"CYFRA-21-1 protein evaluation through *in-silico* methods"

To be submitted as Major Project Report in partial fulfilment of the requirement for the degree of

Master of Technology

Bioinformatics

Submitted by

RITESH KUMAR

(Roll No. 2k14/BIO/14)

Delhi Technological University, Delhi, India

Under the supervision of

Prof. Bansi D. Malhotra

Department of Biotechnology,
Delhi Technological University
Shahbad Daulatpur, Main Bawana Road, Delhi-42, India.

DECLARATION

I certify that the project report entitled **"CYFRA-21-1 protein evaluation through *in-silico* method"** submitted by me is in partial fulfilment of the requirement for the award of the degree of Master of Technology in Bioinformatics, Department of Biotechnology, Delhi Technological University. It is a record of original research work carried out by me under the supervision of **Prof. Bansi D. Malhotra** Department of Biotechnology, Delhi Technological University, Delhi-42.

The matter embodied in this project report is original and has not been submitted for the award of any Degree/Diploma.

Date

Ritesh Kumar
2k14/BIO/14
Department of Biotechnology
Delhi Technological University
Delhi-110042

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my supervisor and mentor, ***Prof. Bansi. D. Malhotra*** *who has guided me through thick and thin. Without his valuable guidance, incessant help, calm endurance, constructive criticism and constant encouragement the completion of this minor project would have been impossible, the congenial discussion with him made me confident and gave enough independence in performing the experiment which enabled in me a higher efficiency to complete the work.*

I express my kind regard and gratitude of ***Prof. D. Kumar*** *(HOD, Department of Biotechnology), and all faculty members for helping in my project.*

I am highly thankful to all my teachers who motivate me and give their support in my study.

I highly indebted to Mr. Suveen Kumar, Mr. Saurabh Kumar, Ms. Shine Augustine and Mr. Neeraj Kumar (Research Scholar) for their guidance and constant supervision as well as for providing necessary information regarding the instruments and experiments and also for their support in completing the reports.

I would also like to thanks Mr. Chhail Bihari, Mr. Jitendre Singh, Mr. Mukesh Kumar and all the staff members of department of biotechnology for providing necessary things & maintained laboratory in good condition.

At last but never the least, words are small trophies to express my deep sense of gratitude and affection to my loving friend and my parents who gave me infinite love to go for this achievement.

Ritesh Kumar
2K14/BIO/14

CERTIFICATE

This is to certify that the dissertation entitled **"CYFRA-21-1 protein evaluation through *in-silico* method" (DTU/14/M.TECH/123)** in the partial fulfilment of the requirements for the reward of the degree of Master of Technology, Delhi Technological University is an authentic record of the candidate's own work carried out by him/her under my guidance. The information and data enclosed in this thesis is original and has not been submitted elsewhere for honouring of any other degree.

Date:

Prof. D. Kumar
Head
Department of Biotechnology
Delhi Technological University
Delhi-110042

Prof. Bansi D. Malhotra
(Project Mentor)
Department of Biotechnology
Delhi Technological University
Delhi- 110042

CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

KRT-19/K-19	-	Keratin 19
NCBI	-	National Center for Biotechnology Information
I-TASSER	-	Iterative Threading ASSEmbly Refinement
ExPASy	-	Expert Protein Analysis System
SAVES-SERVER	-	Structure Analysis Verification and Server
ProSA-Web	-	Protein structure Analysis
PDB	-	Protein Data Bank
MWCNT	-	Multiwall carbon nanotubes
XRD	-	X-Ray diffraction
CTAB	-	Cetyl trimethyl ammonium bromide
APTES	-	3-aminopropyl triethoxysilane
ITO	-	Indium tin oxide
EPD	-	Electrophoretic deposition
BSA	-	Bovine serum albumin

LIST OF TABLES

Tables **Page No.**

"CYFRA-21-1 protein evaluation through *in-silico* method"

RITESH KUMAR
2K14/BIO/14
Delhi Technological University, Delhi, India
E-mail ID: - er.riteshkumar007@gmail.com

ABSTRACT

We report results of the studies relating to the structure prediction and function characterization of keratin, type I cytoskeleton-19 proteins, we highlights our study by downloading sequence of Keratin 19 from NCBI (National Centre of Biotechnology information) and PDB or PMDB. Keratin 19 is a member of the keratin family; that is intermediate filaments protein responsible for the structural integrity of epithelial cells and are subdivided into cytokeratin and hair keratins, Keratin 19 is also known as CYFRA -21-1. The Accession No. **NP_002267** for physico-chemical properties that is using to predict the 3D structure by I-TASSER and also computed the Expasy's ProtParam server. In order to select the best suitable templates, the FASTA sequence was subjected to protein BLAST against PDB database. The evaluation of Model and homology modelling were done by SWISS modelling i.e., an online tool. PROCHECK and ProSA analyses would be calculated to validate the 3D orientation of model, stereo-chemical properties and for minimization and validation that is done by SWISS PDB viewer tool (Version 4.0) and detection of salivary cyfra-21-1 protein via biosensors by synthesis and functionalize process of multi walled carbon nano-tubes (MWCNT)/ZrO2 nano-composite with APTES to predict the results.

INTRODUCTION

Oral cancer is one of the highly prevalent cancers known till date, and it occurs more often in men than women. This cancer occurs as a sore in mouth that doesn't easily heal and can be life threatening if it is not detected and treated early. It occurs on the floor of the mouth, tongue, cheeks, lips etc. There are some common causes also behind it like smoking, chewing tobacco, alcohol consumption, etc [(1) A.Gorschinski.et.al, *J. Mater. Chem.*, 2009,19,8829 (2) Winn.et.al, *J. Med.*, 1981, 304, 745]. Oral cancer can be detected and monitored easily by using some techniques like biopsy, cytopathology, visualization adjuncts etc. [Patton et. al., *J. Am. Dental* Assoc, 2008, 139, 896]. Thus this biomarker, for the detection of oral cancer in human saliva gives a non-invasive and pain-free alternative. Some biomarkers such as (Interleukin-8) IL-8, (Interleukin-6) IL-6, vascular endothelial growth factor (VEGF), epidermal growth factor receptor (EGFR) and CYFRA-21-1, have been recently proposed for the detection of oral cancer [Z.Wang. et.al,. *Biosense, Bioelectron,* 2013, 45, 108]. The detection of oral cancer having some non-invasive medium like saliva, urine, sweat & tear. We work on CYFRA-21-1 as a biomarker for the detection of oral cancer because CYFRA-21-1 also known as keratin-19 is a protenacious biomarker that represents the fragment of 40kDa of keratin-19 [R.Nagler et. al., *Clin, Cancer Res.*2006, 12, 3979]. Keratin-19 is also known as "CYFRA-21-1" or "Keratin, type 1 cytoskeletal 19" or cytokeratin-19 is encoded by K-19 gene in human [Schweizer et.al., *the journal of cell biology,* 2006,174,169-74]. KRT-19 is a member of keratin family. K-19s are intermediate filament proteins which are responsible for structural integrity of epithelial cells [Entrez Gene: KRT19 keratin 19, *NCBI*]. In normal human, the cut-off concentration of CYFRA-21-1 in saliva is 3.8 ng/mL, while in oral cancer patients have found in between 17.46 ± 1.46 ng/mL [Rajkumar. et. al., *Oral Dis.* 2013, 27, 90]. The concentration of CYFRA-21-1 in

saliva is very high as compared to lower level of biomarkers [Bandodkar. et. al., *Trends biotech.* 2014,32,363]. During this study, we used nanostructured zirconia ($nZrO_2$) has been found to own attention-grabbing characteristics for biosensing applications. These characteristics embody biocompatibility, excellent electrical and surface charge properties which will be beneficial for integration of the immobilized biomolecules [M. Das.et.al., *Appl. Phys. Lett*, 2011, 99, 143702].

In a bioinformatics way, the structure prediction & characterization of CYFRA-21-1 by using various bioinformatics tools may help in determining its 3D structure, protein interactions & immuno-complex formations. Because CYFRA-21-1 doesn't have 3-dimensional structure till now. So structural genomics helps within the determination of the 3D structures of all proteins of a given organism, by using some experimental strategies like various techniques like X-ray Crystallography, NMR spectroscopy or computational approaches like similarity modeling are currently being used for protein structure predication and characterisation. This produces new types of challenges in structural bioinformatics, i.e. determinative protein functions from its 3D structure [Baker.et.al, *sciences,* 2001, 294, 93-96]. Where an approach like Homology modeling is employ in order to produce structures for these proteins [Roy, Sudeep.et.al,. *Bioinformation,* 2011, 6(8): 315-319]. Generally, In order to select the best suitable templates, the FASTA sequence was subjected to protein BLAST against PDB databases. I-TASSER or SWISS modelling used for the evaluation of Model and homology modelling like an online tool. To validate the 3D oriented of model widely analysed by using PROCHECK and ProSA, and some others properties like stereo-chemical properties for minimization and validation which is also done by SWISS PDB viewer tool (Version 4.0).

REVIEW OF LITERATURE

The word *Keratin, type 1 cytoskeleton 19* also known as "cytokeratin-19", it is a 40kDa protein which is encoded by the KRT 19 gene. It is a member of keratin family which is also known as "*CYFRA-21-1*". It is a water-soluble proteinaceous biomarker and it is present in human saliva which is used to detect Oral Cancer.

According to [suveen et al., 2015] carcinoma is one in all the extremely lethal type of cancer that is usually happens men than women. However, with passage of your time, some symptoms like mouth ulceration, loosening of teeth, and gruff voice square measure renowned to develop. CYFRA-21-1 may be a soluble proteinaceous biomarker that is gift in human spit, employed in the detection of carcinoma. Throughout this content, nanostructure zirconium oxide ($nZrO_2$) has been found to possess attention-grabbing characteristics for biosensing applications. At last, The Conclusion of connected study has made-up a straightforward, efficient, label-free, and non-invasive nanostructure oxide based mostly biosensing platform for carcinoma detection.

ZrO_2 nanoparticles are synthesized via the hydrothermal methodology and its silanization has been achieved mistreatment APTES. This film of APTES/ZrO_2/ITO is made-up via electrophoratic deposition and followed by valence immobilization of antibodies. The comparison with alternative rumoured carcinoma detection ways together with biosensors, the BSA/anti-CYFRA-21-1/APTES/ ZrO_2/ITO biosensor is straightforward, exhibits a wider detection varies of 2-16 ng/mL.

In a bioinformatics way the study gives [sudeep et al., 2011] an analogous study related to structure prediction and purposeful characterization of secondary proteins (*Ocimum species*) explains the bioinformatics technique to expect the 3D structure. In step with him,

Experimental structures like X-ray and nucleon resonance of molecule do not appear to be even so gettable inside the macromolecule Databank (PDB). These proteins play an awfully vital role in varied metabolic pathways in protein.

The 3D structures of the proteins square measure essential to check most of their functions. Similarity modelling approach was used therefore on derive structures of these proteins. A program meant for comparative modelling-Modeller 9v7 was used for the aim. The stacked proteins were any valid by Procheck and Verify-3d and Errat servers. Compound composition and polarity of these proteins explained by CLC-Protein work bench tool. Expasy's Prot-param server and Cys_rec tool were used for physico-chemical and purposeful characterization of these proteins. Studies of secondary structure of these proteins were administered by procedure program, Profunc. Swiss-pdb viewer was accustomed visualize and analyze these similarity derived structures. The structures square measure finally submitted in macromolecule Model information (PMDB) so as that they become accessible to different users for any studies.

In another study related to structure prediction and functional characterization of proteins also explains by [Ravi et. al., 2013], According to him, He explained about characterization of some hypothetical proteins like Rice (*Oryza Sativa L.)* by using some softwares. Because Structural genomics consists within the determination of the 3 dimensional structures of all protein given organism, by experimental strategies like X-ray crystallography, Nuclear magnetic resonance spectroscopy or computational approaches like similarity modeling. This raises new challenges in structural bioinformatics, i.e. crucial protein operate from its 3D structure. This genomic sequencing project gives linear AA sequences.

In general, the author retrieves the sequences of 10 hypothetical protein sequences of rice from NCBI with some unique IDs and produce physicochemical characterizations like

theoretical Isoelectric point (pI), molecular weight, instability index, aliphatic index and grand average hydropathy (GRAVY) by using Expasy's Protparam. After that this protein sequence are used to find out the Conserve Domain Database (CDD) and this software are used to scan a set of pre-calculated position-specific scoring matrices with a protein query. In same way, these proteins need to predict the Transmembrane Proteins which was used to characterize that the protein is soluble or transmembrane in name by using SOSUIserverbp.nuap.nagoyau.ac.jp/sosui_submit.html. At last, The result of the above study shows that the physiochemical parameters were computed by sing the Expasy's Portparam tool and this tool are used to find out whether it's acidic nature or not.

And the conclusion of above study explains that the structure prediction and enzyme functions finds 10 hypothetical proteins of rice and this is important to annotate and find the structural and functional properties of theses hypothetical proteins.

In 2011, [Subazini K.T & Kumar Ramesh] gives another Lovastatin biosynthetic cluster proteins characterization related to strain *A. terreus*, the author retrieve the sequences which is involved in lovastatin biosynthesis in ATC 20542 stains from NCBI (www.ncbi.nlm.nih.gov) database. And this primary sequence analysis by using sequence manipulation suite, like physicochemical properties like grand average of hydropathy (GRAVY), pH, acidicity, basicity, extinction coefficient, aliphatic index, molecular weight etc and the prediction of secondary structure are predicted by GOR tools for helical content, beta sheet formation and turns, loops, and coil regions. The author also explained the functional annotation by using KOGnitor, which is used for incorporate orthologus information. ScanProsite for identifying biological significant, ProDom for domain based analysis and BLAST for homology. The tertiary structure also predict for the 3D structure of the given proteins by using MODELLER 9V8, where as the template for given proteins is produces by PDB, Swiss-PDB and finally refined models were subjected for validation by

Ramachandran Plot (SAVES server) and the Z-score which is obtained by Prosa web server. The conclusion of above study gives the information that the *insilico* analysis given protein can be done by physiochemical, structural and functional properties.

Another computational and functional analysis of hypothetical proteins of *S.aureus* is also explained by [mohan et. al., 2012]. The same procedure author is followed by author, to retrieve sequences of 10 hypothetical proteins of *S. Aureus* from NCBI database and produces physicochemical and functional characterization like GRAVY, aliphatic index, molecular weight, Isoelectric point, extinction coefficient etc were computed by using Expasy's Protparam server. The author is also used some tools like PFAM which is basically used for multi-protein-sequences alignment, hidden markov models, and a set of manually curated and annotated models. CDD-BLAST is used for quickly scan a set of pre-calculated position-specific scoring matrics. The authors also used STRING as a tool for the protein-protein interaction, SOSUI server for the prediction of transmembrane proteins, DISUFIND server used for stability of protein, PS Square (PS^2) for the protein structure prediction and Q-Site Finder used for ligand binding site prediction and this tool works by binding hydrophobic probes to the proteins. The conclusion shows that the structural and functional properties of hypothetical proteins in the *S. Aureus* produce many virulence factors and cause serious infections and disease.

At last, [Songfeng Wu et. al., 2011] gives their views about the ProPAS tool which is a standalone software used to analyse the properties of proteins. According to him, the identification of Proteins is a basic work for researchers, because physiochemical properties usually need to separate the proteins by their functions. ExPASy is most curated available software for calculation of properties. Some properties like Isoelectric point, hydrophobicity, molecular weight etc.

METHODOLOGY

There are various types of databases and online tools are used for structure prediction and function characterization of protein. These are as follow:-

1. **National Center for Biotechnology Information :-**

   Web page: www.ncbi.nlm.nih.gov

NCBI is a part of United State National Library of medication (USNLM), which is a branch of National Institute of Health (NIH). NCBI is directed by *David Lipman*; it is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by legislator Claude Pepper. Some major databases includes like GenBank for DNA sequences and PubMed, and others like NCBI Epigenomics database. All these databases are offered on-line through the Entrez computer program. The NCBI has software tools that are offered by computer network browsing or by FTP. Example, BLAST is a sequence similarity searching program. BLAST can do sequences comparisons against the GenBank DNA acid information in less than 15 seconds.

2. **Iterative Threading ASSEmbly Refinement (I-TASSER):-**

   Web page: zhanglaInb.ccmb.med.umich.edu/I-TASSER/

It stands for "Iterative Threading ASSEmbly Refinement" (I-TASSER) is a hierarchic methodology for protein molecule structure and performance prediction. It is a unified platform for automated protein structure and function prediction, it has as on-line server in-built the Yang Zhang science laboratory at the University of Michigan, AnnArbor. It is a bioinformatics method for predicting 3-D structure model of protein molecule molecules from organic compound sequences. It detects structure templates from the PDB by a technique called fold recognition or threading [Ambrish Roy.et.al., 2010]. I-TASSER server having 4 basic stages:-

a) **Threading**: - Threading refers to a bioinformatics procedure for identifying template protein molecules from resolved structure databases that have a similar structure or similar structural motif because of the protein sequence. In the first stage of I-TASSER, the query sequence is matched against a non-redundant sequence information by position-specific iterated BLAST (PSI-BLAST), to identify evolution- vary relatives. The quality of the template alignments (and so the issue of modeling the targets) is judged supported the applied math

significance of the simplest threading alignment, i.e., the Z-score, which is outlined as the energy score in variance units relative to the applied math mean of all alignments[Altschul, S.F. et al., *Nucleic Acids Res*. 25, 3389–3402 (1997)].

b) **Structural assembly:-** In the second stage, continuous fragments in threading alignments are excised from the template structures, and are used to assemble structural conformations of the sections that aligned well, with the unaligned regions (mainly loops/tails) built by ab initio modeling [Wu, S., Skolnick, J. & Zhang, Y. *BMC Biol*. 5, 17 (2007)]. To improve the efficiency of conformational search, I-TASSER adopts a reduced model to represent the protein chain, with each residue represented by its Cα atom and side-chain center of mass. Because the regions not aligned throughout the threading method sometimes have a lower modeling accuracy, the structure modeling in these regions is confined to a lattice system of grid size 0.87Å, which helps to scale back the entropy of conformational search [Zhang, Y., Kolinski, A. & Skolnick, J. touchstone II: *Biophys. J*. 85, 1145–1164 (2003)].

c) **Model selection and refinement: -** In the third stage, the fragment assembly simulation is performed again beginning from the chosen cluster centroids. Although the inherent I-TASSER potential remains unchanged in the second run, external constraints area unit pooled from the LOMETS threading alignments and the PDB structures that are structurally highest to the cluster centroids, as identified by TM-align [Zhang, Y. & Skolnick, J. TM-align., *Nucleic Acids Res*. 33, 2302–2309 (2005)]. The purpose of the second iteration is to get rid of steric clashes and to refine the worldwide topology of the cluster centroids.

d) **Structure based functional annotation: -** In the last stage, the function of the protein molecule is inferred by structurally matching the foreseen 3D models against the proteins of glorious structure and performance within the PDB. For this purpose, 3 protein molecule structure/function libraries have been created severally and biweekly updated; at this time, these include a library of 5798 non-redundant entries with known EU numbers [Barrett, A.J. *Eur. J. Biochem.* 250, 1–6 (1997)], a library of 26,045 non- redundant entries with known GO terms [Ashburner, M. et al. *Nat. Genet.* 25, 25–29 (2000).] and a library of 19,658 non-redundant entries with known ligand-binding sites.

3. Expert Protein Analysis system (ExPASy):-

Web Page: www.expasy.org

The ExPASy (the Expert protein Analysis System) provides the service and info concerning the bioscience community by a multidisciplinary team at the SIB (Swiss Institute of Bioinformatics. Some ExPASy databases like SWISS-PROT and TrEMBL, SWISS-2D PAGE, PROSITE, ENZYME and the SWISS-MODEL includes repository. It provides access to a variety of databases and analytical tools that dedicates the protein molecules and genetics. These tools are offered for specific tasks relevant to genetics, post-translational modification prediction, pattern and profile searches, similarity searches, topological prediction, sequence alignments, primary, secondary and tertiary structure. These tools or databases are interlinked every different [Elisabeth Gasteiger.et.al., 2003].

Some Important tools are used in Expert Protein Analysis System (ExPASy):

- **Swiss Model:**

  It is a web-server dedicated structural bioinformatics tools used for homology modelling of protein 3-D structure. Homology modeling is presently the most correct technique to come up with reliable 3-dimensional macromolecule structure models and is habitually utilized in several sensible applications. Homology (or comparative) modeling technique makes use of experimental macromolecule structures (template) to build models for biological process connected proteins (targets). Today, SWISS-MODEL consists of three tightly integrated parts:

- ✓ The SWISS-MODEL pipeline- a suite of software tools and databases for machine-driven protein structure modeling.
- ✓ The SWISS-MODEL Workspace – a web- based mostly graphical user work bench.
- ✓ The SWISS-MODEL Repository- a continuously updated info of similarity models for a set of model organism proteomes if high medical specialty interest.

- **ProtParam:**

  Web page: - web.expasy.org/protparam/

  ProtParam is a tool, used for the computation of many physical and chemical parameters for a protein that hold on in Swiss-Prot or TrEMBL and for a user entered protein sequences. In other words, it computes various physic-chemical properties,

can be deduced from a macromolecule sequence. The protein will nominative as Swiss-Prot/TrEMBL accession range or ID, or in the type of a raw sequences. Some calculated parameters like; molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathcity (GRAVY).

4. **SAVES Server:**

Web page: services.mbi.ucla.edu/SAVES/

It is stands for "*Structure Analysis Verification and Server*", is a metaserver which helps to run 6 programs for checking and validating protein structures during and after model refinement which are as :-

- PROCHECK: This tool used for check the stereochemical quality of protein structure by analyzing residue by residue geometry and overall structure geometry.
- WHAT_CHECK: derived from a subset of protein verification tools from WHATIF program, this does extensive checking of many sterochemical parameters of the residues in the model.
- ERRAT: analyses the statistic of non-bonded interactions between different atom types and plots the value of the error function versus position of a 9-residue sliding window.
- VERIFY_3D: determines the compatibility of an atomic model (3D) with its own amino acids sequence (1D) by assigning a structural class based on its location and environment (alpha, beta, loop, polar, non polar etc.) and comparing the results to good structures.
- RAMACHANDRAN PLOT: Produce an interactive Ramchandran plot.

5. **ProSA-Web:**

Web page: prosa.services.came.sbg.ac.at/prosa.php

ProSA-Web basically stands for "Protein structure Analysis" that provides associate straightforward to use interface to the program ProSA, which is often utilized in protein structure validation. This tool use to calculate an overall quality score for a specific input structure. If the resultant score is outside a range characteristic for any pure protein, means the structure in all probability contains error. A plot of local quality scores points to problematic elements of the model that area unit conjointly

highlighted during a 3D molecule viewer to facilitate their detection. It uses only the C-alpha atoms of input structure, hence it will conjointly apply to low resolution structures and some models obtained early within the structure determination method. In the output, Z-score indicates overall model quality. Its value is displayed in a plot that contains z-score of all determined proteins chains in current PDB. It also visualizes the 3D structure of input protein mistreatment the molecule viewer Jmol.

6. String:-

Web page: string-db.org

STRING may be a bioinformatics method for finding out the patterns in DNA sequences, like principally individuals having similar factor sequences during which some regions of DNA sequences vary from one person to a different with high frequency. Once comparison the variation 2 completely different DNA samples, It permits the researchers answer the queries of weather DNA samples comes from an equivalent person of not. A DNA sequences additionally represents the order that contained among associate organism, wherever the order may be a set of sequences that explains regarding the proteins to make among the organism.

RESULTs and DISCUSSIONs

Minimization and Validation

1. **Search and retrieval of target protein**

   The sequence of human protein "keratin, type I cytoskeleton 19" (**Accession No. – NP_002267**) was taken from National Centre for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov) in FASTA format.

   Procedure:-

   - ✓ Firstly open the Browsers i.e., www.google.co.in
   - ✓ Now search the NCBI (National Center of Biotechnology Information) and open the home page.
   - ✓ In the home page, we select the database as "Protein" and put the name in the search box i.e., "CYFRA 21-1"
   - ✓ Another page shows the summery of total results related to given protein, we select assume protein like "**keratin, type I cytoskeletal 19 [Homo sapiens]**" with desired (400AA) amino acids.
   - ✓ Then we give the command to change the desired sequence in FASTA format and get the sequence as output like:-

   keratin, type I cytoskeletal 19 [Homo sapiens]

   NCBI Reference Sequence: NP_002267.2

   GenPept Identical Proteins Graphics

   ```
   >gi|24234699|ref|NP_002267.2| keratin, type I cytoskeletal 19 [Homo sapiens]
   MTSYSYRQSSATSSFGGLGGGSVRFGPGVAFRAPSIHGGSGGRGVSVSSARFVSSSSSGAYGGGYGGVLT
   ASDGLLAGNEKLTMQNLNDRLASYLDKVRALEAANGELEVKIRDWYQKQGPGPSRDYSHYYTTIQDLRDK
   ILGATIENSRIVLQIDNARLAADDFRTKFETEQALRMSVEADINGLRRVLDELTLARTDLEMQIEGLKEE
   LAYLKKNHEEEISTLRGQVGGQVSVEVDSAPGTDLAKILSDMRSQYEVMAEQNRKDAEAWFTSRTEELNR
   EVAGHTEQLQMSRSEVTDLRRTLQGLEIELQSQLSMKAALEDTLAETEARFGAQLAHIQALISGIEAQLG
   DVRADSERQNQEYQRLMDIKSRLEQEIATYRSLLEGQEDHYNNLSASKVL
   ```

2. **I-TASSER**

   It is stands for "Iterative Threading Assembly Refinement", is a server which is an integrated platform for automated protein structure and function prediction based on the sequences to structure and structure to function. Firstly take an amino acid sequences from NCBI portal, and I-TASSER helps to produce 3D atomic models from multiple threading alignments and Iterative structural assembly simulations. Some other functions are:-

   - ✓ It is a hierarchical method for protein structure and function prediction.
   - ✓ It is a bioinformatics method for predicting 3D structural model or protein molecules and Amino Acid sequences.

- ✓ This tool used to detect the structure templates from the PDB by a technique called fold recognition or threading.

Procedure:-

- ✓ Firstly open the Browser www.google.co.in
- ✓ *Put "zhanglaInb.ccmb.med.umich.edu/**I-TASSER**/" in the Google box.*
- ✓ *Now open another browser for NCBI home page to get the sequence of cytoskeleton 19.*
- ✓ In the home page of I-TASSER we make an account with valid email ID.
- ✓ After that we paste our desired sequence in I-TASSER and then submit it on its website.
- ✓ After 24-48 hours we can get the result in the form of 3-D structure in email account.

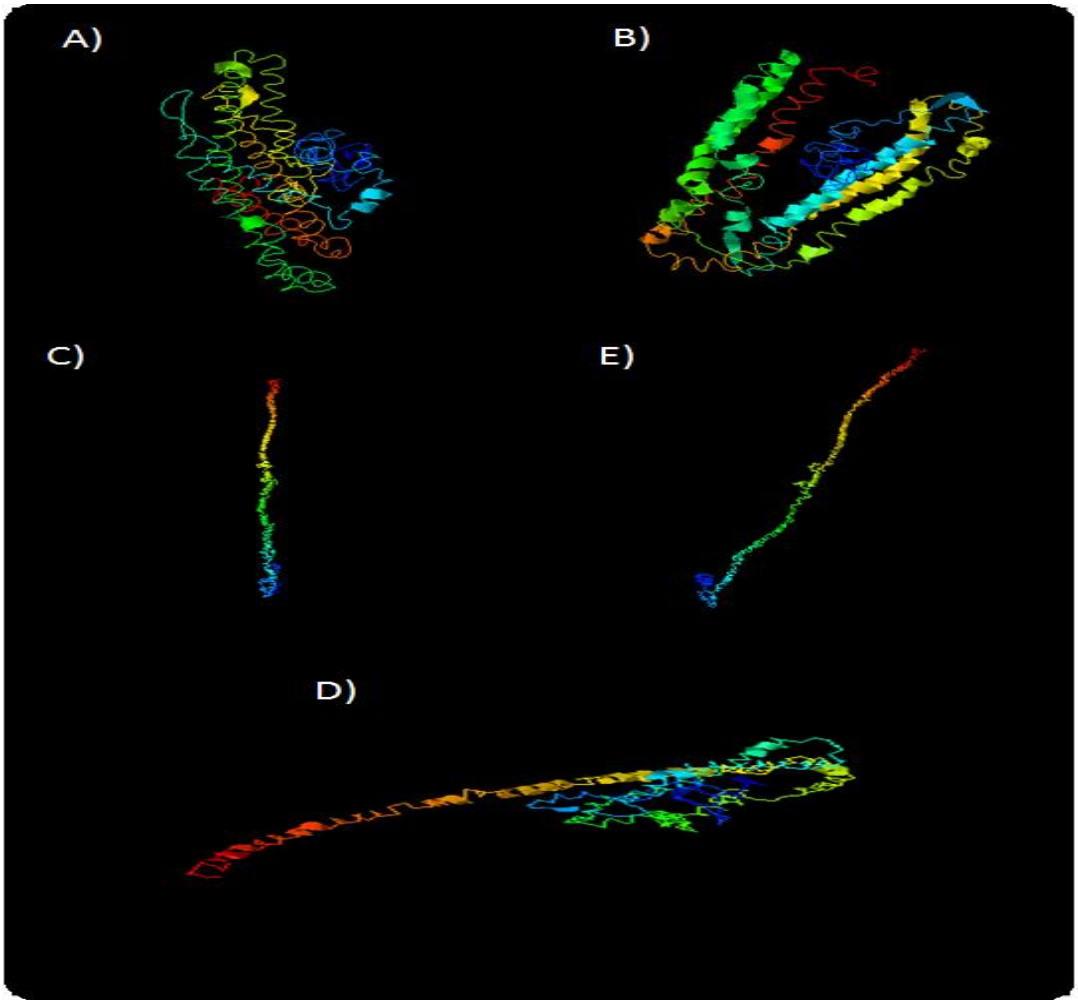


Fig 1: - 3D images of CYFRA-21-1 predicted by using I-TASSER

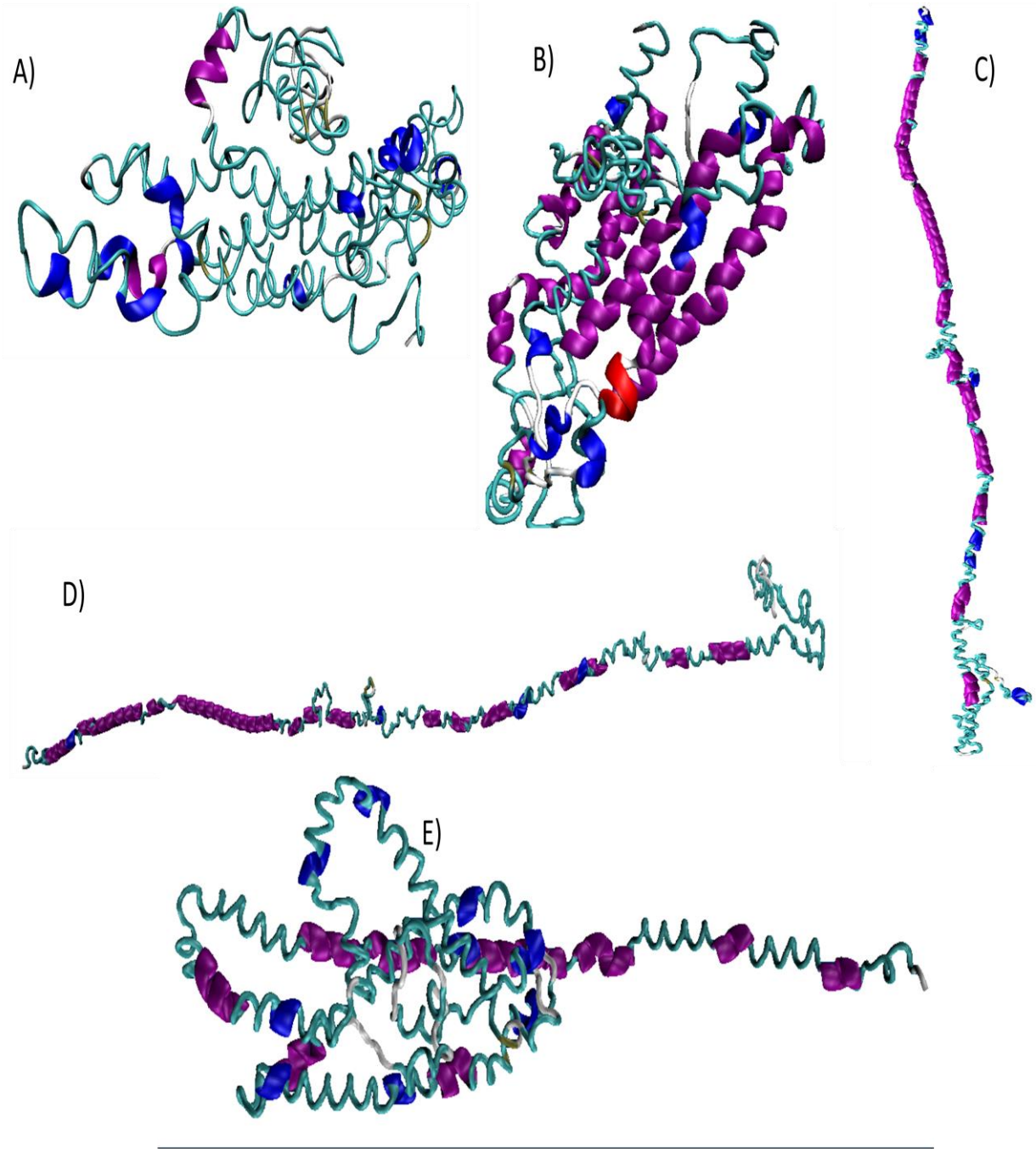
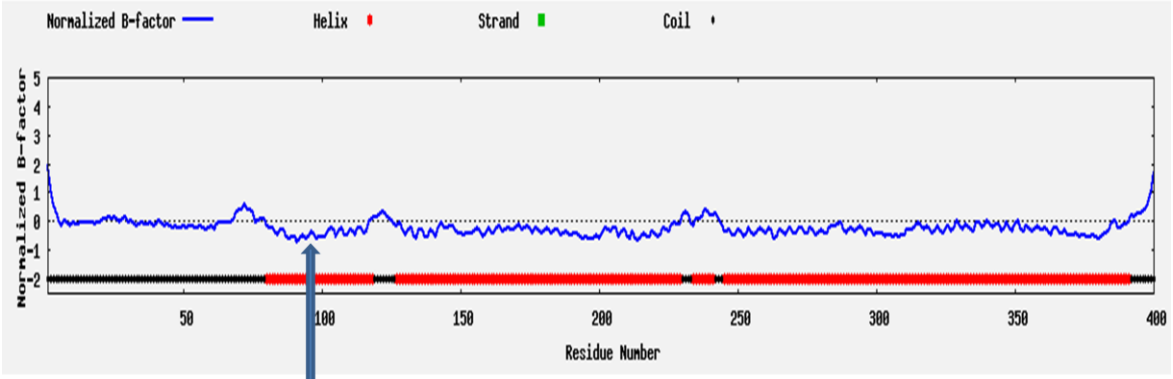


Fig 2: 3D structure designed by VMD software

#Predicted Normalized B-factor# B-factor may be a price to point the extent of the inherent thermal quality of residues/atoms in proteins. In I-TASSER, this price is deduced from threading example proteins from the PDB together with the sequence profiles derived from sequence databases. The rumoured B-factor profile within the figure below corresponds to the normalized B-factor of the target macromolecule, outlined by B=(B'-u)/s, wherever B' is that the raw B-factor price, u and s square measure severally the mean and variance of the raw B-factors on the sequence.



This portion shows the negative(-ve) values means the residue is relatively more stable in this structure.

NAME	C-SCORE	Exp. TM SCORE	Exp. RMSD(Å)	NO. OF DECOYS	CLUSTERS DENSITY
MODEL 1	-2.84	0.39±0.13	13.7±4.0	483	0.0302
MODEL 2	-2.83	-	-	426	0.0304
MODEL 3	-3.14	-	-	278	0.0224
MODEL 4	-4.22	-	-	120	0.0076
MODEL 5	-2.56	-	-	100	0.0402

Fig 3: - Table shows all scores value of CYFRA-21-1 by I-TASSER

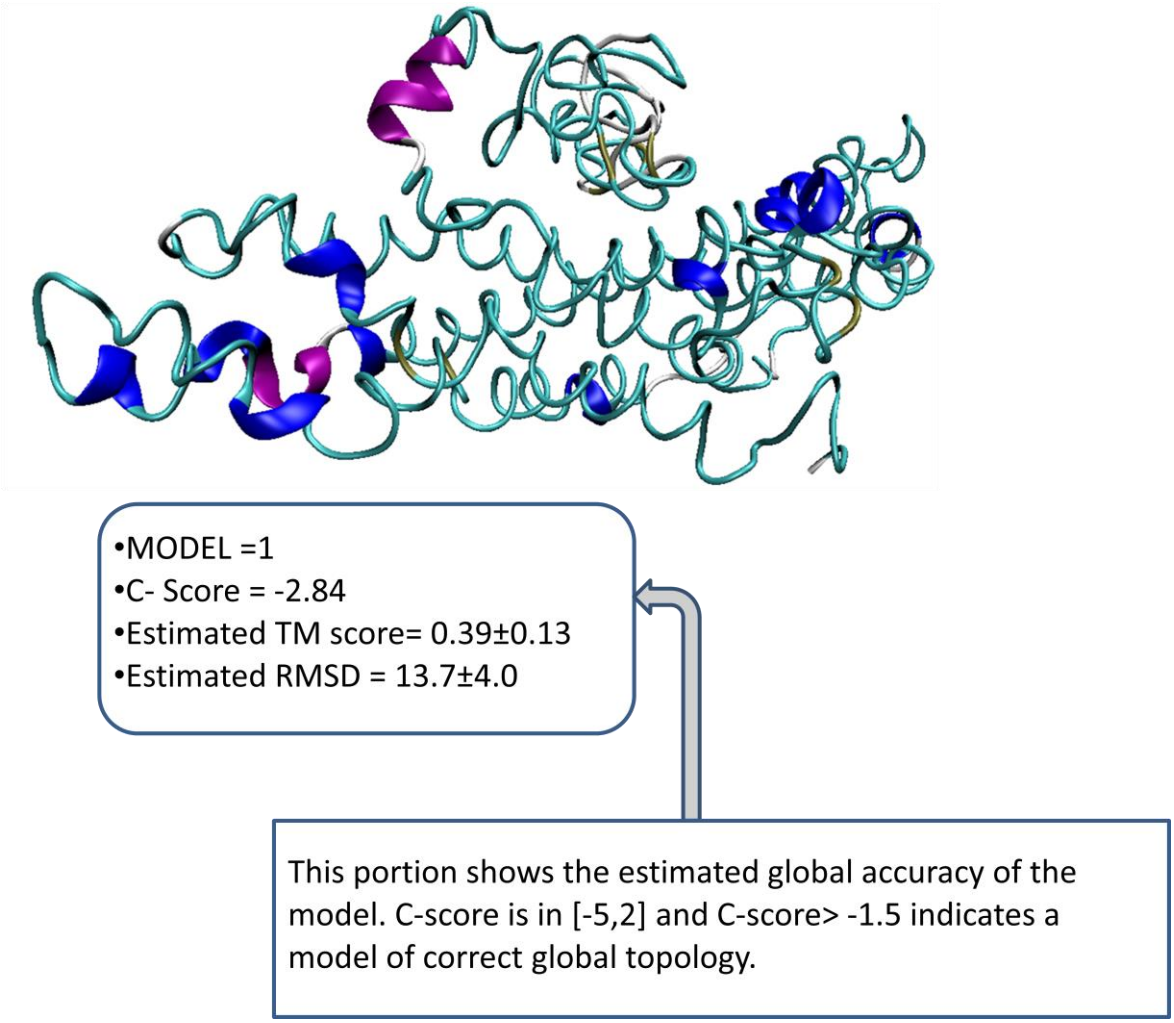


Fig 4: - An example of predicted 3D model with estimated global and local accuracy

#Protein structurally close to the target in the PDB#

After the structure assembly simulation, I-TASSER uses the TM-align structural alignment program to match the primary I-TASSER model to any or all structures within the PDB library. This section reports the highest 10 proteins from the PDB that have the nearest structural similarity, i.e. the very best TM-score, to the expected I-TASSER model. Owing to the structural similarity, these proteins typically have similar perform to the target. In general, After the structure assembly I-TASSER gives 10 best protein structure which is closely to the target in the form of PDB Hit with it's TM score, RMSD value, Identity, and Coverage value are clearly shown in following table:-

RANK	PDB Hit	TM-SCORE	RMSD	IDENTITY	COVERAGE
1	1byt	0.405	6.36	0.052	0.630
2	3bndA	0.404	6.31	0.056	0.627
3	2pomA2	0.401	6.21	0.044	0.610
4	3rdeA	0.401	6.25	0.044	0.613
5	2fnqB	0.401	6.33	0.050	0.620
6	3dy5A	0.399	6.24	0.047	0.615
7	1lox	0.398	6.33	0.041	0.613
8	4g32A	0.398	6.46	0.054	0.620
9	3o8yA2	0.398	6.37	0.037	0.618
10	4nreA	0.398	6.50	0.54	0.620

Table 1:- Table shows Protein structurally close to the target in the PDB with score values

ExPASy:-

For Physico-chemical properties were computed using the Expasy's ProtParam server. Prosite database of protein families and domains was used to compute the functional sites of protein. The basic procedures to get the result are as follows:-

- ✓ Firstly open the Browser www.google.co.in
- ✓ Then put "*web.**expasy**.org/**protparam***" in the Google box.
- ✓ Now open another browser for NCBI home page to getting the sequence of K-19 and get copied it.
- ✓ After open the home page, Paste the sequence of desired protein (K-19) which can get from the NCBI home page in "Protparam sequence box".
- ✓ Now compute the parameters and get the result are as follows:-

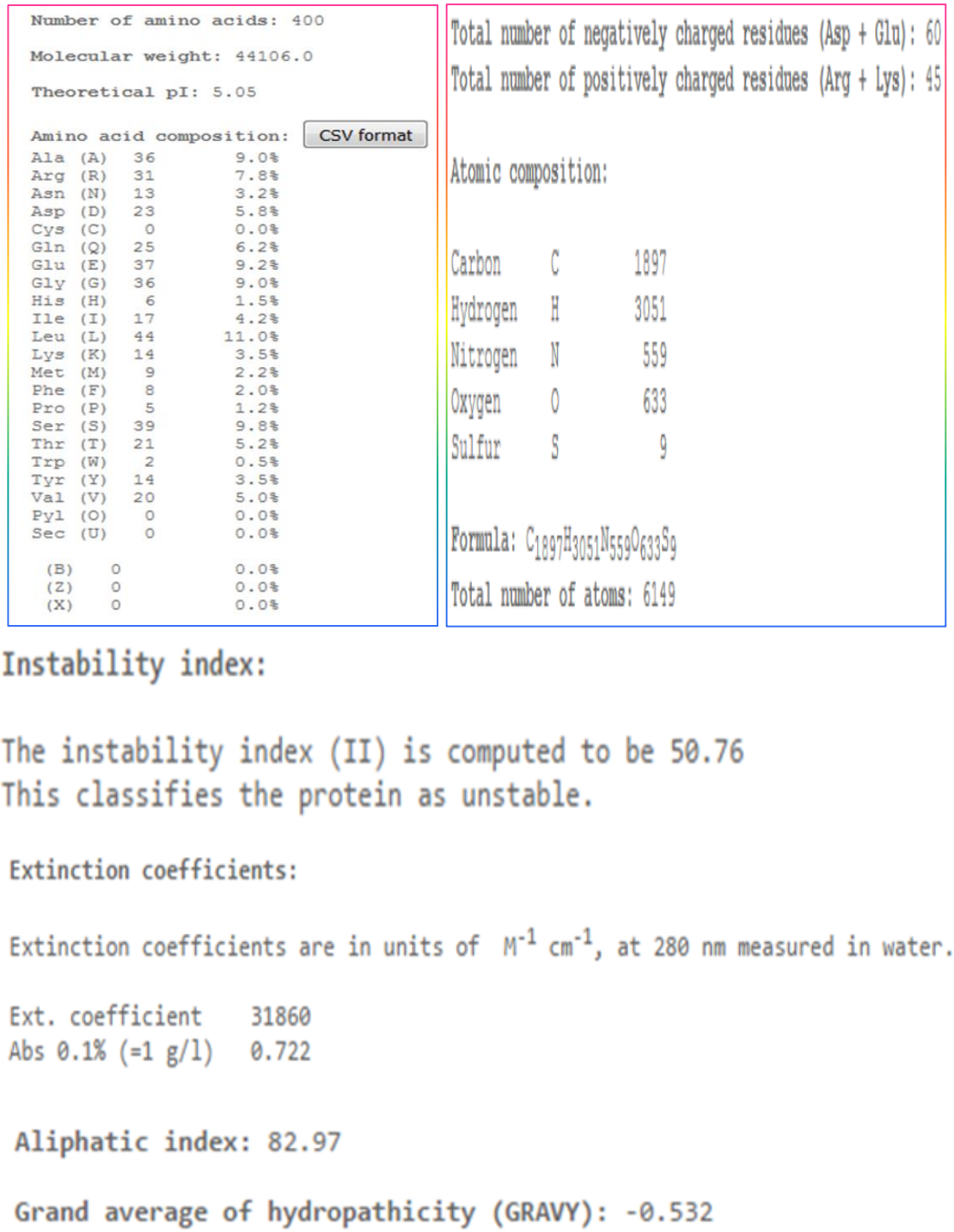


Fig 6:- Physico-chemical characterization of CYFRA -21-1 protein

PROCHECK: - It is used to predict the stereo chemical quality of the selected/ particular model. The model should be in the form of PDB format. This PROCHECK gives the information about all the amino acids with main chain parameter, side chain parameter and secondary structure assignment which are clearer in the diagram.

```
Check of stereochemical quality - PROCHECK v.3.5 (Apr 1998)
-----------------------------------------------

Coordinates file:   [ model1.pdb ]

Chain:              [ A ]

Resolution:          1.5


Running clean-up on file: model1.pdb

 * Side chain atoms swapped for residues:
 **** ILE residue has wrong chirality at CB   36
 * THR A   2  TYR A   4  ARG A   7  LEU A  18  PHE A  31  ARG A  32  TYR A  61
 * TYR A  65  THR A  70  LEU A  82  ARG A  90  ARG A  99  GLU A 102  GLU A 107
 **** ILE residue has wrong chirality at CB  151
 * ARG A 113  ARG A 125  TYR A 127  TYR A 131  ARG A 138  ASP A 139  PHE A 165
 * ARG A 166  PHE A 169  ARG A 176  ASP A 182  ARG A 187  ARG A 188  ARG A 197
 * GLU A 205  TYR A 213  ARG A 226  GLU A 236  ASP A 238  ARG A 253  GLU A 257
 * ARG A 264  ASP A 266  ARG A 274  ARG A 280  ARG A 293  ARG A 300  GLU A 307
 * LEU A 310  GLU A 328  ARG A 330  PHE A 331  GLU A 346  ARG A 353  ASP A 355
 * ARG A 358  TYR A 363  ARG A 365  ARG A 372  GLU A 376  TYR A 380  LEU A 384
 * GLU A 385  GLU A 388  TYR A 391
 * Program completed

.....

Secondary structure assignment

 * Program completed

.....

Non-bonded interactions

* Program completed

.....

Calculation of bond lengths and bond angles

 * Program completed

.....

Phi-psi and chi1-chi2 distributions


 Main Ramachandran plot          * File: model1_01.ps
 All-residue Ramachandran plots  * File: model1_02.ps
 All-residue chi1-chi2 plots     * File: model1_03.ps
 * Program complete

.....

Stereochemical quality plots and residue-by-residue listing
```

```
Stereochemical quality plots and residue-by-residue listing


 Main-chain parameters          * File: model1_04.ps
 Side-chain parameters          * File: model1_05.ps
    * D-amino acid: THR A    2
    * D-amino acid: SER A   10
    * D-amino acid: ALA A   11
    * D-amino acid: SER A   13
    * D-amino acid: SER A   35
    * cis-peptide:  GLY A   39
    * D-amino acid: SER A   46
    * D-amino acid: SER A   49
    * cis-peptide:  SER A   54
    * D-amino acid: ALA A   60
    * cis-peptide:  GLY A   64
    * cis-peptide:  TYR A   65
    * D-amino acid: VAL A   68
    * D-amino acid: THR A   70
 Residue properties plot         * File: model1_06.ps
    * D-amino acid: TRP A  115
    * D-amino acid: PRO A  123
    * D-amino acid: HIS A  129
    * cis-peptide:  GLU A  147
    * D-amino acid: ILE A  151
    * D-amino acid: ILE A  308
    * D-amino acid: LYS A  370
 *
 * Detailed listing: model1.out
 *
 * Program complete

.....................................................................
```

```
Main-chain bond-lengths and angles, and planar groups


 Main-chain bond lengths         * File: model1_07.ps
 Main-chain bond angles          * File: model1_08.ps
 RMS distances from planarity    * File: model1_09.ps
 Distorted geometry              * File: model1_10.ps
    * Distorted main-chain bonds:        121
    * Distorted main-chain angles:       146
    * Distorted planar groups:           24
 * Program complete
 *
 * Summary page: model1.sum
 *


.....................................................................
```

VERIFY 3D:-

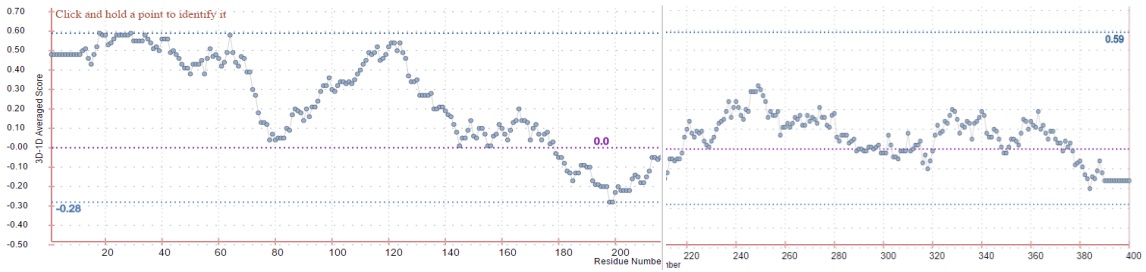


Fig 7:- Verify 3D plot of model 1

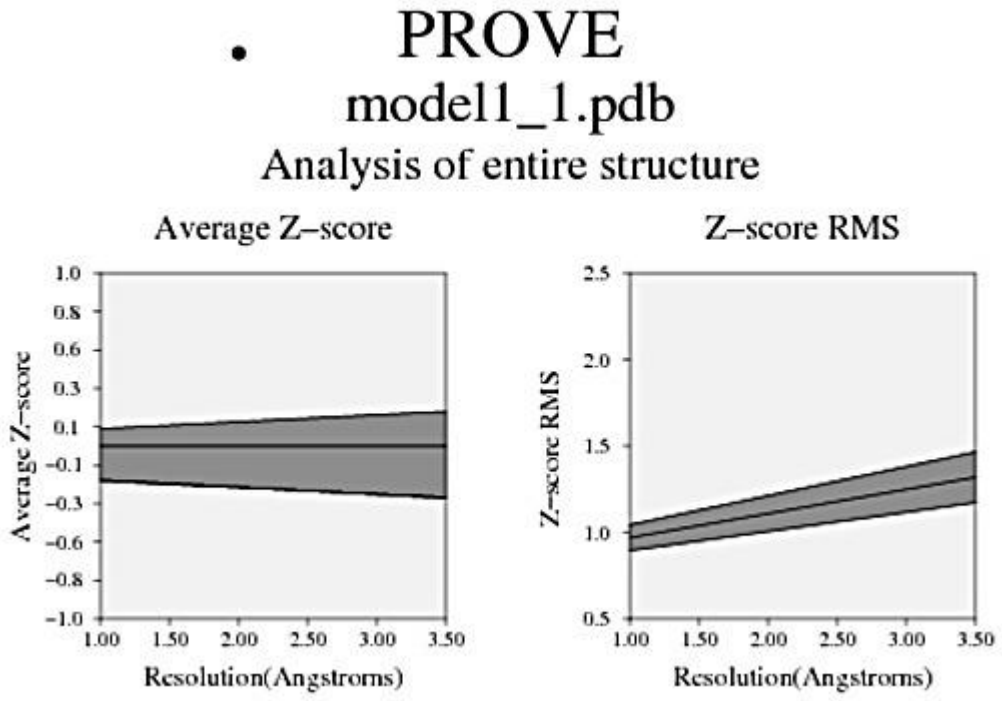


Fig 8:- PROSA analysis shows the average Z-score and RMS value

Prosa Analysis:-

- ✓ Firstly open another window in Browser and follow https://prosa.services.came.sbg.ac.at/prosa.php URL. Here we need to upload Keratin 19 **pdb** file and hit to analyze.
- ✓ Now we get overall model quality, local model quality and 3-D structure of Keratin 19 show highest energy and lowest energy area.

Result: - The following diagram shows that varies of z-score of all model proteins usually found for pure or native proteins of same size that shows the nice quality of

the model. The Energy Plot for CYFRA-21-1 with chain length (400 AA) and Z-score (-3.07) is clearly shown in figure.

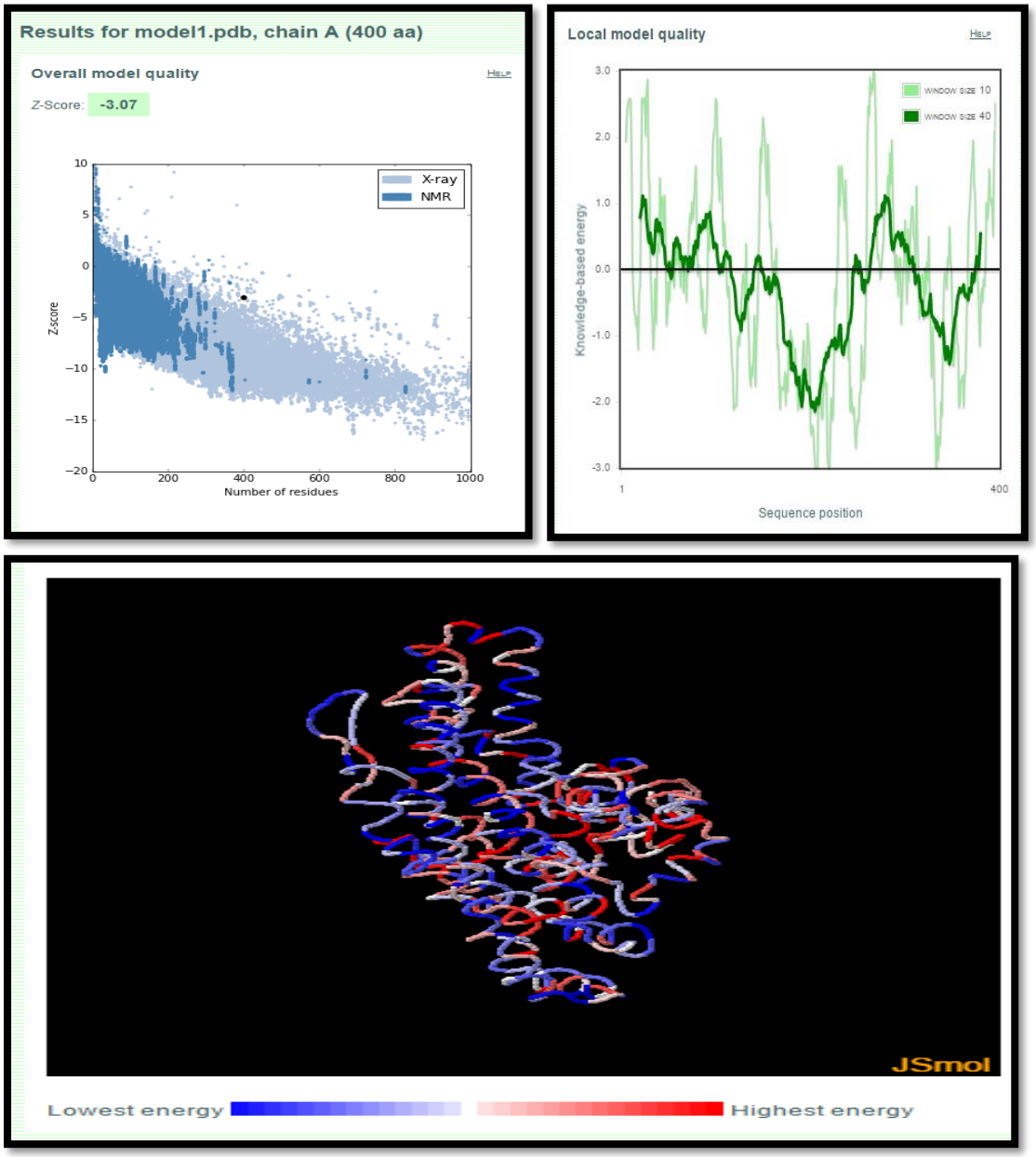


Fig 9:- PROSA analysis of keratin 19 by Z-score and Local model quality

Ramachandran Plot:-

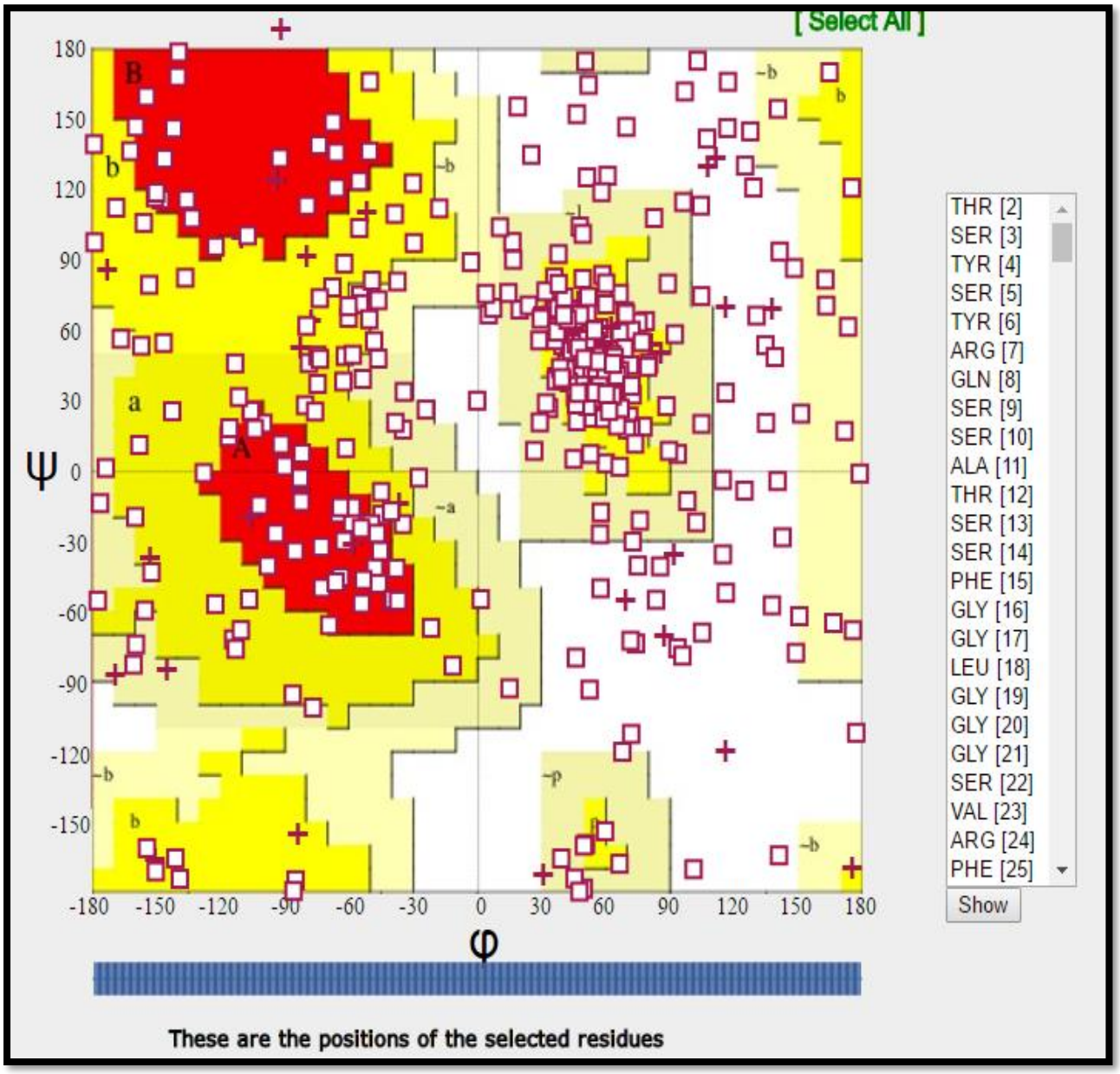


Fig 10:- Ramachandran plot of CYFRA-21-1

String:-

The string is a studying pattern in DNA sequences which use to evaluate two algorithms used for DNA comparison i.e., Longest Common Substring and Sequences (LCS, LCSS). Comparing variation in these regions allows scientists to answer the question of whether two different DNA samples come from the same person.

The procedure of STRING is as follows:-

1. Firstly *www.string-db.org* in Google box.
2. And then select the search option where “Protein by sequences” option should be select.
3. Now open NCBI in other window and select the NCBI sequence of Keratin 19 having 400AA.
4. Paste this sequence in “Protein by sequences” box and select the organism like *“homo-sapiens”* and click search.
5. And then we can get result like:-

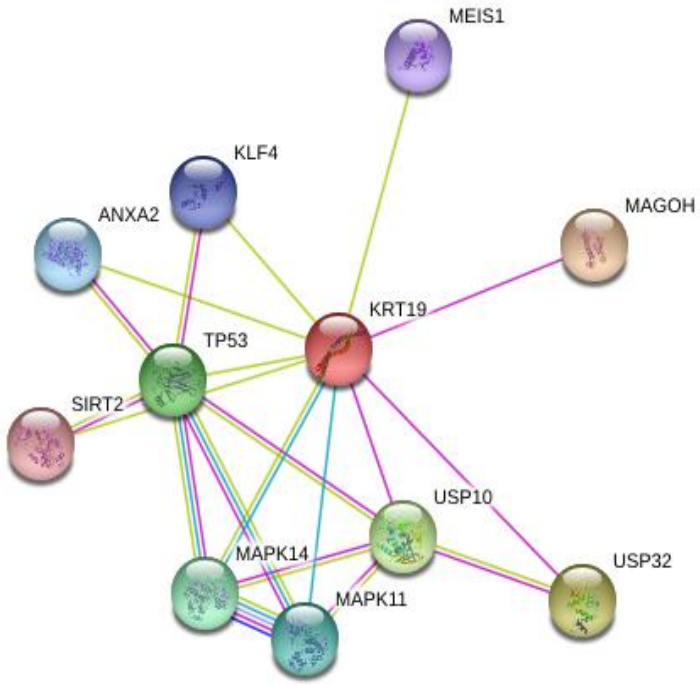


Fig 11:- Interconnection of KRT-19 protein with others evidence

In the above figure, KRT-19 is interconnected with 11 other proteins which like an evidence MAGOH, MEIS-1, KLF-4, TP53, USP32, USP10, ANXA2, SIRT2, MAPK14, MAPK11, USP32. These some proteins are connected with each other, and different colours represent types of evidence for association. Network status shows the number of nodes, edges, with its average node degree, clustering coefficient, expected number of edges and PPI enrichment p-value.

Network Stats

number of nodes:	11	expected number of edges:	13
number of edges:	20	PPI enrichment p-value:	0.0328
average node degree:	3.64		
clustering coefficient:	0.85		

your network has significantly more interactions than expected (what does that mean?)

Functional enrichments in your network

Note: some enrichments may be expected here (why?)

Biological Process (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0032268	regulation of cellular protein metabolic process	9	0.000585
GO:2000377	regulation of reactive oxygen species metabolic process	4	0.0029
GO:0030097	hemopoiesis	5	0.00822
GO:0048646	anatomical structure formation involved in morphogenesis	6	0.00822
GO:0001932	regulation of protein phosphorylation	6	0.00878

(more ...)

Cellular Component (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0043218	compact myelin	2	0.0167
GO:0043220	Schmidt-Lanterman incisure	2	0.0167

KEGG Pathways

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
05014	Amyotrophic lateral sclerosis (ALS)	3	0.000642
04722	Neurotrophin signaling pathway	3	0.00364
05160	Hepatitis C	3	0.00364
05169	Epstein-Barr virus infection	3	0.0087
05205	Proteoglycans in cancer	3	0.0106

(more ...)

INTERPRO Protein Domains and Features

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
IPR008352	Mitogen-activated protein (MAP) kinase, p38	2	0.00504
IPR003527	Mitogen-activated protein (MAP) kinase, conserved site	2	0.0302

Detection of salivary CYFRA-21-1 protein via biosensor

After the predicting 3D structure of CYFRA-21-1 by bioinformatics tools, we also used to detect the salivary CYFRA-21-1 protein via biosensor. For this process firstly we use to do synthesis process with MWCNT and $nZrO_2$ with APTES

❖ **Synthesis process of MWCNT-$nZrO_2$ Nanocomposite**:-

- Firstly Take 25mg of MWCNT (carbon nanotubes) was dispersed in 40ml of deionized water and sonicate it, till MWCNT dispersed.
- Secondly, we make the solution that contains :-
  - ✓ 0.2M Zirconium ethoxide
  - ✓ 0.8M sodium hydroxide
  - ✓ 0.1M solution CTAB(**C**etyl **T**rimethyl**a**mmonium **b**romide)
- Mixed both the solution in deionized water.
- Now, solution of sodium hydroxide added in zirconium ethoxide solution and stirred for 2 hours.
- After 2 hours, add CTAB solution dropwise in above solution and kept stirring for 2 hours.
- Thus prepared mixture was further added in dispersed CNT solution and thereafter it was stirred overnight at 60 ºC.
- Above solution is transferred to Teflon vessel containing hydrothermal pressure tank at 170ºC for 17-18 hours.
- After cooling, the resulting product was washed with deionized water till pH become neutral.

❖ **Functionalization of $nZrO_2$/MWCNT nanocomposite with APTES:-**

- Firstly Take 50mg of $nZrO_2$/MWCNT in isopropane (35ml) and stirred at 40ºC at 300rpm till dissolving time.
- Add 10ml of ddH2O
- Now add 40mg of APTES in the above solution dropwise and then kept for stirring for 48 hours at same temperature and 300 rpm.
- Now washed the solution for 5-6 times by centrifugation machine.
- After that filter the solution by Whattman paper No-1 then,
- Dry it at 60 ºC till dry in hot air oven.
- Sample formed/obtained in powder form.
- Crush it by pestle & Mortal and store in vial and marks it for your identification.

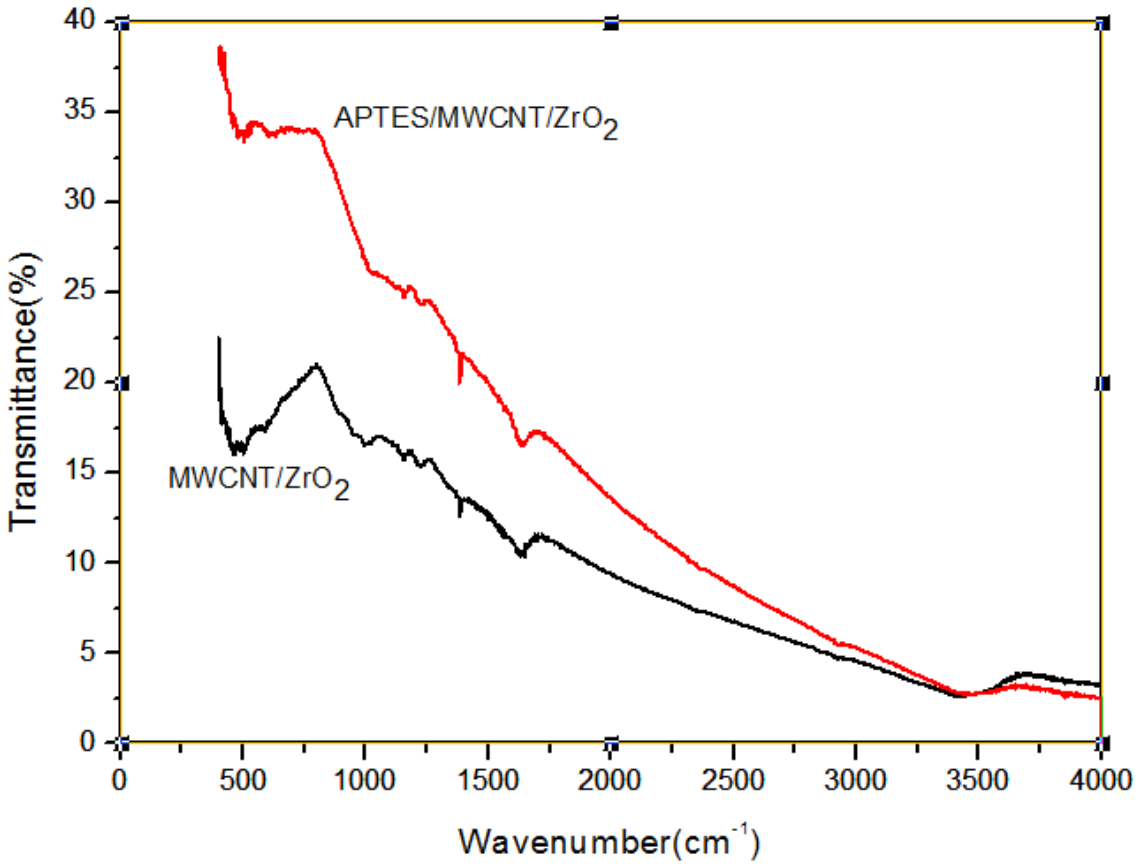


Fig 12:- FT-IR graph of (a) ZrO_2/MWCNT/APTES, (b) MWCNT/ZrO_2

- **Fabrication of biosensing platform:**
- Functionalized nanocomposite was deposited on ITO coated glass substrate using EPD technique.
- CFYRA-21-1 antibody was immobilized on to the fabricated electrode using EDC/NHS chemistry.
- Finally BSA was immobilized for blocking non-specific active sites.

X-Ray Diffraction may be a technique that is employed to calculate info on the layer spacing, the impurities and also the structural strain. Generally, we have a tendency to justify that XRD is maybe the extremely used analytical techniques for characterization materials. Within which the sample is employed in powder type, that consisting of fine grains of crystalline materials for the study. The figure shows the standard XRD pattern of the MWCNTs and nZrO2/MWCNTs. the best or strongest optical phenomenon peak at the angle (2θ) of 26.130°, 42.934° and 43.872° it may be indexed as (002), (100) and (101) it shows the reflection of hexagonal graphite structure which indicates as MWCNTs was acid-oxidized

without significant damage since any decreases [Rosca et al., 2005; Saleh et al., 2011; lutetium et al., 2008] and peak 34.486° & 45.249° attributed to ZrO_2. MWCNTs build the XRD peaks broader and facilitate to shifts the height towards lower angles. Peak 33 shift from 34 are clearly shown in figure, XRD of the $nZrO_2$/MWCNT/APTES verified the characteristic peaks occurring at 2θ = 31.628°, 34.685°, 41.767°, 45.249°, 49.503° and 50.445° corresponding to the lattice planes (111), (020), ($\bar{1}$21), ($\bar{2}$02), (220) and ($\bar{2}$21).

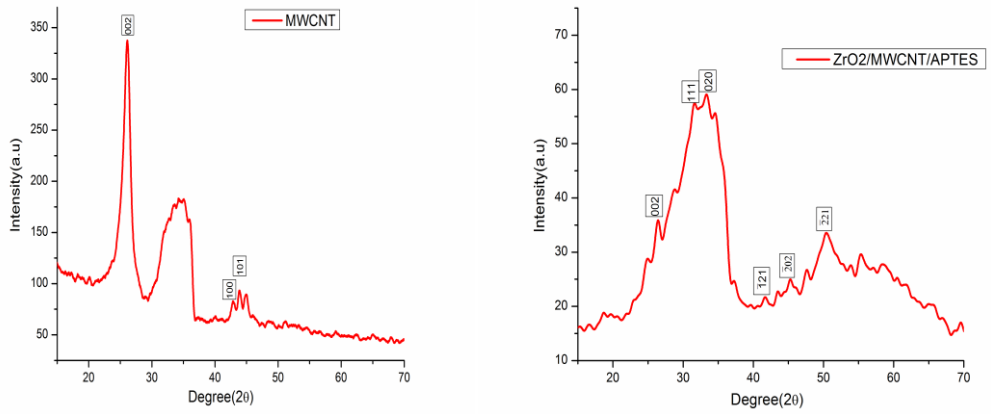


Fig 13: - XRD (a) MWCNT and (b) ZrO_2/MWCNT/APTES

CONCLUSION & FUTURE PROSPECTS

The conclusion of the above study related to CYFRA-21-1 shows that, 3D structure of CYFRA-21-1 has been predicted by using I-TASSER and Swiss-Prot tools. Some other parameters like isoelectric point, molecular weight, extinction coefficient, instability index, aliphatic index and GRAVY (grand average hydropathy) are used to determine its physiochemical characteristics. Some other characterizations like Ramachandran plot and verify-3D have also been obtained through SAVE-sever. Further we have also synthesis as well as functionalized ZrO_2-MWCNT for fabrication of biosensing platform for detection of CYFRA-21-1 biomarker present in saliva sample of oral cancer patient.

Some future prospectuses which can also be done are:-

- The obtained 3-D structure needs to submit in PMDB.
- We will also investigate the reaction mechanism between CYFRA-21-1 antigen and anti-CYFRA-21-1 molecules
- Further we will also fabricate biosensing platform and use for the electrochemical detection of CYFRA-21-1 biomarker.

References:-

1. www.ncbi.nlm.nih.gov
2. *zhanglaInb.ccmb.med.umich.edu/**I-TASSER***
3. www.expasy.org
4. S. Kumar, S. Kumar, S. Tiwari, Dr. S. Srivastava, Dr. S. Kumar, Dr. J. G. Sharma, Prof. S. Maji, Prof. B. D. Malhotra Nanobioelectronics Laboratory
5. Department of Biotechnology, Delhi Technological University New Delhi 110042, India
6. A. Gorschinski , G. Khelashvili , D. Schild , W. Habicht , R. Brand , M. Ghafari , H. Bönnemann , E. Dinjus , S. Behrens , J. Mater. Chem. 2009, 19, 8829.
7. D. M. Winn , W. J. Blot , C. M. Shy , L. W. Pickle , A. Toledo, J. F. Fraumeni Jr. , N. Engl. J. Med. 1981 , 304 , 745 .
8. M. Gonzalez , M. Artimez , L. Rodrigo , C. Lopez-Larrea , M. Menendez , V. Alvarez , R. Perez , M. Fresno , M. Perez , A. Sampedro , J. Clin. Pathol. 1997, 50, 212.
9. L. L. Patton , J. B. Epstein , A. R. Kerr , J. Am. Dental Assoc. (1939) 2008 , 139 , 896 .
10. Rosenberg, Dara, and Shan Cretin. "Use of meta-analysis to evaluate tolonium chloride in oral cancer screening." Oral Surgery, Oral Medicine, Oral Pathology 67.5 (1989): 621-627.
11. Arya, Sunil K., and Shekhar Bhansali. "Lung cancer and its early detection using biomarker-based biosensors." Chemical reviews 111.11 (2011): 6783-6809.
12. L. Wang , Y. Wang , J. I. Wong , T. Palacios , J. Kong , H. Y. Yang , Small 2014, 10 , 1101 .
13. P. R. Solanki , A. Kaushik , V. V. Agrawal , B. D. Malhotra , NPG Asia Mater. 2011, 3, 17.
14. M. Das , C. Dhand , G. Sumana , A. Srivastava , N. Vijayan , R. Nagarajan , B. Malhotra , Appl. Phys. Lett. 2011, 99, 143702.
15. N. Zhao , D. Pan , W. Nie , X. Ji , J. Am. Chem. Soc. 2006 , 128 , 10118 .
16. F. Bellezza , A. Cipiciani , M. A. Quotadamo , Langmuir 2005 , 21, 11099 .
17. S. Kumar, S. Kumar, M. Ali, P. Anand , V. V. Agrawal , R. John , S. Maji , B. D. Malhotra , Biotechnol. J. 2013, 8, 1267.

18. Z. Wang , J. Zhang , Y. Guo , X. Wu , W. Yang , L. Xu , J. Chen , F. Fu , Biosensor Bioelectron. 2013, 45, 108.
19. C.-Y. Yang, E. Brooks, Y. Li, P. Denny, C.-M. Ho , F. Qi , W. Shi, L. Wolinsky, B. Wu, D. T. Wong, Lab Chip 2005, 5, 1017.
20. R. Malhotra , V. Patel , J. P. Vaqué , J. S. Gutkind , J. F. Rusling , Anal Chem. 2010, 82, 3118.
21. T. Li, M. Yang, Sens. Actuators B: Chem. 2011, 158, 361.
22. R. Nagler , G. Bahar , T. Shpitzer , R. Feinmesser , Clin. Cancer Res. 2006, 12, 3979.
23. J. Jose, P. Sunil, Madhavan R. Nirmal, S. S. Varghese, Oral Maxillofacial Pathol. J. 2013, 4, 368.
24. K. Rajkumar , R. Ramya , G. Nandhini , P. Rajashree , Ramesh A. Kumar, Nirmala, S. Anandan, Oral Dis. 2013, 21, 90.
25. D. M. Winn , W. J. Blot , C. M. Shy , L. W. Pickle , A. Toledo, J. F. Fraumeni Jr., N. Engl. J. Med. 1981, 304, 745 .
26. M. Gonzalez , M. Artimez , L. Rodrigo , C. Lopez-Larrea, M. Menendez , V. Alvarez , R. Perez , M. Fresno , M. Perez, A. Sampedro , J. Clin. Pathol., 1997, 50, 212.
27. D. Rosenberg, S. Cretin, Oral Surg. Oral Med., Oral Pathol. 1989, 67, 621.
28. S. K. Arya, S. Bhansali, Chem. Rev. 2011, 111, 6783.
29. P. R. Solanki, A. Kaushik, V. V. Agrawal, B. D. Malhotra, NPG Asia Mater. 2011, 3, 17.
30. M. Das, C. Dhand, G. Sumana, A. Srivastava, N. Vijayan, R. Nagarajan, B. Malhotra, Appl. Phys. Lett. 2011, 99, 143702.
31. N. Zhao , D. Pan , W. Nie , X. Ji , J. Am. Chem. Soc. 2006 , 128 , 10118 .
32. F. Bellezza, A. Cipiciani, M. A. Quotadamo, Langmuir 2005, 21, 11099.
33. Anderson N. L. & Anderson N. G, molecular and cellular proteomics, 2002; 1(11):pp. 845–867.

34. Baldini C, Giusti L, Ciregia F, Da Valle Y, Giacomelli C, Donadio E, Ferro F, Galimberti S, Donati V, Bazzichi L, Bombardieri S, Lucacchini., J Transl Med, 2011; 9: 188.

35. A. J. Bandodkar, J. Wang, *Trends Biotechnol.* 2014, 32, 363.

36. M. Das.et.al., *Appl. Phys. Lett,* 2011, 99, 143702.

37. Roy, Sudeep, et al. "Structure prediction and functional characterization of secondary metabolite proteins of Ocimum." *Bioinformation* 6.8 (2011): 315-319.

37. Subazini, Thankaswamy Kosalai, and Gopal Ramesh Kumar. "Characterization of Lovastatin biosynthetic cluster proteins in Aspergillus terreus strain ATCC 20542." Bioinformation 6.7 (2011): 250-254.

38. Wu, Songfeng, and Yunping Zhu. "ProPAS: standalone software to analyze protein properties." Bioinformation 8.3 (2012): 167-169.

39. Venugopal, Ramadevi Mohan& Subhashree. "Computational structural and functional analysis of hypothetical proteins of Staphylococcus aureus." (2012).

40. Ojha, Ravi Kant, Ajay Kumar Shah, and Roshan Kumar Mandal. "Structure Prediction and Functional Characterization of Some Hypothetical Proteins of Oryza Sativa." Biomirror 4.9 (2013).

41. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402 (1997).

42. Wu, S., Skolnick, J. & Zhang, Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol. 5, 17 (2007).

43. Zhang, Y., Kolinski, A. & Skolnick, J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys. J. 85, 1145–1164 (2003).

44. Li, Y. & Zhang, Y. REMO: a new protocol to refine full atomic protein models from C-α traces by optimizing hydrogen-bonding networks. Proteins 76, 665–676 (2009).

45. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000).

46. Rosca, I. D., Watari, F., Uo, M., & Akasaka, T. (2005). Oxidation of multiwalled carbon nanotubes by nitric acid. *Carbon*, 43, 3124-31.

47. Saleh, Tawfik A. "The role of carbon nanotubes in enhancement of photocatalysis." Syntheses and Applications of Carbon Nanotubes and Their Composites (2013).