International Journal of Advanced Trends in Computer Applications
www.ijatca.com

Analysis of Protein Structure based Drug Design Using Comparative Molecular Field Analysis

[1]Mrs. B.Kalaiselvi, [2]Dr.M.Thangamani
[1]Assistant Professor,
Mahendra Engineering College for Women
Namakkal, Tamilnadu,
India.
[2]Assistant Professor,
Kongu Engineering College,
Perundurai, Tamilnadu,
India.
[1]kalairs2003@gmail.com, [2]manithangamani2@gmail.com

Abstract: *Proteins are molecular devices in the nanometer scale where biological function is exerted the building blocks of all cells in our bodies. DNA molecule is encoded in the dynamic process of life maintenance; replication, defense and reproduction are carried out by proteins. There are twenty natural amino acids whose frequency is higher and with the particular functions are forming polypeptide chains or proteins in different ways determined by the genetic code and limited by stereo chemical properties. Molecular Function is describes the tasks performed by individual proteins and can be broadly divided into twelve subcategories such as cellular processes, metabolism, DNA replication/modification, transcription/translation, intracellular signaling, cell-cell communication, protein folding/degradation, transport, multifunctional proteins, cytoskeletal/structural, defense and immunity, and miscellaneous functions. Biochemical reactions of cell breathing, oxygen and carbonic gas transport, food absorption, energy usage, energy storage, heat or cold physiological reactions, or any life process one can carry out by a protein or a protein complex. The primary structure of a protein referred to its complete covalent structure but it is more frequently interpreted as the sequence of amino acids of each polypeptide chain of which the protein is composed. All processes in a living organism have proteins are acting as the developed under the natural selection. All proteins functions are dependent on their structure which depends on physical and chemical parameters. The genome sequences make it more difficult to identify our target sequence to find the putative gene directly or DNA sequences containing the putative gene. The bioinformatics have been working together in a new area known as molecules, classical biological, physical, chemical, mathematical and informatics to allow a new level of knowledge about life organization.*

Keywords: *Acid sequence, polypeptide chain, terminus, GenBank.*

I. INTRODUCTION

Proteins have viewed the protein of interest had its coding sequence identified and cloned in proper expression vector. The cloning, expression and purification of pure proteins quantities are used to prepare solutions for NMR spectroscopy or to grow crystals for structure determination by X-ray crystallography from single proteins to whole set of proteins of an organism. Bioinformatics tools are used to perform the tasks in all biological databases of which GenBank at NCBI and the RCSB/PDB for sequences and structures respectively. The GenBank database contains the available DNA sequences for more than 205,000 different species are obtained largely from large-scale sequencing projects. The Research Collaboratory for Structural Bioinformatics (RCSB) and Protein Data Bank (PDB) includes tools and resources for understanding the relationship between sequence, structure and function of biological macromolecules. The GenBank nr sequences are use a 90% sequence identity the cutoff to eliminate redundancy in the protein structure data bank with 12,611 distinct sets of structures. It assume the different types of known protein folds and compare it to the different types of known protein sequences from the nr database of GenBank conclude only 0.38% of the available protein sequences are folds. Protein sequences can be determined (from CDS) as compared to the determination of novel protein 3D structures. The computational methods are complementary to experimental methods for the determination of protein

structures and these can be provided by structural bioinformatics efforts. Bioinformatics structure is biology in terms of molecules sense of physical chemistry and applying informatics techniques derived from mathematics, computer science and statistics to organize and explore the structural information associated to these molecules on a large scale. The computational methods for protein structure including homology modeling, fold recognition via threading, and ab initio methods. In the following sections the hierarchy of protein structure and its application to drug design in tropical disease will be described. The last section is to apply several bioinformatics tools starting from a partial DNA sequence or protein accession number and yielding a model of its structure and function.

II. RELATED WORKS

Protein structure describes more complex levels of organization typically uses the four levels are referred to as the primary, secondary, tertiary and quaternary structure of a protein. The intervening levels between the secondary and tertiary structures which are referred to as super secondary structures and domains of the three-dimensional structure of proteins for the drug design process.

2.1 Polypeptide Chain Properties

A polypeptide chain is a uni-dimensional heteropolymer composed of amino acid residues. The first twenty amino acids are directly encoded by the corresponding gene in the stop codons can be used for the incorporation of two additional amino acids. The amino acids are α-amino acids which possess the generic structure to all such amino acids is the amino group, carboxylic acid group and hydrogen bound to the central carbon atom. With the twenty different amino acids to produce the wide diversity of functions of proteins perform in living organisms. The α-carbon is bound to four different chemical moieties amongst the naturally occurring amino acids whose R-group hydrogen atom is making the α-carbon symmetric. The remaining amino acids are always found to be L-amino acids has chiral structures observed in proteins at all levels of the hierarchy. Polypeptide chain is generated by a series of condensation reactions between the carboxyl group of one amino acid and the amino group of the next yielding a covalent bond is the trivial name of a peptide bond of polypeptide chains. The union of two amino acids results possesses a free amino group (N-terminus) and carboxyl group (C-terminus) allowing the condensation reaction to continue ad infinitum in both directions. The polypeptides observe in the extant species are the products selected from a tiny fraction of all possible combinations that were expressed along the evaluative process.

The C-N distance is shorter and longer than a normal single bond and a normal double bond. The α-carbon atoms of two adjacent amino acids together with the carbonyl (C=O) and NH groups of the peptide group itself within the same plane. The torsion angle associated with this bond is termed ω which in order to be planar must be either 180^o (Trans) or 0^o (cis). Cis peptide bonds are common when the amino acid proline lies on the C-terminal side of the bond which forms a covalent bond with its own main chain nitrogen generating a closed ring. Polypeptides obey the standard stereochemistry and all the bond lengths, angles are effectively fixed, varying only minimally around their standard values. Polypeptide possesses comes from the torsional rotation around its single bonds is not freely routable and ω is effectively fixed in the trans configuration. The two remaining single bonds along with the main chain and from these bonds may associate the torsion angles ϕ and ψ. The two-dimensional coordinate system represents the both ϕ and ψ vary between -180 and $+180^o$.

2.2 Representation of Protein structure

If the ϕ and ψ angles are repeated systematically for all residues within a stretch of polypeptide result will be a helix. An infinite number of types of helix depend on the combination of ϕ and ψ. The two parameters n and d are represents the number of residues per helical turn and the shift parallel to the helical axis per residue relationships which relate ϕ and ψ to n and d. The globular proteins to the α-helix ($\phi \approx -63^o$, $\psi \approx -42^o$, n = +3.6, d = 1.5Å) and the β-strand ($\phi \approx -120^o$ and $\psi \approx +135^o$, n = -2.3, d = 3.3Å). The negative n value in the β-strand indicates that is left-handed helix and the α-helix are right-handed. The presence of consecutive hydrogen bonds between the carbonyl of residue I and the NH of residue i+4 and the radius of the helix is compatible with Van de Waals contacts across the helical axis. The possibility of forming ion pairs (salt bridges) between oppositely charged residues separated by 3 or 4 residues along the sequence. Due to the ideal separation between donor (nitrogen) and acceptor (oxygen) atoms is linearity and dipolar alignment.

The alignment of the dipoles is peptide unit leads to a macroscopic dipole associated with the whole helix is the order of 0.5 units of positive charge at the N-terminus and 0.5 units of negative charge at the C-terminus. The free NH groups within the first turn makes the N-termini of α-helices ideal binding sites for anions as phosphate. In the absence of N-terminus is

capped by the side chain of the first helical residue (the N-cap) is often an asparagine or aspartic acid. The β-strands are unable to form internal hydrogen bonds assemble into sheets that may be parallel, ant parallel or mixed in nature depending on the relative orientation of the strands are composed by the formation between strands. The β-sheets are twisted to view perpendicular to the strands has a left-handed twist and viewed parallel to the strands it is right-handed. Planar sheets (with n = 2) do exist but are rare compared with twisted ones. The globular proteins are α-helices and β-strands will traverse the structure have linear axes and other types of structure are needed to cause chain reversal.

2.3 CATH Classification Process

The individual elements of secondary structure of a polypeptide must pack against one another in order to form stable, compact and biologically active tertiary structure. The folding process result is the full three-dimensional structure of the polypeptide depends on the amino acid sequence and the atomic details of the structure. The structure is removing the atomic detail and trace of the main chain in space referred to as the fold of the protein exist and two of the most widely used are SCOP and CATH use a hierarchical approach classifies folds into classes in the highest level based on secondary structure content. The second layer of the CATH classification hierarchy deals with different architecture that describes the overall arrangement of secondary structural elements but without concern for their connectivity within a given class of 30 major types. Some folds are more common that have been termed super folds include the TIM barrel, the jelly-roll and the up-and-down 4-helix bundle. The asymmetric α-carbon atom can be seen at most if not all levels of protein structure are the right-handed α-helix is not energetically favored over its left-handed partner nor t β-strands are left-handed. The right-handed α-helices using side chain inter-digitations leads to the common packing angle of 20^{o} these helices coil around one another to form coiled structures will be left-handed. At the right handed connections found in βxβ units with the twist on β-strands means that all $(\beta\alpha)_8$ barrels (TIM barrels) have the same chirality.

Proteins are composed only single polypeptide chain and the homo- or hetero-oligomers or macromolecular assemblies or polymers. The recent evaluation of proteins encoded by the E. coli genome is 20% expected to be monomeric and nearly 40% of all proteins followed by tetramers is 21%. Two polypeptide chains subunits interact to form an interface two possibilities exist and interface may be isologous or heterologous may be either symmetric, leading to ring-like structures possessing symmetry axes of higher order (>2) or asymmetric. The resulting structures are generally spherical and used as storage molecules or as viral capsids. The translational symmetry can be added to the rotation axes to produce helical symmetry as in microtubles, tobacco mosaic virus and other filamentous structure. A single molecule introduces the possibility for a cross-linking activity and mentioned many oligomeric structures serve to generate cavities with a specific biological function.

III. DOCKING SEARCH ALGORITHM

The 3D structures of proteins are developed before the techniques are obtained by required exhaustive procedures of syntheses and evaluation. The drugs were discovered by modification of compounds with known biological activities that discovered by chance or random screening based on trial and error. In the traditional medicinal chemists focused the relationships between structures and activities of known active compounds and proposing modifications would improve some biological property of the molecule. The techniques developed for visualizing, manipulating and superposing 3D structures of active compounds in order to obtain faster information about the structure-activities relationship of known active compounds. Different computer based Quantitative Structure-Activities Relationship (QSAR) approaches like HQSAR, CoMFA and related 3D QSAR approaches became popular amongst medicinal chemists. The structure-based ligand design still faces several major challenges and limitations such as a lack of availability of 3D structures for many receptors, membrane-bound receptors, adequate computational approaches to deal with induced-fit in ligand-receptor interactions and selectivity between different isoforms of the same receptor. The ligand design stage of the drug discovery process is the initial steps in multidisciplinary effort that requires several rounds of refinement between the identification of an active principle and the selection of a viable drug candidate. Structural based design can assist the continuous process by pinpointing opportunities for structural modifications do not interfere with binding may improve affinity and selectivity properties of the molecule.

A large number of conformations are generated for the small molecule either prior to docking or during the docking routine. Each conformation is positioned in the active site in a variety of orientations the combination of conformation and orientation being known as a "pose". Many poses are selected and ranked by a scoring function in order to determine the best overall pose. To make the docking problem tractable most

programs address adequately the question of ligand flexibility and its scoring functions to score and rank multiple ligands against a single target according to their experimental binding affinities. The scoring function that can select the correct pose for one ligand will not necessarily rank a series of potential ligands in the correct order. The important that search parameters are chosen that give docking speeds useful for virtual screening applications with an acceptable loss in docking accuracy.

3. 1 Protein structures with Virtual Screening and 3D-QSAR

Virtual screening results are dependent on database need to be tuned and that means the effort should be put into filtering out molecules with undesirable physical properties and chemical functionalities. Virtual screening is common to have available more than one structure of the same protein in complex with different ligands or without any ligand bound. Procheck and Whatcheck can be used to analyze the stereochemistry and chemical environment of each residue of the protein. The virtual screening procedures usually take days to finish and if a mistake is made in the preparation stages then the entire work may be compromised and process especially during the visual inspection stage.

Protein structures can be used to improve the results of Quantitative Structure-Activity Relationship (QSAR) techniques. Two most common techniques are the Comparative Molecular Field Analysis (CoMFA) and the Comparative Molecular Similarity Indices Analysis (CoMSIA). Both methods assume the ligands have binding mode and 3D structures are Partial Least Square (PLS) analysis is performed to try to find a correlation between the compounds biological activities and their 3D structures. The target protein structure is a valid approach is to dock the ligands into the protein biding site is use the result as structural alignment can also used to the interpretation of the results. Comparative Binding Energy (COMBINA) requires the structure of a ligand-protein complex and binding data for the training set is used for model development not for evaluation of the model of ligand-protein complexes. The training set must be performed an independent set of data to model all remaining complexes of the training sets are superimposed and energy is minimized. The physical properties and location in the molecule are computed and decomposed based on interaction energy between ligand and receptor in each complex. PLS analysis is used to build a model correlating the biological binding data with sum of weighted and selected components of the interaction energy.

IV. PROTEIN CHARACTERIZATION FROM PARTIAL GENE SEQUENCE

Tropical diseases are responsible for worldwide health problem is affected by millions of people annually. The bifunctional enzyme DHFR-TS from parasites as Toxoplasma gondii, Cryptosporidium parvum and four species of Plasmodium plays a crucial role in pyrimidine biosynthesis has been used as target for structure-based drug design. Plasmodium falciparum DHFR structural data to develop novel inhibitors with higher potency and selectivity data would be useful for molecular modeling DHFRs from species of parasites. Crystallographic analyses of the Leishmania major in binary and ternary complexes identified the interactions can be exploited in a structure-based development of novel enzyme inhibitors with potential therapeutic value. A structural comparison with the human enzyme identified a potentially important difference in the substrate-binding cleft assist to the development of TbGalE specific inhibitors. The xenobiotic electrophiles have identified antischistosomal antimalarial and antifilarial activity of compounds known for their GST inhibiting activity.

4.1 BLASTN interface in Nucleotide Sequence

The bioinformatics tools used to computationally characterize a complete protein starting with partial DNA sequence of its corresponding gene. The GenBank searching involves the whole gene sequence and translating it into protein sequence and searching for protein structure.

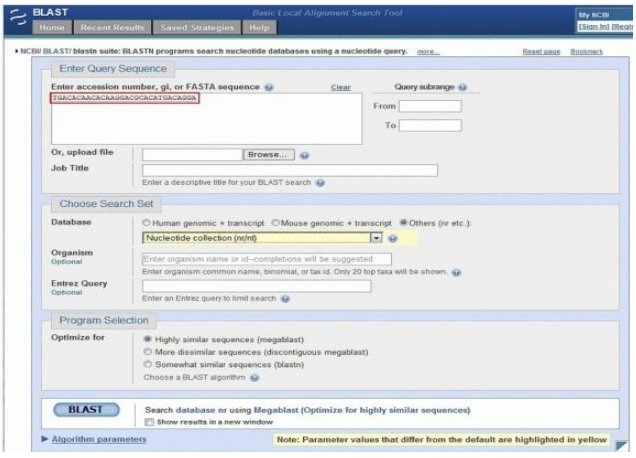


Figure 1: The Nucleotide Database using Nucleotide Query

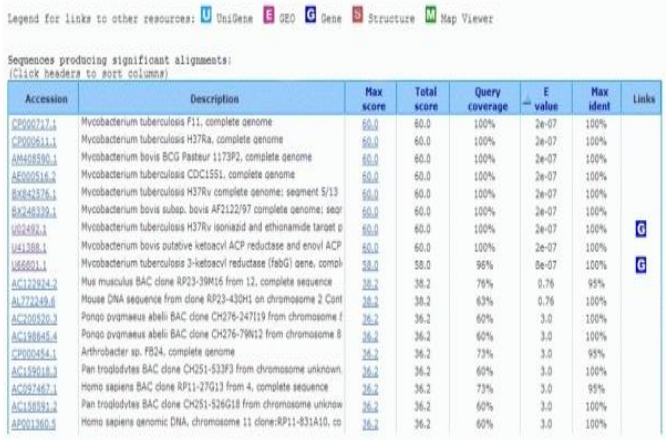
When no structure is available, then student should search for similar sequences with 30% or greater identity of known 3D structures to get familiar with protein architecture.

The gene can be completely and unambiguously characterized but it will require a more careful analysis have a nucleotide sequence as our initial data set to the NCBI BLAST page select the nucleotide blast program and search the nucleotide database using a nucleotide query.

This BLASTN interface is is divided into four major sections.

1. The first allows the user to paste the query sequence for analysis containing 30 nucleotides is highlighted in the red rectangle was pasted into the search box.
2. The second section permits the database to be searched and optional sequence range coordinates to the search is non-redundant (nr) nucleotide (nt) database.
3. The third section offers optimization alternatives to the search is use standard parameters and settings for our search will highly similar sequences using megablast.
4. This section shows a summary of our search features to show the results in a new window it allows us to re-run BLAST searches with different sequences parameters maintaining in the former results.

To perform the search, just click on BLAST icon after a few seconds the results page will display a new window.


Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
CP000717.1	Mycobacterium tuberculosis F11, complete genome	60.0	60.0	100%	2e-07	100%	
CP000611.1	Mycobacterium tuberculosis H37Ra, complete genome	60.0	60.0	100%	2e-07	100%	
AM408590.1	Mycobacterium bovis BCG Pasteur 1173P2, complete genome	60.0	60.0	100%	2e-07	100%	
AE000516.2	Mycobacterium tuberculosis CDC1551, complete genome	60.0	60.0	100%	2e-07	100%	
BX842576.1	Mycobacterium tuberculosis H37Rv complete genome: segment 5/13	60.0	60.0	100%	2e-07	100%	
BX248339.1	Mycobacterium bovis subsp. bovis AF2122/97 complete genome: segr	60.0	60.0	100%	2e-07	100%	
U02492.1	Mycobacterium tuberculosis H37Rv isoniazid and ethionamide target p	60.0	60.0	100%	2e-07	100%	G
U41388.1	Mycobacterium bovis putative ketoacyl ACP reductase and enoyl ACP	60.0	60.0	100%	2e-07	100%	
U66801.1	Mycobacterium tuberculosis 3-ketoacyl reductase (fabG) gene, compl	58.0	58.0	96%	8e-07	100%	G
AC122924.2	Mus musculus BAC clone RP23-39M16 from 12, complete sequence	38.2	38.2	76%	0.76	95%	
AL772249.8	Mouse DNA sequence from clone RP23-430H1 on chromosome 2 Cont	38.2	38.2	63%	0.76	100%	
AC200520.3	Pongo pygmaeus abelii BAC clone CH276-247I19 from chromosome 5	36.2	36.2	60%	3.0	100%	
AC198645.4	Pongo pygmaeus abelii BAC clone CH276-79N12 from chromosome 8	36.2	36.2	60%	3.0	100%	
CP000454.1	Arthrobacter sp. FB24, complete genome	36.2	36.2	73%	3.0	95%	
AC159018.3	Pan troglodytes BAC clone CH251-533F3 from chromosome unknown	36.2	36.2	60%	3.0	100%	
AC097467.1	Homo sapiens BAC clone RP11-27G13 from 4, complete sequence	36.2	36.2	73%	3.0	95%	
AC158591.2	Pan troglodytes BAC clone CH251-526G18 from chromosome unknow	36.2	36.2	60%	3.0	100%	
AP001360.5	Homo sapiens genomic DNA, chromosome 11 clone:RP11-831A10, co	36.2	36.2	60%	3.0	100%	

Figure 2: BLAST output lists of hits and alignments

The first 11 hits are very low E-values (below 10^{-5}) which represent the most possibly significant and demanding 100% of coverage in the sequence alignment. Then the first eight hits are relevant for search and the characterization of sequence. The first six hits are related to genome sequences which makes it difficult us to identify the target sequence to find DNA sequences containing the putative gene. The **partial** DNA sequence is seeking to identify most likely part of a gene belonging to either Mycobacterium tuberculosis or to Mycobacterium bovis, or both.

4.2 Use of Translate Tool

The tool Translate can be translates a DNA (RNA) sequence into its six different possible open reading frames (ORFs) and accessed through the ExPASy web site to translate our inhA gene sequence into its corresponding protein. Protein sequence is the GenBank flatfile including the number of amino acids. BLASTp parameters for protein-protein comparison are being used the substitution matrix BLOSUM62 and nr database. The bit score and the E values indicate the protein and red squares on the right with the letter "S" inside indicate our proteins have 3D structure. To obtain the protein access annotation through its GenBank accession number NP_21600 is similarly as the gene accession number. The PDB file to a local directory in our computer and work with our preferred molecular modeling and visualization package. The use of molecular visualization software of the object to be visualized the principles of protein structure and function. The InhA structure appears static will initially display the image rolling in three-dimensionality. Pressing the left mouse button and moving it left and right and up and down allows the user to keep full control of the rotation of the molecule. Pressing the left mouse button at the same time as the control or alt keys and moving the mouse will change the zoom. The principle states of protein sequence of target protein have sequence identity greater than or equal to 30% with another protein with known structure to its whole length and target protein 3D structure can be modeled based on one or on a combination of several template molecules.

V. CONCLUSION

The full three-dimensional structure of a protein is generally to understand its biological activity. It depends on the atomic detail and often subtle modifications to a structure can dramatically affect its activity. Furthermore, collaboration with synthetic chemists and experimental biologists will probably continue to be a critical determinant of success. Protein-ligand structure analysis can be enough to suggest modifications in the ligand to improve affinity. Protein families should become evident that the knowledge of a high resolution atomic structure of a target protein is an essential pre-requisite for structure based drug design. Virtual screening can provide ligand structures completely different from the already known ligands. QSAR models can aid the design of modifications to improve a biological property of a ligand and can predict activities with a reasonable error. Docking procedures can be used to model

binding modes and estimate binding energies. To overcome the ranking problem researchers have different methods ranging from brute force, which means using highly expensive computer clusters to do all the heavy computational work to using consensus scoring. Some docking tools also have another useful characteristic the estimation of binding energy. The best ligands found on virtual screening are submitted to this procedure and these calculations are not very accurate if it can provide the correct rankings of candidate molecules it will provide valuable information. The recent study used to structures and gene sequences of the enzyme glyoxalase I to create a model of the Leishmaniadonovani enzyme. Docking algorithm indicates the key electrostatic binding interactions between the inhibitor and the enzyme for the design of the next generation of inhibitors.

REFERENCES

[1] H.A. Scheraga, S. Rackovsky "Homolog detection using global sequence properties suggests an alternate view of structural encoding in protein sequences" ProcNatlAcadSci U S A, 111, pp. 5225–5229, 2014.

[2] C.J. Oldfield, J. Meng, J.Y. Yang, M.Q. Yang, V.N. Uversky, AK.Dunker "Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners" BMC Genomics, 9 2009.

[3] K. Motomura, M. Nakamura, J.M. Otaki "A frequency-based linguistic approach to protein decoding and design: simple concepts, diverse applications, and the SCS package" CSBJ, 5, pp. 1–9, 2013.

[4] Kouranov A. , Xie L. , de la Cruz J. , Chen L. , Westbrook J , Bourne P. E. , Berman H. M. " The RCSB PDB information portal for structural genomics", Nucl. Acids Res. 2006;34:D302–D305. [PMC free article] [PubMed]

[5] Z.P. Feng, C.-T.Zhang "A graphic representation of protein sequence and predicting the subcellular locations of prokaryotic proteins" Int J Biochem Cell Biol, 34, pp. 298–307, 2002.

[6] K.A. Dill, S.B. Ozkan, T.R. Weikl, J.D. Chodera, V.A. Voelz "The protein folding problem: when will it be solved" Curr Opin Struct Biol, 17 pp. 342–346, 2007.

[7] D.S. Marks, T.A. Hopf, C. Sander " Protein structure prediction from sequence variation" Nat Biotechnol, 30, pp. 1072–1080, 2011.

[8] Verdonk M. L. Cole J. C. Hartshorn M. J. Murray C. W, Taylor R.D "Improved Protein–Ligand Docking Using GOLD. PROTEINS: Structure, Function, and Genetics" 2003; 52:609–623. [PubMed]

[9] Wade, C. R., Henrich, S., Wang, t "Using 3D protein structures to derive 3D-QSARs" Drug Discovery Today, Vol.1, No 3. 2004. [PubMed]

[10] Padmanabhan P.K. et al. "Glyoxalase I from Leishmaniadonovani: A potential target for anti-parasite drug" Biochemical and Biophysical Research Communications. 2005;337:1237–1248. [PubMed]

[11] Edwards Y. J., Cottage A. "Bioinformatics methods to predict protein structure and function, A practical approach" Mol.Biotechnol. 2003;23:139–166. [PubMed]

Mrs. B. Kalaiselvi has completed Master of Engineering in Computer Science and Engineering (2009) in Anna University, Coimbatore. Her research expertise covers medical data mining, machine learning, cloud computing, big data, fuzzy, soft computing and ontology. She has published journal articles in the above fields and she has presented 15 papers in national and international conference in the above field. . She has got junior educator and researcher award from National foundation for entrepreneurship development award. She has handled a session in National level workshop at Kongu Engineering College, Perundurai. She has published 3 Book chapters in Jupiter publications consortium, R.K Publications and Mercury Publications Consortium. She is currently working as assistant professor in Mahendra Engineering College for women.

Dr. M. Thangamani possesses nearly 23 years of experience in research, teaching, consulting and practical application development to solve real-world business problems using analytics. Her research expertise covers Medical data mining, machine learning, cloud computing, big data, fuzzy, soft computing, ontology development, web services and open source software. She has published nearly 80 articles in refereed, indexed, SCI Journals, books and book chapters and presented over 67 papers in national and international conferences in above field. She has delivered more than 60 Guest Lectures in reputed engineering colleges and reputed industries on various topics. She has got best paper awards from various education related social activities in India and Abroad. She has received the many National and International Awards. She is also Board member in Taylor & Francis Group and seasonal reviewer in IEEE Transaction on Fuzzy System, international journal of advances in Fuzzy System and Applied mathematics and information journals. She is currently working as Assistant Professor at Kongu Engineering College at Perundurai, Erode District.