$\xi(k) = \frac{1}{2}\sum_{j=1}^{p} e_j^2(k)$$

where P is the number of nodes at the output layer.

The weight vector can be updated by least mean square (LMS) algorithm, as

$$w(k+1) = w(k) - \frac{\mu}{2}\hat{\nabla}(k)$$

Where $\hat{\nabla}$ (k) is an instantaneous estimate of the gradient of ξ with respect to the weight vector W(k) . It is derived as

$$\hat{\nabla}(k) = \frac{\partial \xi}{\partial w} = -2e(k)\frac{\partial y(k)}{\partial w} = -2e(k)\frac{\partial [w(k)s(k)]}{\partial w}$$
$$= -2e(k)s(k)$$

2.7 PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal Components Analysis (PCA) is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data. The other main advantage of PCA is that once you have found these patterns in the data, and you compress the data, ie. by reducing the number of dimensions, without much loss of information. This technique used in image compression.

Method

Step 1: Get some data

The data set which needs dimensionality reduction may be treated as a data set, provided all the data's should be independent variables. We have considered variables of about 28 x 969 dimensions.

Step 2: Subtract the mean

For PCA to work properly, one has to subtract the mean from each of the data dimensions. The mean subtracted is the average across each dimension. So, all the x values have x' (the mean of the x values of all the data points) subtracted, and all the y values have y' subtracted from them. This produces a data set whose mean is zero

Step 3: Calculate the covariance matrix

covariance is always measured between 2 dimensions. If we have a data set with more than 2 dimensions, there is more than one covariance measurement that can be calculated. For example, from a 3 dimensional data set (dimensions x,y,z) you could calculate cov(x,y), cov(y,z) and cov(z,x). In fact, for an n -dimensional data set, you can calculate [n! / ((n-2)!*2)] different covariance values. A useful way to get all the possible covariance values between all the different dimensions is to calculate them all and put them in a matrix

Step 4: Calculate the eigenvectors and eigenvalues of the covariance Matrix

It is important to notice that these eigenvectors are both *unit* eigenvectors ie. Their lengths are both 1. This is very important for PCA, but luckily, most maths packages, when asked for eigenvectors, will give a unit eigenvectors. So what do they mean?, they provide us with information about the patterns in the data. The highest eigenvectors goes through the middle of the points, like drawing a line of best fit? That eigenvector is showing us how these two data sets are related along that line. The second eigenvector gives us the other, less important, pattern in the data, that all the points follow the main line, but are off to the side of the main line by some amount. So, by this process of taking the eigenvectors of the covariance matrix, we have been able to extract lines that characterise the data. The rest of the steps involve transforming the data so that it is expressed in terms of them lines.

Step 5: Choosing components and forming a feature vector

Here is where the notion of data compression and reduced dimensionality comes into it. On looking at the eigenvectors and eigenvalues from the previous section, one can notice that the eigenvalues are quite different values. In fact, it turns out that the eigenvector with the *highest* eigenvalue is the *principle component* of the data set.

It is the most significant relationship between the data dimensions. In general, once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives you the components in

Order of significance. Now, to reduce one can decide to *ignore* the components that have lesser significance. Only some information, may be lost but if the eigenvalues are small. What needs to be done now is one need to form a *feature vector*, which is just a fancy name for a matrix of vectors. This is constructed by taking the needed eigenvectors, and forming a matrix with these eigenvectors in the columns.

Step 6: Deriving the new data set

This, the final step in PCA, and is also the easiest. Once we have chosen the components (eigenvectors) that we wish to keep in our data and formed a feature vector, we simply take the transpose of the vector and multiply it on the left of the original data set, transposed.

Final data=Row feature vector x Row data adjust

|where row feature vector is the matrix with the eigenvectors in the columns *transposed* so that the eigenvectors are now in the rows, with the most significant eigenvector at the top, and row data adjust is the mean-adjusted data *transposed*, ie. the data items are in each column, with each row holding a separate dimension. It will give us the original data *solely in terms of the vectors we chose*. In the case of when the new data set has reduced dimensionality, ie. We have left some of the eigenvectors out, the new data is only in terms of the vectors that
we decided to keep. To show this on our data, the final transformation can be taken with each of the possible feature vectors. the transpose of the result in each case to bring the data back to the nice table-like format. The other transformation we can make is by taking only the eigenvector with the largest eigenvalue. So what have we done here is we have transformed our data so that is expressed in terms of the patterns between them, where the patterns are the lines that most closely describe the relationships between the data. This is helpful because we have now, the values of the data points tell us exactly where (ie. above/below) the trend lines the data point sits.

2.8 STEADY STATE GENETIC ALGORITHM (SSGA)

In the implemented SSGA the individual chromosome has three genes representing the ligand translation, four genes representing the ligand orientation and the other genes representing the ligand conformation. The translational genes are the X, Y, Z reference atom coordinates (usually

the closest atom to the ligand center of mass), the orientational genes are a quaternion constituted by a unit vector and one orientational angle. The conformational genes are the ligand dihedral angles (one gene to each dihedral angle). The ligand-protein energy function used is the GROMOS96 classical force field implemented in the THOR program of molecular mechanics/dynamics. The force field parameters are adjusted to reproduce experimental results (*e.g.*, structural and thermodynamic properties) or higher level *ab initio* quantum calculations. The GROMOS force field is given by:

$$\sum_{\text{Protein}}\sum_{\text{Ligand}}\left(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}+\frac{q_i q_j}{Dr_{ij}}\right)+\sum_{\text{Ligand}}\sum_{\text{Ligand}}\left(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}+\frac{q_i q_j}{Dr_{ij}}\right)+\sum_{\substack{\text{Dihedral}\\ \text{Angles}}}\gamma_k(1+\cos(\varpi_k\theta_k-\theta_{0k}))$$

where rij is the distance between the atoms i and j; Aij and Bij are the Lennard-Jones parameters; qi and qj are atomic charges and D is a sigmoidal distance-dependent dielectric constant. The first term of the equation corresponds to van der Waals interaction and electrostatic interaction between the protein and the ligand molecule, and the last two terms correspond to the ligand internal energy interaction, which also have one term for van der Waals interaction and one term for electrostatic interaction. The ligand-protein docking problem involves millions of energy evaluations, and the computational cost of each energy evaluation increases with the number of the atoms of the complex ligand-protein which has thousands of atoms. To reduce the computational cost, we implemented a grid-based methodology where the protein active site is embedded in a 3D rectangular grid and on each point of the grid the electrostatic interaction energy and the van der Waals terms for each ligand atom type are pre-computed and stored, taking into account all the protein atoms. In this way the protein contribution at a given point is obtained by tri-linear interpolation in each grid cell. A random initial population of individuals is generated inside the grid. For translational genes, random values between the maximum and minimum grid sizes are generated. For flexible docking, we also generated the initial population using a Cauchy distribution. The individual translational genes are generated by adding a random perturbation (drawn from a Cauchy distribution) to the grid center coordinates. In this way

individuals are generated with higher probability near the grid center, while still permitting that individuals be generated far from the center. The Cauchy distribution is given by:

$$C(\alpha,\beta,x)=\frac{\beta}{\pi(\beta^2+(x-\alpha^2))}$$
$$\alpha\geq 0,\ \beta>0,\ -\infty<x<\infty$$

Where α and β are Cauchy distribution parameters. In this work we used α = 0 and β = 0.75. For genes orresponding to angles (dihedrals and/or orientationals), random values ranging from 0° to 360° are generated. Finally, for the genes corresponding to the orientational unit vector, random values between -1 and 1 are used. The individuals are evaluated, and then are selected to suffer recombination or mutation. A rank-based selection scheme was used. A new individual is inserted in the population if its fitness is better than the fitness of the worst individual in the population. The algorithm evolves until the maximum number of the energy evaluations is reached. The reproduction operators used are classical two-point crossover and non-uniform mutation operators. The non-uniform mutation operator, when applied to an individual i at generation n generation, mutates a randomly chosen variable ci according to the following:

$$c_i^{new}=\begin{matrix}c_i+\Delta(ngen,\,b_i-c_i),\ \text{if}\ \tau=0\\ c_i+\Delta(ngen,\,c_i-a_i),\ \text{if}\ \tau=1\end{matrix}$$
$$c_i\in(a_i,b_i),\quad \Delta(ngen,y)=y(1-r^{\left(1-\frac{ngen}{max\ gen}\right)^b})$$

where ai and bi are respectively the lower and upper bounds for the variable ci, τ is randomly chosen as 0 or 1, r is randomly chosen in [0,1] and the parameter b set to 5. In the flexible docking, initially one randomly decides if a conformational gene will be mutated or not. Then a gene in the chosen group (conformational or not) is randomly selected for mutation. In this way, the seven translational/ orientational genes have the same probability of being mutated as the conformational ones.

2.9 QUANTITATIVE STRUCTURE ACTIVITY MODELING (QSAR) USING PCA AND FLANN FOR SARS

2.9.1 DATASET

A Dataset of 28 non-peptide inhibitor molecules collected from the literature are used in this paper for study. All the molecules studied have the same parent skeleton. All these compounds share similar chemical structure and a core structure of N-phenyl-2-(pyrimidiny-2-ylsulfanyl) acetamide has been identified. This core structure is used as a query structure to search for analogues. The selected protein and a single ligand (CMK) found to be crystallized with the original protein target. The SARS coronavirus protease (1UK4) is collected from Protein Data Bank (http://*www.rcsb.org*). IC_{50} is the concentration of the compound leading to 50% inhibitory effect. The logarithm transformation of this parameter is used as biological end points (log IC_{50}) in the QSAR study so as to move the data to a nearly normal distribution.

2.9.2 GENERATION AND SELECTION OF MOLECULAR DESCRIPTORS BY PRINCIPAL COMPONENT ANALYSIS (PCA)

The collected 28 druggable molecules are built and energy minimized using PRODRG software and saved as the mol (output) files. The output files are loaded into PreADMET which is a web-based application for predicting ADME data and building drug-like library using in- silico method. It calculates 955 diverse molecular descriptors which include 60 constitutional descriptors, 61 electrostatic descriptors, 16 geometrical descriptors, 130 physicochemical descriptor, and 688 topological descriptors. The other structural features of the molecules like polar surface area, logP and 14 other descriptors are also calculated from ADME-tox server. A total no. of 969 descriptors are calculated for each compound.

To reduce the redundancy existing in the descriptors data matrix the principal component analysis technique (PCA) is used where almost three fourth of the total dimension is reduced. Further by clustering the remaining one fourth descriptors it is observed that in some clusters,

one descriptor shows maximum explained variance. Such descriptors are first chosen. Subsequently any one descriptor from the remaining clusters is selected. All the selected twenty variables by this process constitute a feature vector of a compound. The same process is repeated for all 28 compounds.

2.9.3 QSAR TECHNIQUE

Quantitative structure-activity relationship (QSAR) is a ligand based approach, which does not use any structural information of the target, but attempts to correlate the biological activity of the ligands with structural description. Selected 18 drug molecules are classified into a training set and rest 10 into a test set. The test set includes compound 4, 7, 8, 11, 12, 14, 15, 18, 21, 23 and the remaining is used as a training set. Artificial neural network (ANN) is performed on these data, which correlates biological activity with physicochemical descriptors. The refinement of model is based on the consideration of statistical parameters such as correlation coefficient (r^2), cross validated correlation coefficient (q^2) and the mean squared error (MSE).

2.9.4 DOCKING STUDY

Using Autodoc 4.0 software the overall lowest binding energy output and the predicted binding energy (which is the summation of Intermolecular energy and torsional energy) are computed and taken as the criterion for ranking. Using this table eight top conformers based on their lowest energies are selected and these are further carefully checked according to ligand's location and size. The objective is to find ligands with good steric complimentarity. Furthermore the intermolecular hydrogen bonds are also investigated for these top conformers. The main features shared by these eight molecules are used to design analogues for Compound 1, Compound 6, and Compound 8. Further modifications of the analogues are made by replacing each one's existing amide bond with all possible substituting peptide bonds collected from the literature. These two sets of peptide and amide bonds are collected . All the modified compounds are made to pass through ADME/tox filters which screen them via simple filtering rules such as molecular weight,

polar surface area, logP or number of rotatable bonds. Only 11 out of 123 overall modified ligands retrieved after the screening are found to pass ADME/tox filters with improved binding efficiency and steric compatibility.

2.10 MODELLING OF PARAMETERS USING DOCKING FOR MIGRAINE

2.10.1 TARGET INDENTIFICATION

Human receptor activity-modifying protein-1 (hRAMPI) protein is sufficient for functionality of ***CGRP*** and is identified as a ***drug target*** for migraine. By inhibiting the activity of ***hRAMPI*** the migraine attacks can be substantially reduced. The structure of human ***receptor activity-modifying protein-1 (hRAMPI)*** **is taken from** RCSB Protein data bank whose PDB code is 2yx8.

2.10.2 TARGET VALIDATION

Receptor structure plays a central role in the target based drug design. The crystal structure of 2yx8 fetched from the PDB site and identification of secondary structure is done through the STRIDE WEB INTERFACE (http://webclu.bio.wzw.tum.de/stride) which generates protein secondary structure assignment from atomic co-ordinates based on the combined use of hydrogen bond energy and statistically derived backbone tensional angle information to identify the number of secondary structure helix, sheet and coil. To check whether any side chain is missing or not we use SCWL 3.0 software. (SCWRL3.0 is based on graph theory to solve the side chain prediction problem.) The missing residues were fixed using Deep View (The Swiss PDB Viewer available at (http://us.expasy.org.spdbv/).

2.10.3 ENERGY MINIMIZATION OF TARGET RECEPTOR

Before energy calculations can be performed, it is necessary to correct structural inconsistencies, add hydrogen and associated atoms with force field parameters. Minimization routines are

provided by MMTK, which is included with Chimera. The Kollman charges were added to each atom of the remained promoter The Amber ff99 force field is used for standard residues, and Amber Antechamber module (also included with Chimera) is used to assign parameters to nonstandard residues.

2.10.4 PREPARATION OF LIGANDS

The compounds are drawn and are energy minimized in **Prodrg server (**http://davapc1.bioch) using the GROMOS 96.1 force field. Since ligands are not peptides, Gasteiger charge was assigned and then non-polar hydrogen‟s were merged.

Docking: The GA-LS (Lamarckian genetic algorithm) was chosen to search for the best conformers. During the docking process, the docking parameters were set as Maximum Number of GA runs 100, Population size of 150, Maximum number of evaluation 250000, Rate of Gene mutation 0.02, and Rate of Crossover 0.8, for each Compound. The parameters were set using the software Autodock Tools. The Calculations of Autogrid and Autodock were performed on Linux Operating system having. Evaluation of modified ligands by a flexible docking procedure followed by the Screening of ***ADME/Tox filters.***

At the end of the docking run, it outputs a result which is the lowest energy conformation of the ligands, it found during that run. This conformation is a combination of translation, quaternion and Torsional angles and is characterized by intermolecular energy, internal energy and Torsional energy. The first two of these combined give the „Docking energy‟ while the first and third give „Binding energy‟ .Autodock 4.0 also breaks down the total energy into Vander walls (vdW) energy and an electrostatic energy for each atom. We used the overall lowest binding energy output by Autodock 4.0 and the Predicted binding energy, as the criterion for ranking. Therefore after ranking the 3 top conformers according to their lowest energies, the Ligands were further carefully checked according to the factor such as ligands location, size of the Ligand; to yield ligands with good steric complementarily. Furthermore they were also investigated for the 3 top conformers. The main features shared by these 3 top conformers were study with help of **LigndScout2.0** .and 16 derivatives were design on (basis of the important

pharmacophores of three top conformer) PRODRG using the GROMOS 96.1 forcefield .All the 16 derivatives are passed through ADME/Tox Filter and also ligands are passing through Lipinski filter to check whether ligands satisfy the 5 Lipinski rules using. The derivatives are docked to the receptor using autodock4.0 tools, best 4 ligand are taken into consideration. The important pharmacophores and critical amino acids which are common in all four ligand protein complex are predicted using **LigndScout2.0.**

2.11 SIMULATION RESULTS FOR QSAR STUDY

Table 2.1 Statistical measurement of the model using ANN and FLANN method

Sl. No	r^2		q^2		MSE		No. of descriptors
	ANN	FLANN	ANN	FLANN	ANN	FLANN	
1	0.7956	0.7960	0.7955	0.7956	0.0111	0.0110	1
2	0.9543	0.9620	0.9542	0.9608	0.0036	0.0034	2
3	0.9738	0.9762	0.9735	0.9758	0.0017	0.0012	3
4	0.9825	0.9856	0.9821	0.9852	0.0007	0.0005	4
5	0.9878	0.9900	0.9872	0.9900	0.0003	0.0001	5

Table 2.2 Binding free Energies for the 28 molecules (Unmodified Compounds)

Sl. No	Compounds	Lowest Binding Energies (Kcal/mol)	Binding energies (Torsional +Intermolecular) (kcal/mol)	Rank 1	Rank 2
1	Compound1	-6.65	-7.57	8	1
2	Compound2	-6.84	-6.6	5	7
3	Compound3	-6.66	-6.41	7	8
4	Compound4	-6.56	-6.89	11	4
5	Compound5	-6.51	-5.9	12	14
6	Compound6	-7.45	-6.92	2	3
7	Compound7	-7.00	-6.79	4	5

8	Compound8	-7.66	-7.42	1	2
9	Compound9	-5.67	-5.57	26	17
10	Compound10	-7.06	-6.64	3	6
11	Compound11	-6.23	-5.12	19	21
12	Compound12	-6.25	-5.07	18	22
13	Compound13	-5.97	-4.68	22	25
14	Compound14	-6.47	-4.96	13	24
15	Compound15	-6.31	-4.55	16	27
16	Compound16	-6.44	-5.22	14	19
17	Compound17	-6.71	-6.00	6	12
18	Compound18	-6.61	-5.87	9	15
19	Compound19	-5.9	-5.86	25	16
20	Compound20	-6.13	-6.08	20	13
21	Compound21	-4.45	-3.92	28	28
22	Compound22	-6.4	-6.13	15	11
23	Compound23	-5.94	-5.40	24	18
24	Compound24	-6.29	-6.23	17	10
25	Compound25	-6.59	-6.26	10	9
26	Compound26	-5.96	-4.68	23	26
27	Compound27	-5.24	-5.02	27	23
28	Compound28	-6.09	-5.16	21	20

Table 2.3 Peptide bonds substituents substituted over amide bonds of non-peptide ligands

S. No	Amide bonds of non-peptide ligands (R1-CO-NR2-R3)	Various groups substituting peptide bonds
1	R1-CO-NH-R3	R1-CH2-CH2-R2
2	R1-CO-NH-R3	R1-CO-O-R2
3	R1-CO-NH-R3	R1-CHF-S-R2
4	R1-CO-NH-R3	R1-CF2-NH-R2
5	R1-CO-NH-R3	R1-CO-CF2-R2
6	R1-CO-NH-R3	R1-CO-CF2-CO-R2
7	R1-CO-NH-R3	R1-NH-CH(CF3)-R2
8	R1-CO-NH-R3	R1-NH-R2
9	R1-CO-NH-R3	R1-O-R2
10	R1-CO-NH-R3	R1-NH-N=CH-CO-R2
11	R1-CO-NH-R3	R1-CO-CO-NH-R2
12	R1-CO-NH-R3	R1-CH2-NH-R2
13	R1-CO-NH-R3	R1-CH(OH)-CH2-NH-R2
14	R1-CO-NH-R3	R1-C(OH)=CH-R2
15	R1-CO-NH-R3	R1-CH(OH)-CH(OH)-R2
16	R1-CO-NH-R3	R1-CH(OH)-CH2-R2
17	R1-CO-NH-R3	R1-CH=CH-R2
18	R1-CO-NH-R3	R1-CMe=CMe-R2
19	R1-CO-NH-R3	R1-CH=C(Pro)
20	R1-CO-NH-R3	R1-C(CF3)=CH-R2
21	R1-CO-NH-R3	R1-C(CH3)=CH-R2

Table 2.4 Observed and the predicted (FLANN) values of logIc_{50} for the (training + test) set of the 28 derivatives

Sl. No	Actual logIC_{50}	Predicted logIC_{50} (ANN)	Residual Error
1	0.477121	0.4908	-0.01368
2	1	1.0089	-0.0089
3	1.04139	1.0491	-0.00771
4	1.07918	1.0908	-0.01162
5	1.14613	1.1529	-0.00677
6	1.17609	1.1829	-0.00681
7	1.17609	1.1826	-0.00651
8	1.17609	1.1787	-0.00261
9	1.47712	1.4796	-0.00248
10	1.60206	1.6035	-0.00144
11	1.60206	1.6035	-0.00144
12	1.65321	1.6512	0.00201
13	1.77815	1.7796	-0.00145
14	1.77815	1.7733	0.00485
15	2	1.9968	0.0032
16	2.30103	2.298	0.00303
17	2.30103	2.2962	0.00483
18	2.30103	2.2989	0.00213
19	2.30103	2.2974	0.00363
20	2.30103	2.2983	0.00273
21	2.39794	2.3928	0.00514
22	2.47712	2.4693	0.00782
23	2.47712	2.4726	0.00452
24	2.47712	2.4738	0.00332
25	2.54407	2.541	0.00307
26	2.60206	2.5977	0.00436
27	2.69897	2.6955	0.00347
28	3	2.9883	0.0117

Table 2.5 Calculated Biological Activity values for the 11 modified ligands

S. No	Modified Compound	Lowest Binding Energy (Kcal/mol)	Predicted Log IC_{50} (ANN)
1	C1_Ana1_CD_11	-8.00	2.6743
2	C1_Ana3_CD_18	-8.85	2.4517
3	C1_Ana3_CD_19	-8.77	2.8433
4	C8_Ana5_CD_19	-8.86	2.2656
5	C8_Ana3_CD_14	-8.06	2.8484
6	C8_Ana3_CD_18	-8.19	2.9301
7	C8_Ana3_CD_19	-8.62	0.4789
8	C6_Ana3_CD_11	-8.15	2.7054
9	C6_Ana3_CD_13	-8.09	0.2906
10	C6_Ana3_CD_18	-8.44	2.8777
11	C6_Ana3_CD_19	-9.24	0.1880

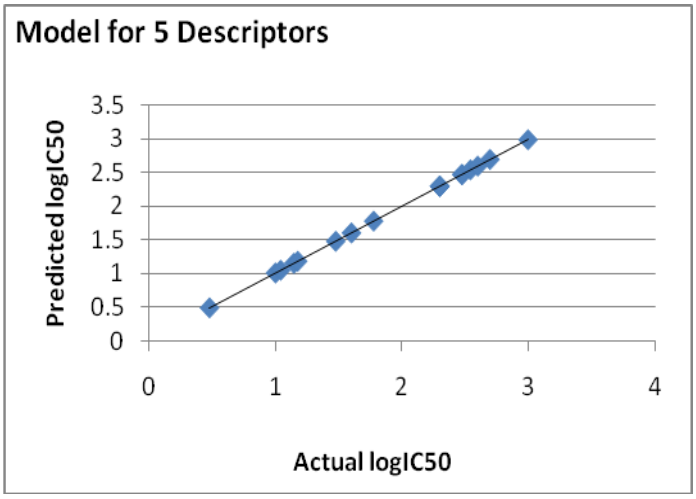


Fig 2.5 the predicted vs. actual activities of training (18) sets

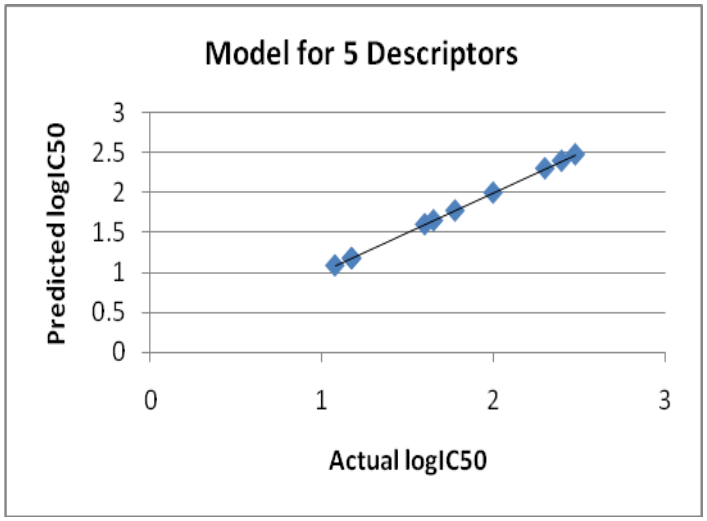


Fig 2.6 the predicted vs. actual activities of test (10) sets

2.12 INFERENCES DRAWN FROM QSAR STUDY AND PREDICTING POTENTIAL DRUG MOLECULES FOR SARS Co-3CL VIRUS

The docking results are ranked according to the ascent of the docking energies of the 100 conformers for each of the ligands. A comparison of the results between rank 1 and rank 2 suggests that some large sized ligands suffer from more loss of torsional freedom upon binding. The top eight compounds are selected based on the comparison of scoring function, which predicts the ranking of different ligands in approximate order of ligand size and order of affinity.

From observing the hydrogen bonds formed between compounds 1, 2, 3, 4, 6, 7, 8, 10 (having lowest binding energy) and 3CL Protease respectively, it reveals that Asn142 occurs for four times, Cys145 for twelve times, His41 for three times, Gly143 occurs for sixteen times, His163 occurs for two times and Glu166 occurs for six times. From the frequency of the residues occurrence in the formation of hydrogen bonding, His41 and Cys145 plays important role, this may be due to the carbonyl group of the amide bond and the nitrogen present in the centered benzene ring. The other residues like Asn142, Gly143, His163 and Glu166 also take part in the hydrogen bonding with relatively high frequency. This information is helpful for our drug design in which the potential drug should interact well with these residues, especially with His41 and Cys145. His 41 and Cys145 is of great concern in anti-SARS drug design because it is believed to be the Catalytic dyad. Therefore we designed analogues for the first three compounds (1, 6 and 8) keeping the amide bond and the center benzene ring locked such as its catalytic function is depressed. For each of the best performing analogues, residue His41 interacting with the carbonyl group of the amide bond whose frequency of occurrence observed only three times is replaced by groups of various possibilities of substituting peptide bonds shown in Table 3. This is done with the aim to increase the overall interaction in the form of total binding free energy. This docking study results 11 modified better ligands out of 123 overall ones, which are also passed the ADME-tox filters with improved binding efficiency and steric complementarity. The 11 modified compounds found to have improved binding efficiency is screened through the model equation with the aim to identify the unique characteristic of compounds and to build the relationship between the structure and biological activity. Table 5 reports the calculated

biological activity values for the 11 modified ligands. It is found that the C6_Ana3_CD_19 with lowest binding energy (-9.24) had a very high biological activity of (0.1880), followed by C6_Ana3_CD_13 (0.2906) and C8_Ana3_CD_19 (0.4789) respectively. The above result clearly says that C6_Ana3_CD_19, C6_Ana3_CD_13 and C8_Ana3_CD_19 may be considered as potential drugs to treat SARS and further wet lab studies are recommended.

2.13 SIMULATION RESULTS FOR DOCKING STUDY USING SSGA

During the docking process, the docking parameters were set as the following for each Compound.

Maximum Number of GA runs 100
Population size of 150
Maximum number of evaluation 250000
Rate of Gene mutation 0.02
Rate of Crossover 0.8

Table 2.6 Comparisons Of Experimental Binding Affinities And Docking Scores Using Autodock4.

COMPOUNDS	LOWEST BINDING ENERGY	CALCULATED Pki conc	RANK
Compound1	-5.58	80.62μM	RANK14
Compound2	-5.75	61.3μM	RANK7
Compound3	-5.31	127.47μM	RANK16
Compound4	-5.57	82.31μM	RANK15
Compound5	-5.7	66.87μM	RANK10
Compound6	-5.82	54.41μM	RANK5
Compound7	-5.74	62.01μM	RANK8
Compound8	-5.75	60.65μM	RANK6
Compound9	-6.51	16.88μM	RANK2
Compound10	-6.16	30.62μM	RANK3
Compound11	-5.62	76.47μM	RANK13
Compound12	-6.07	35.82μM	RANK4
Compound13	-6.6	14.45μM	RANK1
Compound14	-5.68	68.79μM	RANK11
Compound15	-5.66	71.24μM	RANK12
Compound16	-5.73	62.58μM	RANK9

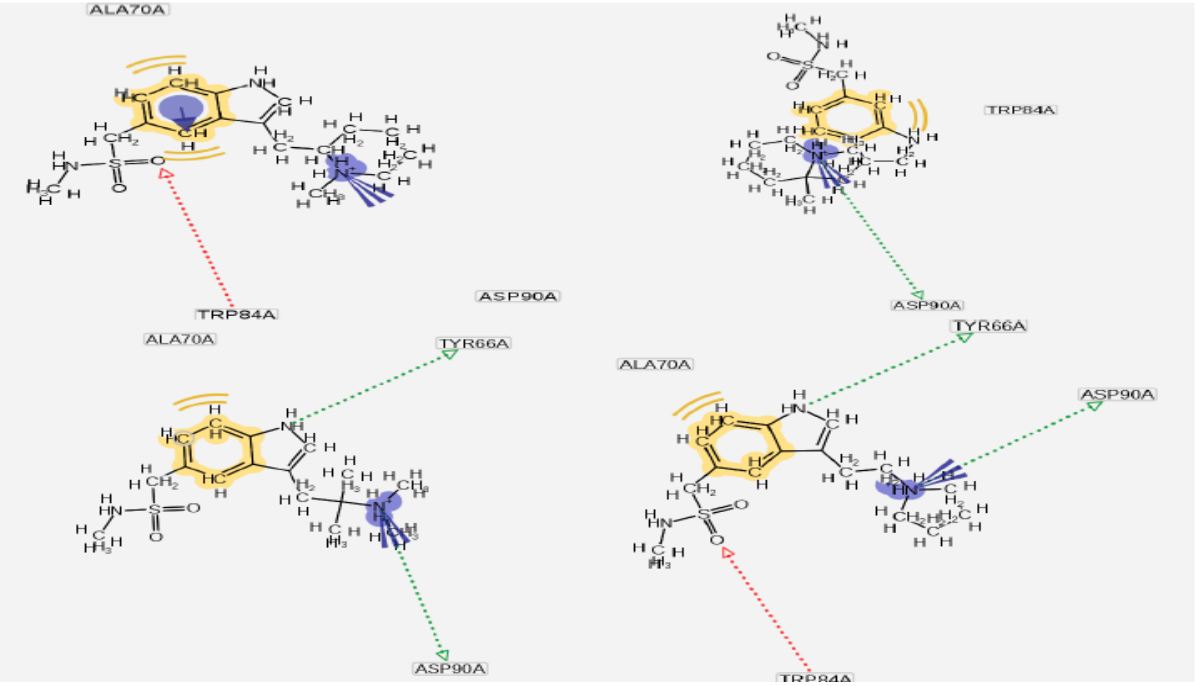
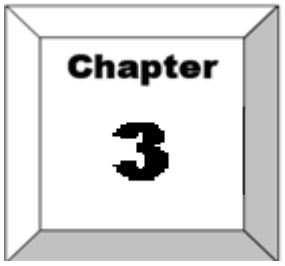


Fig 2.7 Interaction between hRAMPI receptor and best ligands

2.14 INFERENCES DRAWN FROM DOCKING STUDY USING SSGA

The docking results are ranked according to the ascent of the docking energies of the 100 conformers for each of the ligands. Ranking the energy results according to the Binding Energy which Included the Intermolecular Energy and the Torsional terms, It was found that most of the Ligands Interacted quite well with the receptor in the pocket. A Comparison of the results between Rank1, Rank2, Rank3 and Rank4 suggests that some large sized Ligands suffer from more loss of Torsional freedom upon binding. The top three compounds was selected based on the comparison of Scoring function, which predicts the ranking of different ligands in approximate order of Ligand size, order of affinity, allows selectivity and was helpful in setting up the priorities. It shows that compound no-13, 9, 10, 12 has lowest binding energy and the compound were Also checked for their conformations from the output dlg file of a docking run, found at the compounds were also checked for their conformation from the output dlg file of a

docking run, and are sorted out according to their RMSD values. We found a cluster of 2 conformations in compound13, a cluster of 8 conformations in Compound 9, a cluster of 19 conformations in Compound10, a cluster of 35 conformations in compound12 all observed with a RMSD tolerance set to 2.0.The recurrence of the identical conformations of one ligand means that fits well to the pockets and is likely to be a good inhibitor candidate.



PROTEIN STRUCTURAL CLASS PREDICTION

3.1 INTRODUCTION

In the post genomic era the study of sequence to structure relationship and functional annotation plays an important role in molecular biology. In this context protein fold prediction is one of the major tasks in protein science. The functions of protein are relevant to its 3D structure. The function of protein can be efficiently determined by the sequence analysis and structure analysis. The knowledge of protein structural class can provide useful information towards the determination of protein structure. The exponential growth of the newly discovered protein sequences by different scientific community made a large gap between the number of sequence-known proteins and the number of structure-known proteins. Hence there is a critical challenge to develop automated methods for fast and accurately determining the structures of proteins in order to reduce the gap. Therefore the development of computational methods for identifying the structural classes of newly found proteins based on their primary sequence is essential. The structural class has become one of the most important features for characterizing the overall folding type of a protein and it has played as important role in molecular biology, pharmacology, rational drug design and many other applications.

The concept of protein structural classes was proposed by Levitt & Chothia on a visual inspection of polypeptide chain topologies in a dataset of 31 globular proteins. He proposed ten structural classes, four principal and six small classes of protein structure. But the biological community follows the first four principal classes and these are: all α, all β, α+β and α/β classes. The all-α and all-β classes represent structures that consist of mainly α-helices and β-strands, respectively. The α/β and α+β classes contain both α-helices and β-strands which are mainly interspersed and segregated. α class: Proteins of this class contain more than **45%** α-helices and

less than **5%** β-strands. β class: Proteins of this class contain less than 5% α-helices and more than **45%** β-strands. α+β class: Proteins of this class contain more than **30%** a-helices and more than 20% P-strands with dominantly anti-parallel β-strands. α/β class: Proteins of this class contain more than **30%** a-helices and more than 20% β-strands with dominantly parallel β-strands. These class definitions have been well accepted and are still in common use by many researchers.

The development of predicting protein structural classes from the primary sequence are mainly focused on the two aspects. First is the effective representation of the protein sequence and then the development of the powerful classification algorithms to efficiently predict the class. Many in-silico structural class prediction algorithms and methods have developed in few decades. Many amino acid indices and feature s are used for the assignment of the protein sequence. Nishikawa *et al.* first indicated that the protein structural classes are strongly related to the Amino Acid composition (AA). Auto-correlation functions based on non-bonded residue energy, polypeptide composition, pseudo AA composition and complexity measure factor have been used by many researchers. Several classification methods are also proposed such as distance classifier, component coupled methods, Principal component analysis and support vector machine. Although promising results have been achieved, the representation of protein with AA lacks sequence-order information and sequence-length information. In this chapter we have proposed a novel optimization approach for the prediction of protein structural class using Genetic Algorithm (GA) and Particle swarm Optimization (PSO). We have used the modified amino acid composition vector proposed by Fei Gu et al. which contains the sequence length information.

3.2 PROTEIN STRUCTURAL CLASS

3.2.1 PROTEIN

Proteins are an important class of biological macromolecules present in all biological organisms, made up of such elements as carbon, hydrogen, nitrogen, oxygen, and sulphur. All proteins are polymers of amino acids. The polymers, also known as polypeptides consist of a sequence of

20 different L-α-amino acids, also referred to as residues. For chains under 40 residues the term peptide is frequently used instead of protein. To be able to perform their biological function, proteins fold into one, or more, specific spatial conformations, driven by a number of non covalent interactions such as hydrogen bonding, ionic interactions, Vander Waals forces and hydrophobic packing. In order to understand the functions of proteins at a molecular level, it is often necessary to determine the three dimensional structure of proteins. This is the topic of the scientific field of structural biology, that employs techniques such as X-ray crystallography or NMR spectroscopy, to determine the structure of proteins.

A number of residues are necessary to perform a particular biochemical function, and around 40-50 residues appears to be the lower limit for a functional domain size. Protein sizes range from this lower limit to several thousand residues in multi-functional or structural proteins. However, the current estimate for the average protein length is around 300 residues. Very large aggregates can be formed from protein subunits, for example many thousand actin molecules assemble into a microfilament.

3.2.2 LEVELS OF PROTEIN STRUCTURE

Biochemistry refers to four distinct aspects of a protein's structure:

- **Primary structure** - the amino acid sequence of the peptide chains.
- **Secondary structure** - highly regular sub-structures (*alpha helix* and *strands of beta sheet*) which are locally defined, meaning that there can be many different secondary motifs present in one single protein molecule.
- **Tertiary structure** - three-dimensional structure of a single protein molecule; a spatial arrangement of the secondary structures. It also describes the completely folded and compacted polypeptide chain.
- **Quaternary structure** - complex of several protein molecules or polypeptide chains, usually called protein subunits in this context, which function as part of the larger assembly or protein complex.

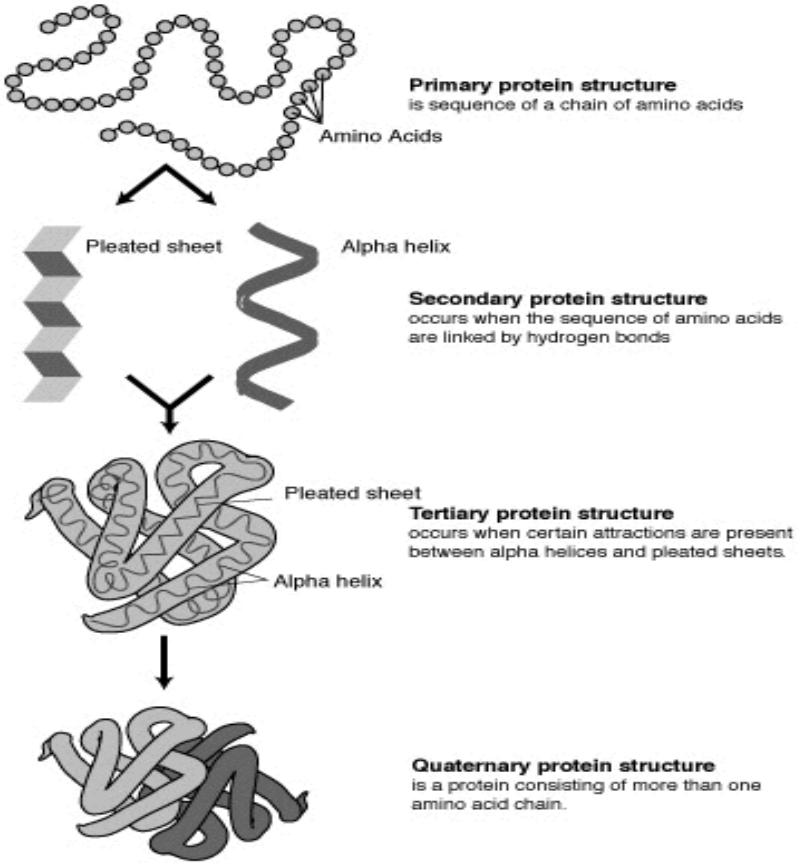


Fig 3.1 Four Types Of Protein Structure

3.2.3 STRUCTURAL CLASSES

A protein (domain) is usually classified into one of the following four structural classes: all-α, all-β, α/β, and $\alpha + \beta$. Structural class categorizes various proteins into groups that share similarities in the local folding patterns. The all-α and all-β classes represent structures that consist of mainly α-helices and β-strands, respectively.

The α/β class represents those proteins containing both α- helices and β- strands that are largely interspersed in forming mainly parallel β- sheets, while the α+β class represents those containing also both α- helices and β- strands but they are largely segregated in forming mainly antiparallel β- sheets. Prediction of structural classes is based on identifying these folding patterns based on thousands of already categorized proteins, and applying these patterns to unknown structures but known amino acid sequences. The structural class of a protein presents an intuitive description of

its overall folding and the restrictions of the structural class have a high impact on its secondary and tertiary structure prediction. Therefore, the prediction of the four principal protein structural classes is the foundation in the field of protein analysis.

3.2.4 AMINO ACIDS IN PROTEINS

An **amino acid** is a molecule containing both amine and carboxyl functional groups. These molecules are particularly important in biochemistry, where this term refers to alpha-amino acids with the general formula $H_2NCHRCOOH$, where R is an organic substituent. In the alpha amino acids, the amino and carboxylate groups are attached to the same carbon, which is called the α–carbon. The various alpha amino acids differ in which side chain (R group) is attached to their alpha carbon. They can vary in size from just a hydrogen atom in glycine through a methyl group in alanine to a large heterocyclic group in tryptophan.

Amino acids are critical to life, and have a variety of roles in metabolism. One particularly important function is as the building blocks of proteins, which are linear chains of amino acids. Amino acids are also important in many other biological molecules, such as forming parts of coenzymes, as in S-adenosylmethionine, or as precursors for the biosynthesis of molecules such as heme. Due to this central role in biochemistry, amino acids are very important in nutrition.

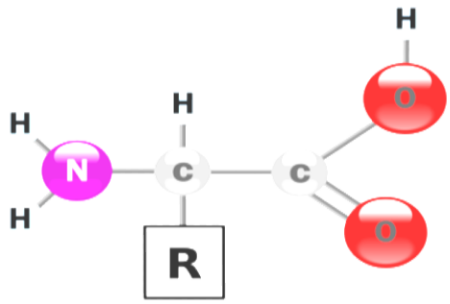


Fig 3.2 Basic Structure of an Amino Acid

Table 3.1 Twenty Different Amino Acids

Amino Acid	3-Letter	1-Letter	Side chain polarity	Side chain charge (pH 7)	Hydropathy index
Alanine	Ala	A	nonpolar	neutral	1.8
Arginine	Arg	R	polar	positive	-4.5
Asparagine	Asn	N	polar	neutral	-3.5
Aspartic acid	Asp	D	polar	negative	-3.5
Cysteine	Cys	C	nonpolar	neutral	2.5
Glutamic acid	Glu	E	polar	negative	-3.5
Glutamine	Gln	Q	polar	neutral	-3.5
Glycine	Gly	G	nonpolar	neutral	-0.4
Histidine	His	H	polar	positive	-3.2
Isoleucine	Ile	I	nonpolar	neutral	4.5
Leucine	Leu	L	nonpolar	neutral	3.8
Lysine	Lys	K	polar	positive	-3.9
Methionine	Met	M	nonpolar	neutral	1.9
Phenylalanine	Phe	F	nonpolar	neutral	2.8
Proline	Pro	P	nonpolar	neutral	-1.6
Serine	Ser	S	polar	neutral	-0.8
Threonine	Thr	T	polar	neutral	-0.7
Tryptophan	Trp	W	nonpolar	neutral	-0.9
Tyrosine	Tyr	Y	polar	neutral	-1.3
Valine	Val	V	nonpolar	neutral	4.2

3.3 LITERATURE REVIEWS OF COMPUTATIONAL TOOLS USED TO PREDICT PROTEIN STRUCTURAL CLASS

3.3.1 EUCLIDEAN DISTANCE

According to this algorithm, the potential function of a query protein in the 20-D composition space is given by

$$U^{(0)}(\mathbf{X}, \bar{\mathbf{X}}^{\xi}) = k^{(0)} \sum_{i=1}^{20} (x_i - \bar{x}_i^{\xi})^2$$

$$(\xi = \alpha,\ \beta,\ \alpha/\beta,\ \alpha + \beta),$$

where $k(0)$ is a force constant, which is trivial here and can be left out in calculation because it is the same for all the structural classes. As we know from a basic law in physics, a system will become the most stable when it is in a state with the lowest potential, or strictly speaking, the lowest free energy. Accordingly, if $U(0)(\mathbf{X}, \mathbf{X}^{\alpha})$ is the smallest among $U(0)(\mathbf{X}, \mathbf{X}^{\xi})$ (ξ=α, β, α/β and α+β), the protein **X** will fold into the all-a structural class; if $U(0)(\mathbf{X}, \mathbf{X}^{\beta})$ the smallest, then it will fold into the all-b class; and so forth. Therefore, the recognition rule should be formulated as

$$U^{0}(\mathbf{X}, \bar{\mathbf{X}}^{\ell}) = \mathbf{Min}\{U^{0}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha}),\ U^{0}(\mathbf{X}, \bar{\mathbf{X}}^{\beta}),\ U^{0}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha/\beta}),\ U^{0}(\mathbf{X}, \bar{\mathbf{X}}^{\alpha+\beta})\},$$

where, can be a, b, a/b, or a 1b, and the operator **Min** means taking the least one among those in the parentheses, and the superscript , represents the very structural class which the protein **X** belongs to. If there is a tie case, is not uniquely determined, but that rarely occurs. As we can see, after leaving out the constant $k(0)$, $U0(\mathbf{X}, \mathbf{X}^{l},)$ actually represents the squared Euclidean distance between **X** and $\mathbf{X}^{\xi}$ as used by Nakashima *et al.* for recognizing the protein structural class.

3.3.2 HAMMING DISTANCE METHOD

- Forming of nucleation. A nucleation can be predicted when 4 of 6 sequential residues in certain segment tend to form helix (the helix former), and this number is 3 of 5 for strand.
- The nucleation regions are extended along both directions of the sequence until the average 4-peptides propensities drops below 1.
- If any extended segment with average propensities $\langle P_{\alpha} \rangle > 1.03$(helical propensities larger than 1.03 are strong alpha former and alpha former) and $\langle P_{\alpha} \rangle > \langle P_{\beta} \rangle$ (subscript α means helical propensities while β corresponds to strand propensities), it can be predicted as helix. And the condition changes to $\langle P_{\beta} \rangle > 1.05$ (strand propensities larger than 1.05 are

strong strand former and strand former) and $\langle P_\beta \rangle > \langle P_\alpha \rangle$ for strand. If both helix and strand are predicted in certain region (overlapped region), the secondary structure conformation with higher average propensities is predicted.

3.3.3 PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA does not use the conventional amino acid components as the bases for the 20D composition space, but instead reorganizes them to a set of orthogonal and normalized base functions through a process of linear combination.

In the PCA approach the 20-dimensional amino acid composition space is reduced to an orthogonal space with fewer dimensions, and the original base functions are converted into a set of orthogonal and normalized base functions. The advantage of such an approach is that it can minimize the random errors and redundant information in protein dataset through a principal component selection, remarkably improving the success rates in predicting protein structural classes.

The essence of the idea of PCA is to optimize the dimensions of the original data matrix so as to exclude the overstaffed information and the random errors. To realize this, let us define the data matrix $\mathbf{D}\tau$ from the composition matrix $M\tau$ as follows:

$$\mathbf{D}_\tau = \begin{bmatrix} d^\tau_{1,1} & d^\tau_{1,2} & \cdots & d^\tau_{1,20} \\ d^\tau_{2,1} & d^\tau_{2,2} & \cdots & d^\tau_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ d^\tau_{20,1} & d^\tau_{20,2} & \cdots & d^\tau_{20,20} \end{bmatrix} = \mathrm{M}^{\mathrm{T}}_\tau \mathrm{M}_\tau$$

Where **T** is the transpose operator and

$$d^\tau_{i,j} = \frac{1}{n_\tau} \sum_{u=1}^{n_\tau} x_{i,u} x_{j,u}, \quad (i, j = 1, 2, \ldots, 20)$$

The data matrix $\mathbf{D}_\tau$ is symmetrical and positive definite matrices. According to linear algebra, it must have 20 non-negative Eigen values, which can be obtained by solving the following equation:

$$\mathbf{D}_\tau \Psi_j^\tau = \lambda_j^\tau \Psi_j^\tau = \lambda_j^\tau \begin{bmatrix} \psi_{j,1}^\tau \\ \psi_{j,2}^\tau \\ \vdots \\ \psi_{j,20}^\tau \end{bmatrix}, \quad (j = 1, 2, \ldots, 20)$$

where λi is the *j*th Eigen value for the τth subset

The PCA approach uses less orthogonal variables to represent the original data as described below. The number of principal eigenvectors is defined by setting a threshold value according to the following equation

$$\frac{\sum_{i=1}^{\mu} \lambda_i}{\sum_{i=1}^{20} \lambda_i} \geq \Theta, \quad (\mu < 20)$$

The query protein **X** is predicted belonging to the subset with which it has the largest predictive value. In other words, suppose

$$P_X^m \equiv \mathrm{Max}\{P_X^1, P_X^2, \ldots, P_X^M\}, \quad (m = 1, 2, \ldots, \text{or } M)$$

Where the operator means taking the maximum for those in the brackets, then we have $\mathbf{X} \in Sm$, meaning the query protein belongs to the *m*th subset.

3.3.4 LOGITBOOST

The LogitBoost Algorithm.
The Pseudo-Code of LogitBoost

1. Input data set $S = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_N, y_N)\}$, where $\vec{x}_i \in X$, $y_i \in Y = \{-1, 1\}$. Input number of iterations T.
2. Initialise the weight $w_i = 1/N (i = 1, \ldots, N)$; initialize committee function $F(\vec{x}) = 0$ and probabilities $p(\vec{x}) = P(y = 1|x) = 1/2$.
3. Repeat $t = 1, \ldots, T$
   a. Compute the weights and working response

$$w_i = p(\vec{x}_i)[1 - p(\vec{x}_i)]$$

$$z_i = \frac{y_i^* - p(\vec{x}_i)}{w_i}, \text{ where } y_i^* = (y_i + 1)/2$$

   b. Fit the function $f_t(\vec{x})$ by a weighted least-squares regression of z_i to $\vec{x}_i$ using weights w_i. In our study we use regression decision tree to fit the data $\{(\vec{x}_1, z_1), \ldots, (\vec{x}_N, z_N)\}$ using weights w_i.
   c. Update $F(\vec{x}) \leftarrow F(\vec{x}) + \frac{1}{2} f_t(\vec{x})$ and $p(x) \leftarrow \frac{e^{F(x)}}{e^{F(x)} + e^{-F(x)}}$.

4. Output the final classifier $LF(\vec{x}) = \text{sign}[F(X)]$.

The working dataset was taken from Chou (1999) that contains 204 protein chains, of which 52 are all-α proteins, 61 all-β proteins, 45 α/β proteins and 46 α+β proteins. Their average sequence similarity scores are 21% for all- α, 30% for all- β, 15% for α/β and 14% for α+β. Therefore, the majority of the proteins are not similar to each other in this dataset. In this study the protein samples are represented by their amino acid compositions, and hence each input of the LogitBoost corresponds to a vector or point in a 20-dimensional space.

3.3.5 SUPPORT VECTOR MACHINE (SVM)

Support vector machine (SVM) is a kind of learning machine based on statistical learning theory. first, map the input vectors into one feature space (possible with a higher dimension), either linearly or non-linearly, which is relevant with the selection of the kernel function. Then, within the feature space from the first step, seek an optimized linear division, i.e. construct a hyperplane which separates two classes (this can be extended to multi-class). SVM training always seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. SVMs have been used in a range of problems including drug design, image recognition and text classification. In this method, Vapnik's support vector machine is applied for predicting the structural classes of proteins.

Suppose we are given a set of samples, i.e. a series of input vectors $Xi \epsilon Rd$ (i=1, ..., N) with corresponding labels $yi\_\{+1,-1\}$ (i=1, ..., N).

Here −1 and +1 are used to represent, respectively, the two classes. The goal here is to construct one binary classifier or derive one decision function from the available samples, which has small probability of misclassifying a future sample. Both the basic linear separable case and the most useful linear non-separable case for most real life problems are considered here. For the SVM, the width of the Gaussian RBFs is selected as that which minimized an estimate of the VC-dimension. The parameter C that controls the error-margin trade-off is set at 150. After being trained, the hyperplane output by the SVM was obtained. This indicates that the trained model, i.e. hyperplane output which is including the important information, has the function to identify protein structural classes.

3.4 GENETIC ALGORITHM (GA)

Genetic algorithm is a part of **evolutionary computing**, which is a rapidly growing area of artificial intelligence. Genetic algorithm is inspired by Darwin's theory of evolution. In this case the problems are solved by an evolutionary process resulting in a best (fittest) solution (survivor).

Evolutionary computing was introduced in the 1960s by I. **Rechenberg** in his work "Evolution strategies". His idea was then developed by other researchers. Genetic Algorithm (GAs) was first imposed by **John Holland** and developed by his students and colleagues and are available in the book "Adaption in Natural and Artificial Systems" published in 1975.

3.4.1 BASIC PRINCIPLES OF GA

The Algorithm begins with a set of possible solutions called chromosomes which are used to assess the cost surface of the problem. The process can be thought of as solution breeding in that it creates a new generation of solutions by crossing two chromosomes. The solution variables or genes that provide a positive contribution to the population multiply and be passed through each subsequent generation until an optimal combination is obtained.

The population is updated after each learning cycle through three evolutionary processes: selection, crossover and mutation. These create the new generation of solution variables.

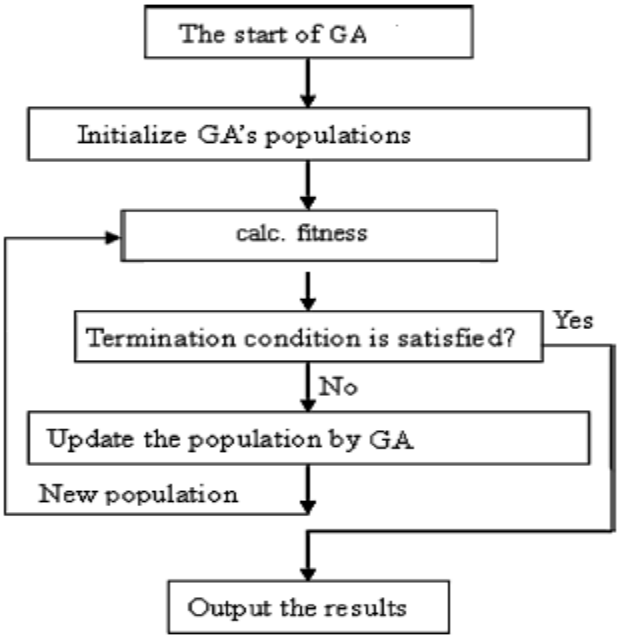


Fig 3.3 A GA iteration cycle

The selection function creates a mating pool of parent solution strings based upon the "survival of the fittest" criterion. From the mating pool the crossover operator exchanges gene information. This essentially crosses the more productive genes within the solution population to create an improved, more productive generation. Mutation randomly alters selected genes, which helps prevent premature convergence by pulling the population into unexplored areas of the solution surface and adds new gene information into the population.

3.4.2 OPERATORS OF GA

As one can see from the iteration cycle of genetic algorithm selection, crossover and mutation are the most important parts of the genetic algorithm. The performance is influenced mainly by these three operators.

- ***Encoding of a Chromosome***

A chromosome should in some way contain information about solution that it represents. The most used way of encoding is a binary string.

A chromosome then could look like this:

Chromosome 1 : - 1101100100110110

Chromosome 2 : - 1101111000011110

Fig 3.4 Chromosome

Each chromosome is represented by a binary string. Each bit in the string can represent some characteristics of the solution. There are many other ways of encoding. The encoding depends mainly on the solved problem. For example, one can encode directly integer or real numbers; sometimes it is useful to encode some permutations and so on.

- ***Initial population generation***

The initial population is generated randomly in the range of each parameter. Therefore, at the beginning of the separating procedure, *N* individuals are generated as random binary string. The GA starts with a group of chromosomes known as the population. The population has *Npop* chromosomes and is an *Npop* * *Nbits* matrix filled with random ones and zeros generated using

Pop =round (rand(*Npop*, *Nbits*));

where the function (*Npop*, *Nbits*) generates a *Npop* * *Nbits* matrix of uniform random numbers between zero and one. The function round rounds the numbers to the closest integer which in this case is either 0 or 1. Each row in the pop matrix is a chromosome.

- ***Evaluation of fitness***

After the initial population generations, the fitness of each individual is determined. Fitness is a numeric index to measure the effectiveness of each individuals of the population as a solution, which is usually utilized to select members of the population for reproduction.

- ***Selection Operation***

A pair of individuals is selected from the current population for mating using tournament selection.

In this, Selection procedure can be easily adjusted by changing the tournament size. If the tournament size is larger, weak individuals have a smaller chance to be selected.

Tournament selection pseudo code:

choose k (the tournament size) individuals from the population at random

choose the best individual from pool/tournament with probability p

choose the second best individual with probability p*(1-p)

choose the third best individual with probability p*((1-p)^2)

Deterministic tournament selection selects the best individual (when p=1) in any tournament. A 1-way tournament (k=1) selection is equivalent to random selection. The chosen individual can be removed from the population that the selection is made from if desired, otherwise individuals can be selected more than once for the next generation.

Tournament selection has several benefits: it is efficient to code, works on parallel architectures and allows the selection pressure to be easily adjusted.

- ***Crossover***

The crossover operator exchanges gene information between two selected chromosomes, where this operation aims to improve the diversity of the solution vectors. The pair of chromosomes, taken from the mating pool, becomes the parents of two offspring chromosomes for the new generation.

A binary crossover operation can be either single point or two point crossover. The simplest way to do single point crossover is to choose randomly some crossover point and copy everything before this point from the first parent and then copy everything after the crossover point from the other parent. In Fig.2.10 the sixth crossover position is randomly chosen, where the first position

corresponds to the left side. The bits from the right of the fifth bit are exchanged to produce two offspring chromosomes.

Single point crossover can be illustrated as follows:

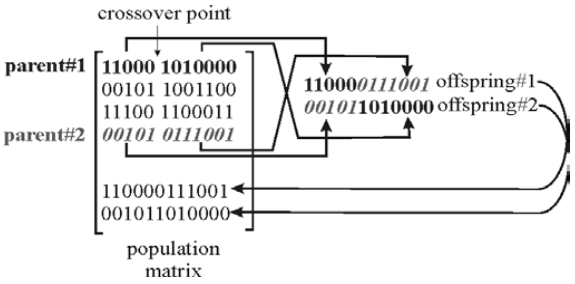


Fig 3.5 Single point Crossover

In two point crossover two points are randomly chosen and the bits in between them are exchanged.

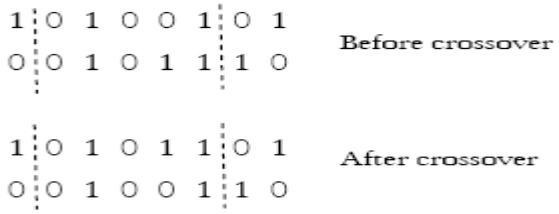


Fig 3.6 Double point Crossover

- ***Mutation Operation***

Random mutation operator is applied to the newly generated offspring to prevent from premature convergence. In this process randomly two parents are chosen and the binary bit positions are also selected randomly. Then mutations are done by changing the bit '0' to '1' in that position or vice versa with a probability expressed by, *m P*, where *m P* is called mutation probability. Thus a new set of offspring's are formed as shown below a '0' in the second bit changed to '1'.

1**1**001001 => 1**0**001001

After all the above processes the final set of best population are chosen from it which replaces the initial population thus continuing the cycle till the best fit population is obtained.

3.5 PARTICLE SWARM OPTIMIZATION (PSO)

3.5.1 BASIC CONCEPT OF PSO

The Particle Swarm Optimization (PSO) was developed by Eberhart and Kennedy in 1995 inspired by swarm intelligence theory such as birds flocking, fish schooling etc. It refers to a relatively new family of algorithms where the individuals evolved through generation by cooperation and competition among each other. In other evolutionary algorithms the evolutionary operators are used to manipulate individuals.

3.5.2 HOW ARE THE VALUES OF 'X AND Y' ARE UPDATED IN EVERY ITERATION?

The vector representation for updating the values for x and y is given in Figure 3.7. Let the position of the swarms be at 'a' and 'b' respectively as shown in the figure. Both are trying to reach the position 'e'. Let 'a' decides to move towards 'c' and 'b' decides to move towards'd'. The distance between the position 'c' and 'e' is greater than the distance between'd' and 'e'. So based on the neighbor's decision position'd' is treated as the common position decided by both 'a' and 'b'. (ie) the position 'c' is the individual decision taken by 'a', position 'd' is the individual decision taken by 'b' and the position 'd' is the common position decided by both 'a' and 'b'.

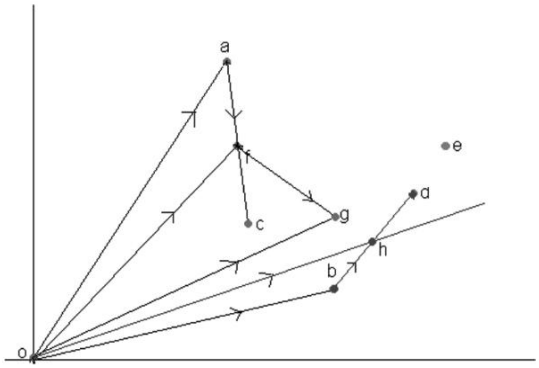


Fig 3.7 Vector Representation of PSO Algorithm

'a' based on the above knowledge, finally decides to move towards the position 'g' as the linear combination of 'oa' , 'ac' and 'ad' . [As 'd' is the common position decided].The linear combination of 'oa' and scaled 'ac' (ie) 'af' is the vector 'of'. The vector 'of' combined with vector 'fg' (ie) scaled version of 'ad' to get 'og' and hence final position decided by 'a' is 'g'. Similarly, 'b' decides the position 'h' as the final position. It is the linear combination of 'ob' and 'bh'(ie) scaled version of 'bd'. Note as 'd' is the common position decided by 'a' and 'b', the final position is decided by linear combinations of two vectors alone. Thus finally the swarms 'a' and 'b' moves towards the position 'g' and 'h' respectively for reaching the final destination position 'e'. The swarm 'a' and 'b' randomly select scaling value for linear combination. Note that 'oa' and 'ob' are scaled with 1 (ie) actual values are used without scaling. Thus the decision of the swarm 'a' to reach 'e' is decided by its own intuition along with its neighbor's intuition. Now let us consider three swarms (A,B,C) are trying to reach the particular destination point 'D'. A decides A', B decides B' and C decides C' as the next position. Let the distance between the B' and D is less compared with A'D and C' and hence, B' is treated as the global decision point to reach the destination faster. Thus the final decision taken by A is to move to the point, which is the linear combination of OA, AA' and AB'. Similarly the final decision taken by B is to move the point which is the linear combination of OB, BB'. The final decision taken by C is to move the point which is the linear combination of OC, CC' and CB'.

3.5.3 PSO ALGORITHM TO MAXIMIZE THE FUNCTION F (X, Y, Z)

1. Initialize the values for initial position a, b, c, d, e.
2. Initialize the next positions decided by the individual swarms as a', b', c', d' and e'.

3. Global decision regarding the next position is computed as follows.

Compute f (a', b, c, d, e), f (a, b', c, d, e), f (a, b, c', d, e), f (a, b, c, d', e) and f (a, b, c, d, e'). Find minimum among the computed values. If f (a',b, c, d, e) is minimum among all, the global position decided regarding the next position is a'. Similarly If f (a, b', c, d, e) is minimum among all, b' is decided as the global position regarding the next position to be shifted and so on. Let the selected global position is represented ad 'global'.

4. Next value for a is computed as the linear combination of 'a' , (a'-a) and (global-a) (ie)

- nexta = a+ C1 * RAND * (a' –a) + C2 * RAND * (global –a)
- nextb = b+ C1 * RAND * (b' –b) + C2 * RAND * (global –b)
- nextc = c+ C1 * RAND * (c' –c) + C2 * RAND * (global –c)
- nextd = d+ C1 * RAND * (d' –d) + C2 * RAND * (global –d)
- nexte = e+ C1 * RAND * (e' –e) + C2 * RAND * (global –e)

5. Change the current value for a, b, c, d and e as nexta, nextb, nextc, nextd and nexte.

6. If f (nexta, b, c, d, e) is less than f (a', b, c, d, e) then update the value for a' as nexta, otherwise a' is not changed.

If f (a, nextb, c, d, e) is less than f (a, b', c, d, e) then update the value for b' as nextb, otherwise b' is not changed.

If f (a, b, nextc, d, e) is less than f (a, b, c', d, e) then update the value for c' as nextc, otherwise c' is not changed.

If f (a, b, c, nextd, e) is less than f (a, b, c, d', e) then update the value for d' as nextd, otherwise d' is not changed.

If f (a, b, c, d, nexte) is less than f (a, b, c, d, e') then update the value for e' as nexte, otherwise e' is not changed.

7. Repeat the steps 3 to 6 for much iteration to reach the final decision.

The values for 'c1','c2' are decided based on the weightage given to individual decision and global decision respectively.

Let $\Delta a(t)$ is the change in the value for updating the value for 'a' in t iteration, then nexta at (t+1)th iteration can be computed using the following formula. This is considered as the velocity for updating the position of the swarm in every iteration. Where

$$nexta\ (t+1) = a(t) + \Delta a(t+1)$$

$$a(t+1) = c1 * rand * (a' - a) + c2 * rand * (gl\ obal - a) + w(t) * \Delta a(t)$$

'w (t)' is the weight at tth iteration. The value for 'w' is adjusted at every iteration as given below, where 'iter' is total number of iteration used.

$$w(t+1)=w(t)-t*w(t)/(iter)$$

Decision taken in the previous iteration is also used for deciding the next position to be shifted by the swarm. But as iteration increases, the contribution of the previous decision is decreases and finally reaches zero in the final iteration.

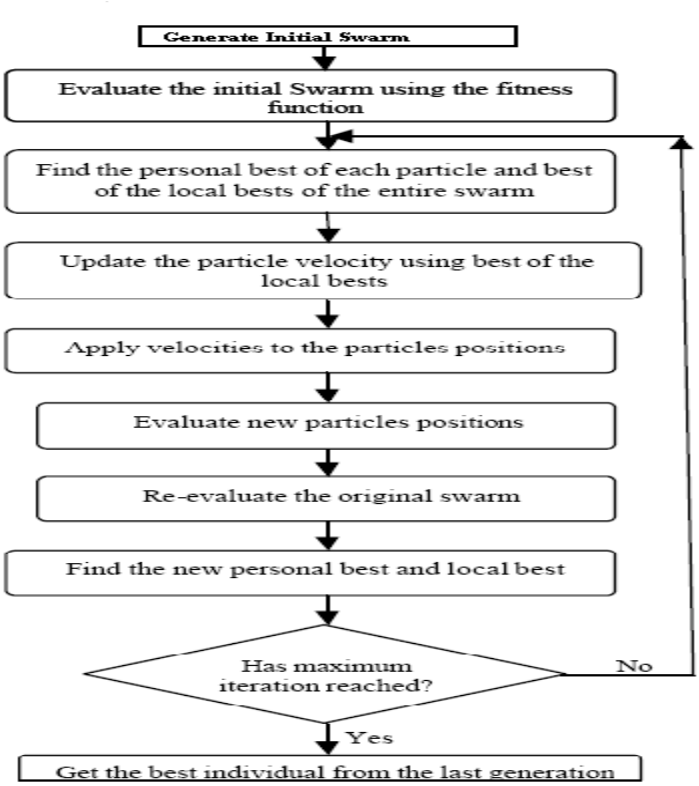


Fig 3.8 a PSO Iteration Cycle

3.6 GA AND PSO APPLIED TO PROTEIN STRUCTURAL CLASS PREDICTION

A protein can be represented by a 20-D unit vector through the following formulation. Suppose ***x*** is a protein to be predicted, and $f_i(x)$ $(i = 1, 2, \ldots, 20)$ represents the occurrence frequencies of

its 20 constituent amino acids. The compositions of the 20 amino acids in the protein ***x*** are given by the following normalization equation:

$$v_k(x) = \frac{f_k(x)}{\sum_{i=1}^{20} f_i(x)} \qquad (k = 1,2,\ldots,20).$$

Thus, in the composition space, the protein ***x*** can be expressed by the vector

$$\mathbf{v}(x) = [v_1(x), v_2(x), \ldots, v_{20}(x)].$$

Because **u1(x),** ***u2 (x),. . . , u20*** **(x)** are normalized as defined **,** the vector ***v(x)*** is a 20-D unit vector. Similarly, we can also use the following four standard unit vectors to represent the norms of the four protein structural classes:

$$\begin{cases} \mathbf{v}(\alpha) = [v_1(\alpha), v_2(\alpha), \ldots, v_{20}(\alpha)] \\ \mathbf{v}(\beta) = [v_1(\beta), v_2(\beta), \ldots, v_{20}(\beta)] \\ \mathbf{v}(\alpha+\beta) = [v_1(\alpha+\beta), v_2(\alpha+\beta), \ldots, v_{20}(\alpha+\beta)] \\ \mathbf{v}(\alpha/\beta) = [v_1(\alpha/\beta), v_2(\alpha/\beta), \ldots, v_{20}(\alpha/\beta)] \end{cases}$$

Where *u 1(a), u 2(a)., . . , u20(a)* represent the average compositions of the 20 amino acids occurring in the **19** known a-type proteins**,** ***u1(P) , u2 (p) , . . . , u20 (P)*** the average compositions of the 20 amino acids occurring in the **15** known &type proteins, and so forth**.**

Now initialize the GA parameters and run Genetic Algorithm to find out the structural classes in the protein. Below is the figure showing the GA process.

No of iteration=60

No of chromosomes=10

No of parameters to optimize=4

- Finding structural classes in proteins using PSO algorithm.

Starting particle positions and velocities were initialized at random. To reduce the problem of premature convergence to relative minima, the Attractive-Repulsive modification has been introduced. This modification defines a measure of global diversity (*D*) among the particles as:

$$D=\frac{1}{S}\sum_{i=1}^{S}\sqrt{\sum_{j=1}^{N}(w_{ij}-\overline{w}_j)^2}$$

where S is the number of particles in the swarm, N is space dimension (the number of networks weights) and $\overline{w}_j$ is the average of the parameter j among the particles.

If D falls below a minimal threshold (t_{min}) the update rule is inverted as follow

$$\begin{cases} v_{i,t+1}=\mu v_{i,t}+(-1)c_1(w_{i,t}^{best}-w_{i,t})+(-1)c_2(w_t^{global}-w_{i,t}) \\ w_{i,t+1}=w_{i,t}+v_{i,t+1} \end{cases}$$

causing the particles to spread in the phase space. If D reaches a maximal threshold (t_{max}) the update rule is restored as in the standard PSO method. We choose $t_{min} = 0.1$ and $t_{max} = 5.0$.

The parameters $c1$, $c2$ and μ were set as in the original PSO method as $c1 = c2 \in [0.0, 2.0]$ and $\mu = 0.7298$. The maximum number of iterations was set to 10000. A population size of 5 particles was chosen.

3.7 SIMULATION RESULTS FOR PROTEIN STRUCTURAL CLASSES STUDY

Table 3.2 Overall comparisons of results

METHODS	α	β	α + β	α / β	OVERALL
EUCLIDEAN DISTANCE	73%	82%	57%	49%	**67%**
HAMMING DISTANCE	71%	89%	57%	49%	**68%**
PRINCIPAL COMPONENT ANALYSIS(PCA)	82%	97%	78%	82%	**85%**

LOGIT-BOOST	90.4%	88.5%	73.9%	80%	**83.8%**
SUPPORT VECTOR MACHINE(SVM)	75%	90%	64%	64%	**74.5%**
GENETIC ALGORITHM(GA)	90%	94%	82%	80%	**86.5%**
PARTICLE SWARM OPTIMIZATION(PSO)	91%	94%	84%	80%	**87.5%**

3.8 INFERENCE DRAWN FROM SIMULATION IN PROTEIN STRUCTURAL CLASS PREDICTION

It can be seen from the simulation results that the prediction based on GA and PSO works very well and overall prediction also gives better results when compared to Euclidean Distance, Hamming Distance , Principal Component Analysis(PCA) , Logit-boost and Support vector machine(SVM).

PROTEIN CODING REGION IDENTIFICATION

4.1INTRODUCTION

Rapidly sequencing of genomic information of variety of organism have thrown considerable interest to both biology and computer science community. A major challenge for genomic research is to establish a relationship among sequences, structures and function of genes. In addition processing and analyzing this information are of prime importance. Basically genes are repositories for protein coding information and proteins in turn are responsible for most of the important biological functions in all cells. The traditional methods to analyze the genes in DNA based on the statistical and Discrete Fourier transform (DFT) are not robust, time consuming and unsuitable for future routine and rapid medical applications. In this chapter we propose a new method based on sliding DFT (SDFT) which is efficient and cost effective for DNA sequencing than the traditional Fourier Transform approach. In addition the DFT approach loses its effectiveness in case of small DNA sequences for which the autoregressive (AR) modeling has been used as an alternative tool. In this chapter we also propose an alternative but promising adaptive AR method for the same purpose and to improve the clarity of exon regions. Simulation study carried out on many DNA sequences subsequently reveals that a substantial savings in computation time is achieved by our methods without degrading the performance.

4.2 SLIDING DFT

The sliding DFT algorithm is an improved version of the traditional DFT process where the spectral bin output rate is equal to the input data rate on sample by sample basis with improved computational efficiency. The sliding DFT employs the idea of Goertzel algorithm and computes

the DFT spectra through implementing an infinite impulse response (IIR) filter. The Z-transform function of the Goertzel filter is given by

$$H_G(z) = \frac{1 - e^{-j2\pi k/N} z^{-1}}{1 - 2\cos\frac{2\pi k}{N} z^{-1} + z^{-2}}$$

Where N is no. of samples of which the DFT is to be calculated. The sliding DFT (SDFT) algorithm performs an N-point DFT on time samples within a sliding window. The time window is then advanced by one or more samples and a new N-point DFT is calculated. The importance of this process is that each new DFT is efficiently computed directly from the result of the previous DFT. The principle used in the SDFT is the DFT shifting theorem or the circular shift property. It states that if the DFT of a windowed time domain sequence is X(k), then the DFT of that sequence, circularly shifted by one sample, is $X(k)e^{j2\pi k/N}$, where k is the DFT bin of interest. This process is expressed as

$$P(k) = P(k-1)e^{(j2\pi k/N)} - x(n-N) + x(n)$$

Where x(n) is the new sample to be included in the windowed sequence, x(n-N) is the first sample of the sequence which is to be discarded and P(k); P(k-1) are the new and previous spectral components of the windowed time sequence respectively. This can be represented by an IIR filter and the Z-domain transform function of the sliding DFT filter is given by

$$H_{SDFT}(z) = \frac{1 - z^{-N}}{1 - e^{j2\pi k/N} z^{-1}}$$

A better representation of the algorithm with correct phase and magnitude is given in

$$P(k) = e^{(j2\pi k/N)}[P(k-1) - x(n-N) + x(n)]$$

The IIR filter implementation of SDFT is shown in Fig.4.1

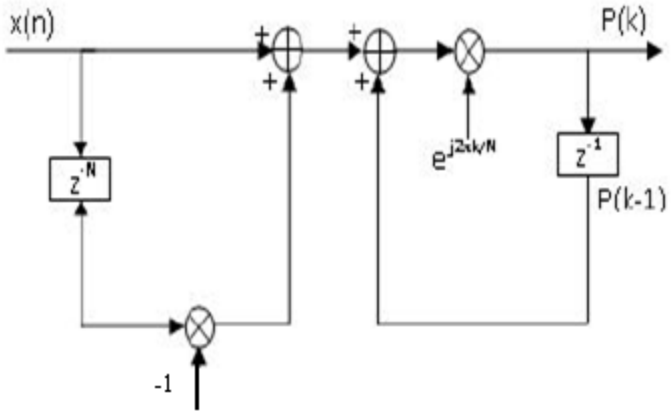


Fig 4.1 An IIR implementation of the Sliding DFT

4.3 AUTOREGRESSIVE MODELING

The Fourier analysis based methods for detecting the protein coding regions in DNA lose their effectiveness in the case of small as well as standard DNA sequences. These methods are constrained by the frequency resolution and spectral leakage effects of the data record. Hence to overcome this problem model based approaches have been evolved which look the spectral analysis problem differently compared to Fourier Transform methods. The autoregressive (AR) model is one such method which has been proposed for detection for the 3-base periodicity in DNA sequence and to predict the protein coding regions. The AR modeling approach has been used successfully in speech processing and radar signal processing. It is a simple and robust method and requires no a priori knowledge of the sequence to be analyzed and also works well with low signal to noise (SNR) ratio. In AR modeling the observed signal x(n) can be modeled as a linear combination of its M past output values x(n - k) and present input u(n) as

$$x(n) = -\sum_{m=1}^{M} a(m)x(n-k) + u(n)$$

Where a(k) represents the coefficient of the model to be estimated. This AR process can be viewed as a recursive all pole digital filter whose transfer function is

$$H(z) = \frac{1}{1 + \sum_{m=1}^{M} a(m) z^{-k}}$$

In this case the number of filter coefficients is equal to the order of the model and is same as the number of filter poles which is to be determined efficiently. The Yule-Walker or Burg methods are usually used to determine these coefficients using the Levinson-Durbin recursive procedure. The spectral estimation for the AR model given as

$$S_x(k) = \frac{\sigma^2}{\left| 1 + \sum_{m=1}^{M} a(m) e^{j2\pi mk/N} \right|^2}$$

where σ^2 is the variance of the input signal.

4.4 NUMERICAL REPRESENTATION AND PRELIMINARY SPECTRAL MEASURE FOR CODING REGIONS

To perform the gene prediction based on period-3 property the total DNA sequence is first converted into four indicator sequences, one for each base. The DNA sequence D(n) is mapped into binary signals uA(n); uC(n); uG(n) and uT (n),which indicate the presence or absence of these nucleotides at location n. For example the binary signal uA(n), attributed to nucleotide A takes a value of 1 at n = n0 if D(n0) = A, else uA(n0) is 0.

Suppose the DNA sequence is represented as

D(n) = [ATGATCGCAT]

Then its numerical representation is given by

u_A(n)=[1 0 0 1 0 0 0 0 1 0]

u_C(n)=[0 0 0 0 0 1 0 1 0 0]

u_G(n)=[0 0 1 0 0 0 1 0 0 0]

u_T(n)=[0 1 0 0 1 0 0 0 0 1]

u_A (n)+ u_C(n)+ u_G(n)+ u_T(n)=1 for n=0,1,2…….,N-1

4.5 MEASURE OF SPECTRAL CONTENT USING SDFT

In this frequency domain method, the DFTs of the four binary indicator sequences are employed to exploit the 3-base periodicity. Let UA[k]; UG[k]; UC[k] and UT [k] represent the Discrete Fourier transform (DFT) of the corresponding binary sequences and is given by

$$U_X[k] = \sum_{n=0}^{N-1} u_X[n] e^{-j2\pi kn/N}$$

for x = A;C;G or T and k = 0; 1; 2; _ _ _ ;N - 1.Then the spectral content at k is given by

$$S[k] = \{U_A[k]\}^2 + \{U_B[k]\}^2 + \{U_C[k]\}^2 + \{U_D[k]\}^2$$

S[k]acts as a preliminary indicator of a coding region giving a peak at the $2\pi/3$ frequency. This coding procedure can be used to detect the probable coding region in the DNA sequence. The progression of S[N/3] can be plotted by evaluating S[N/3] over a window of N samples, then sliding the window by one or more samples and recalculating S[N/3].This process is carried out over the entire DNA sequence. It is necessary that the window length N be sufficiently large (typical sizes are a few hundred eg 351 to a few thousand) so that the periodicity effect dominates the background noise spectrum. The conventional DFT method involves large computations which poses difficulty for online evolution of protein coding regions. The following section deals with the sliding DFT method, a fast approach in DSP literature for spectral estimation of the DNA sequences.

Each of the four binary indicator sequences within the window is passed through this IIR filter and the corresponding DFT output sequence is obtained. The resulting DFTs are represented as PA[k]; PC[k]; PG[k] and PT [k]. Then the spectral content is computed as

$$S[k] = \Sigma [Px(k)]^2$$

where x stands for either of the sequences A, C, T or G. The successive progression of S[k] within the sliding window and the plot of S[N/3] exhibit the coding regions in DNA. Thus the SDFT requires only one complex multiplication and two real additions per output sample. Hence the computational complexity of each successive N-point output is O(N) for the sliding DFT compared to $O(N^2)$ for the DFT. As a result there is substantial computational saving in identifying the coding region by the proposed method.

4.6 PROPOSED ADAPTIVE AR MODELING APPROACH

For achieving online prediction of gene and exon the computational time needs to be reduced. Further the fixed AR method requires all data to be available simultaneously which is not always feasible. With a motive to alleviate these limitations an adaptive AR model based approach is suggested in this section for efficient prediction. The AR process can be viewed as an adaptive prediction error filter (all-zero filter) that adaptively adjusts its coefficients to flatten the spectrum of the signal to be observed. It is a fact that with a proper learning algorithm like LMS and RLS the weight vector of the adaptive prediction error filter converges to optimal AR coefficients.

4.6.1 THE LMS PREDICTION ERROR FILTER

A signal x(n) modeled as a M order AR process can be expressed as

$$\hat{x}(n) = \sum_{m=1}^{M} w_m(n)x(n-m) + e(n)$$

Where e(n) is the prediction error and W_1; W_2;……. ; W_m are AR coefficients.

The LMS prediction error filter illustrated in Fig. 2 is used to adaptively estimate the optimal AR coefficients by minimizing the mean square value of prediction error.

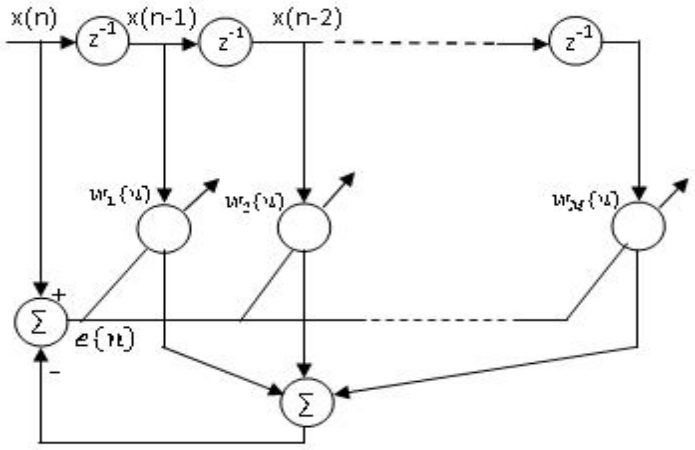


Fig 4.2 LMS Prediction Error Filter

The weight update equation of the filter is given by and μ is the step size that determines the rate of converge and stability of weights.

The power spectra is estimated using the prediction error filter is given as

$$S_x(k) = \frac{\sigma_\epsilon^2}{\left|1 - \sum_{m=1}^{M} w_m e^{j2\pi mk/N}\right|^2}$$

where σ_ϵ^2 is the variance of the prediction error signal.

4.7 RESULTS

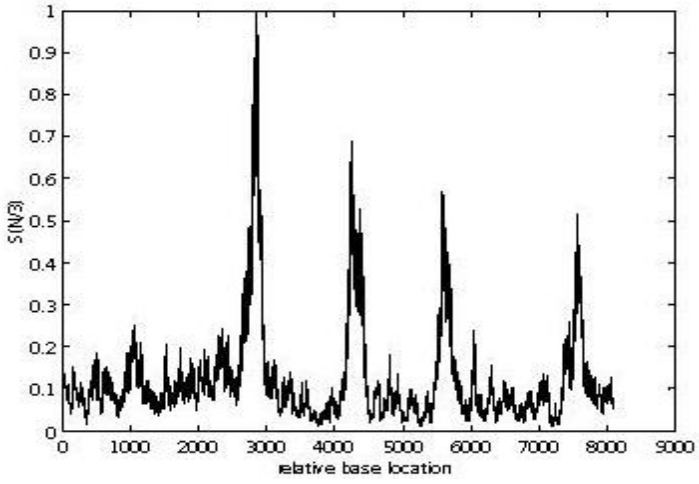


Fig 4.3 Spectral plot of S[N/3] for gene F56F11.4a in the C-elegans chromosome III using DFT

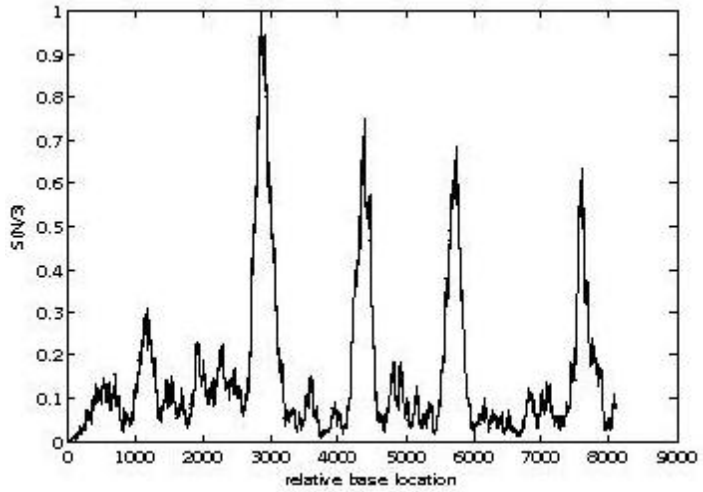


Fig 4.4 Spectral plot of S[N/3] for gene F56F11.4a in the C-elegans chromosome III using SDFT

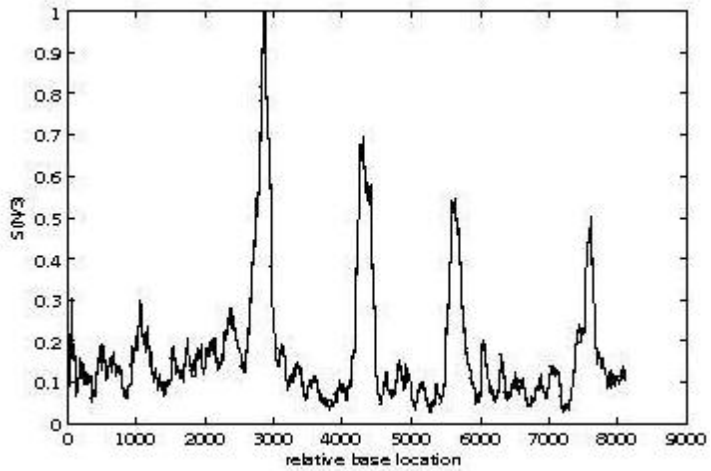


Fig 4.5 Spectral plot of gene F56F11.4a in the C-elegans chromosome III using Fixed AR modeling

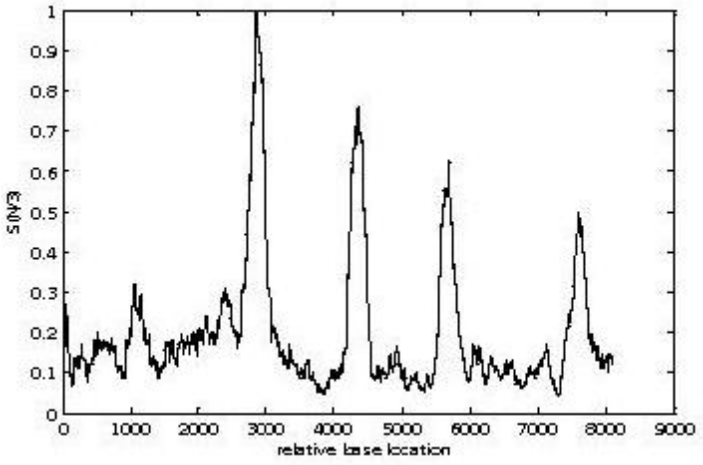


Fig 4.6.Spectral plot of gene F56F11.4a in the C-elegans chromosome III using adaptive AR modeling

SCOPE FOR FUTURE WORK

5.1 SCOPE FOR FUTURE WORK

This kind of computational approach like Docking and QSAR will help us to extract the important descriptors from the existing drug series to add up a new molecule in the SARS Protease drug list and to combat with fast drug resistance property of the virus. Thus it can be inferred that the extracted descriptors and total negative binding energy play a major role in the maintenance of biological activity. Also the descriptors encoding significant structural information such as topological and constitutional environment can be used to get unique characteristics of compounds to build the relationship between the structure and biological activity.

Owing to the fuzzy nature of the protein structural class definition and the possible experimental errors in protein data bank, the naive statistical mathematical methods might not work well for the protein structural class prediction. Application of evolutionary algorithms like BFO and improved GA can be applied to predict the structural class of proteins and coding regions more efficiently.

PUBLICATIONS FROM THESIS

PUBLISHED

Sahu Sitanshu Sekhar; Panda Ganapati and Mahapatra Chinmaya."Development of a new low complexity ANN based Drug discovery method for efficient design of SARS Co3CL Protease Inhibitors". *2nd International Conference on Environmental Research, BITS Pilani, Goa, India, vol.2, 2008.*

Panda Ganapati; Mahapatra Chinmaya; Bissoyi Akalabya. "In silico Prediction of Lead Molecules for Migraine Using Blind Docking and Identifying the Active Amino Acids Contributing to the Protein-Ligand Interaction". *National symposium of Bioinformatics, NIT Raipur, India, 2009.*

References

1. **J Lee, N., Hui, D.Wu, A., Chan, P. Cameron, P. Joynt, G. M.; Ahuja,A. Yung, M. Y.; Leung, C. B.; To, K. F.; Lui, S. F.; Szeto, C. C.;Chung, S.; Sung, J. J. A major outbreak of severe acute respiratory syndrome in Hong Kong. *N. Engl. J. Med.* 2003, *348*, 1986-94.**

2. **Keng-Chang Tsai,Shih-Yuan Chen, Po-Huang Liang, I-Lin Lu,Neeraj Mahindroo,Hsing-Pang Hsieh,Yu-Sheng Chao,Lincoln Liu, Donald Liu, Wei Lien, Thy-Hou Lin, and Su-Ying Wu, Discovery of a Novel Family of SARS-CoV Protease Inhibitors by Virtual Screening and 3D-QSAR Studies, *J. Med. Chem.* 2006, *49*, 3485-3495**

3. **Thiel, V.; Ivanov, K. A.; Putics, A.; Hertzig, T.; Schelle, B.; Bayer, S.; Weissbrich, B.; Snijder, E. J.; Rabenau, H.; Doerr, H. W.; Gorbalenya, A. E.; Ziebuhr, J. Mechanisms and enzymes involved in SARS coronavirus genome expression. *J. Gen. Virol.* 2003, *84*, 2305-2315.**

4. **Chou, K. C.; Wei, D. Q.; Zhong, W. Z. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. *Biochem. Biophys. Res. Commun.* 2003, *308*,148-151.**

5. **Zhang, X. W.; Yap, Y. L.; Altmeyer, R. M. Generation of predictive pharmacophore model for SARS-coronavirus main proteinase. *Eur.J. Med. Chem.* 2005, *40*, 57-62.**

6. **Anand, K.; Ziebuhr, J.; Wadhwani, P.; Mesters, J. R.; Hilgenfeld, R. Corona virus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* 2003, *300*, 1763-1767.**

7. **Bacha, U.; Barrila, J.; Velazquez-Campoy, A.; Leavitt, S. A.; Freire, E. Identification of novel inhibitors of the SARS coronavirus main protease 3CLpro. *Biochemistry* 2004, *43*, 4906-4912.**

8. **Anand, K.,Palm, G. ,J. Mesters, J. R. Siddell, S. G. Ziebuhr, J .Hilgenfeld, R. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *EMBO J.* 2002, *21*, 3213-3224.**

9. **Latifa Douali,Didier Villemin, and Driss Cherqaoui," Neural Networks: Accurate Nonlinear QSAR Model for HEPT Derivatives", *J. Chem. Inf. Comput. Sci.* 2003, *43*, 1200-1207**

10. **Jenwitheesuk, E.; Samudrala, R. Identifying inhibitors of the SARS coronavirus proteinase. *Bioorg. Med. Chem. Lett.* 2003, *13*, 3989-3992.**

11. **Leng, Q.; Bentwich, Z. A novel coronavirus and SARS. *N. Engl. J. Med.* 2003, *349*, 709.**

12. **Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Quantitative Structure-Activity Relationship Analysis. *J. Med. Chem.* 1990, *33*, 2583-2590.**

13. **Bishop, C. M. Neural Networks and their Applications. *Re*V. *Sci. Instrum.* 1994, *65*, 1803-1832.**

14. **Sirois, S.; Wei, D. Q.; Du, Q.; Chou, K. C. Virtual screening for SARS CoV protease based on KZ7088 pharmacophore points. *J.Chem. Inf. Comput. Sci.* 2004, *44*, 1111-1122.**

15. **G. Schneider, P. Wrede, Artificial neural networks for computer-based molecular design, Prog. Biophys. Mol. Biol. 70 (1998) 175–222.**

16. **Patra,J.C. Pal ,R,N. Chatterji,B,N. Panda,G. "Identification of nonlinear dynamic systems using functional link artificial neural networks"IEEE Transactions , Systems, Man and Cybernetics, Part B, Vol 29 ,pp 254 - 262,April-1999**

17. **A.Namatame and N.Ueda."Pattern classification with Chebyshev neural networks," Ind.J.Neural Networks Vol 3,pp23-31,Mar. 1992.**

18. **Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed. Engl.* 1993, *32*, 503-527.**

19. **Manallak, D. T.; Ellis, D. D.; Livingstone, D. J. Analysis of linear and nonlinear QSAR data using neural networks. *J. Med. Chem.* 1994, *37*, 3758.**

20. **J. J. Ding, "time-frequency analysis and wavelet transform course note," the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, 2007.**

21. **R. G. Stockwell, L. Mansinha, and R. P. Lowe, "Localization of the complex spectrum: the S transform," IEEE Trans. Signal Processing, vol.44, no.4, pp.998-1001, Apr.1996.**

22. **Hongmei Zhu, PhD and J. Ross Mitchell, PhD, "The S Transform in Medical Imaging," University of Calgary Seaman Family MR Research Centre Foothills Medical Centre, Canada.**

23. **Jaya Bharata Reddy, Dusmanta Kumar Mohanta, and B. M. Karan, "Power system disturbance recognition using wavelet and s-transform techniques," Birla institute of Technology, Mesra, Ranchi-835215, 2004.**

24. **B. Boashash, "Notes on the use of the wigner distribution for time frequency signal analysis", IEEE Trans. on Acoust. Speech. and Signal Processing , vol. 26, no. 9, 1987**

25. **R. N. Bracewell, The Fourier Transform and Its Applications , McGrawHill Book Company, New York, 1978**

26. **E. O. Brigham, The Fast Fourier Transform , Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1974**

27. **L. Cohen, "Time-frequency distributions - A review", Proc. IEEE, vol. 77, no. 7, July 1989**

28. **I.Daubechies, "The wavelet transform, time-frequency localization and signal analysis", IEEE Trans. on Information Theory, vol. 36, no. 5, Sept. 1990**

29. **M. Farge, "Wavelet transforms and their application to turbulence", Annual Review of Fluid Mechanics, vol. 24, pp. 395-457, 1992**

30. **D. Gabor, "Theory of communication", J. Inst. Elect. Eng. , vol. 93, no. 3, pp. 429-457, 1946**

31. **P. Goupillaud, A. Grossmann, and J. Morlet, "Cycle-octave and related transforms in seismic analysis", Geoexploration, vol. 23 pp. 85-102, 1984**

32. **F. Hlawatsch and G. F. Boudreuax-Bartels, "Linear and quadratic timefrequency signal representations", IEEE SP Magazine, pp. 21-67, April 1992**

33. O. Rioul and M. Vetterli, "Wavelets and signal processing", IEEE SP Magazine, vol. 8 pp. 14-38, 1991

34. R. K. Young, Wavelet Theory and its Applications, Kluwer Academic Publishers, Dordrecht,1993

35. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M.. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535-542, 1977

36. Brutlag, D.L., Dautricourt, J.-P., Maulik, *S.*, & Relph, J. Sensitive similarity searches of biological sequence databases*C. omput.Appl. Biosci. 3*, 237-245,1990

37. Chou, P.Y. Prediction of protein structural classes from amino acid compositions. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., Ed.), pp. 549-586. Plenum Press, New York.,1989

38. Deleage, G. & Roux, B. An algorithm for prctein secondary structure prediction based on class prediction. *Protein Eng. I*, DeSieno, D. (1988). Adding a conscience to competitive learning. *Proc Second Annu. IEEE Intl. Conference on Neural Networks 1*, 117-124,1987

39. Metfessel, B. & Saurugger, P.N. Pattern-recognition in the prediction of protein structural class. Proc. 26th Annu. Hawaii Intl. Conference on System Sciences, pp. 679-688,1993

40. Nakashima, H., Nishikawa, K., & Ooi, T. The folding type of a protein is relevant to its amino acid composition. J. Biochem.tokyo,153-19,1986

41. Qian, N. & Sejnowski, J. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202,865-884, 1988

42. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., & Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins.Science 229, 834-838, 1985

43. Rumelhart, D.E., Hinton, G.E., & Williams, R.J. Learning internal representations by error propagation. In Parallel Distributed Processing, Vol. 1, pp. 318-362. MIT Press, Cambridge, Massachusetts,1986

44. Saurugger, P.N. & Metfessel, B.A. Patterns in protein primary sequences: Classification, display and analysis. Proc. 15th Annu. Symposium on Computer Applications in Medical Care, pp.299-303,1991

45. Sander, C. & Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment,1991

46. Schiffer, M. & Edmundson, A. Use of helical wheels to represent the structures of proteins and to identify segments within. Biophys. J. 7, 121-135,1967

47. Sheridan, R.P., Dixon, J.S., Venkataghavan, R., Kuntz, I.D., Scott, K.P. Amino acid composition and hydrophobicity patterns of protein domains correlate with their structure polymers24,1995-2023,1985

48. Widman, L.E. & Pierce, B.S., 111. Comparison of a knowledge based pattern classification algorithm with the Mahalanobis statistical clustering algorithm. Proc. 14th Annu. Symposium on Computer Applications in Medical Care, pp. 529-533,1990

49. Zhang, 12.-T. & Chou, K.-C. An optimization approach to Predicting protein structural class from amino acid composition. protein secondary structure prediction. J. Mol. Bi01. 225, 1049-1063,1992

50. Zhou, G.P., Assa-Munt, N., Some insights into protein structural class prediction. PROTEINS: Struct. Funct. Genet. 44,57–59,2001

51. Zhou, G.P., Doctor, K., Subcellular location prediction of apoptosis proteins. PROTEINS: Struct. Funct. Genet. 50,44–48,2003

52. Zhou, Z.H., Jiang, Y., Yang, Y.B., Chen, S.F., Lung cancer cell identification based on artificial neural network ensembles. Artif.Intel. Med. 24, 25–36,2002