Journal of Advanced Bioinformatics Applications and Research
ISSN 0976-2604.Online ISSN 2278 – 6007
Vol 4, Issue2, 2013, pp30-42
http://www.bipublication.com

COMPARATIVE PROTEIN STRUCTURE MODELING OF CYTOCHROMEP450 26A1(CYP26A1) AND PREDICTION OF LIGAND BINDING SITES.

Madhu Yadav[1]* and Gurmit Singh[2]

[1]*Department of Computational Biology & Bioinformatics, Sam Higgonbottom Institute of Agriculture,Technology & Sciences, (Formerly Allahabad Agricultural Institute), Allahabad – 211007, Uttar Pradesh, India.
[2]Department of Computer Science & IT, Sam Higginbottom Institute of Agriculture,Technology & Sciences (Deemed University), Allahabad-211007, India
* Corresponding author.- Email : madhuyadav2003@gmail.com.

[Received-02/04/2012, Accepted-22/04/2013]

ABSTRACT :

The Cytochrome P450 superfamily is a large and diverse group of enzymes. The function of most CYP enzymes is to catalyze the oxidation of organic substances. The substrates of CYP enzymes include metabolic intermediates such as lipids and steroidal harmones, as well as xenoboitic substances such as drug and other toxic chemicals. Homology models of Cytochrome P450 RAI (CYP26A1) were constructed using three human P450 structures, CYP2C8, CYP2C9 and CYP3A4 as a templates for the model building. Used Modeller 9v9 software for the lowest energy CYP26A1 and side chain environment. Modeller is computer a computer program that models protein structure by satisfaction of spatial restraints. It can be used in all stages of comparative modeling described so far, including template search, target – template alignment and model building. Further the ligand site optimization of the CYP26A1 using template CYP3A4 was performed by molecular dynamics to generate a final CYP26A1 model. 3DLigandsite is a web server used for the prediction of ligand binding sites. It is based upon successful manual methods used eighth round of Critical Assessment of Techniques for Protein Structure Predictions (CASP8). The modeling of the three dimensional structure of the protein was performed by three homology modeling programs Geno3D, Swiss- model and Modeller. The constructed 3D models energy minimization and optimization was performed using the molecular dynamics program in GROMACS force field using steepest descent minimization algorithms.

KEYWORDS : CYP26A1, Homology modeling, Modeller9v9, all-trans-retinoic acid (atRA), GROMACS.

1.INTRODUCTION

The Cytochrome P450 superfamily is a large and diverse group of enzymes. The function of most CYP enzymes is to catalyze the oxidation of organic substances. The substrates of CYP enzymes include metabolic intermediates such as lipids and steroidal harmones, as well as xenoboitic substances such as drug and other toxic chemicals. The most common reaction catalyzed by Cytochrome P450 is a monooxygenase reaction, e.g. insertion of one atom of oxygen into an organic substrate (RH). While the other oxygen atom is reduced to water $RH + O_2 + 2H^+ + 2e^- \rightarrow ROH + H_2O$ plays a key role in retinoic acid metabolism. Acts on retinoids, including all-trans-retinoic acid (RA) and its stereoisomer 9Cis-RA. Capable of both

4- hydroxylation and 18-hydroxylate. Responsible for generation of several hydroxylated forms of RA, including 4-OH-RA, 4-oxo-RA and 18-OH-RA. This use specificity highest levels in adult liver, heart, pituitary gland, adrenal gland, placenta and regions of the brain[1,3,5,7].

In human, CYP26A1 is expressed in the liver, heart, pituitary gland[25] , testis brain and placenta, and has been mapped to chromosomes 10q23-q24 [26]. It is thought that the principal role of CYP26A1 is homeostatic, that is the regulation of intercellular ATRA steady - state levels via a negative feedback loop of cholecalciferol [27] . The enzyme therefore may have a protective functions, as an important regulator of differentiation and a possible modulator of disease states indirectly by controlling ATRA and other retinoid concentrations[37,40].The need for close regulation of RA concentrations at the cellular and tissue levels is well demonstrated by the observation that both a deficiency and an excess of RA, during critical development periods are teratogenic to the embryo, tissue levels of RA are regulated both by synthesis from retinol and by catabolism **(Figure 1).**In recent year studies have provided new sight into the process by which RA is catabolism is markedly increased in response to nutritional or pharmacological situations in which the concentration of RA rises.

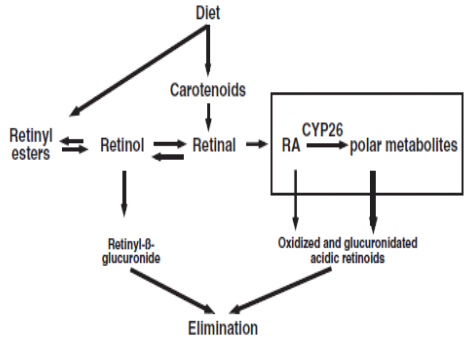


Figure 1. Metabolism of vitamin A, showing production and metabolism of retinoic acid (RA) (boxed). CYP,cytochrome P450.

The active metabolite of Vitamin A, retinoic acid (RA), is a powerful regulation of gene transcription. RA also a therapeutic drug. The oxidative metabolism of RA by certain members of the cytochrome P450(CYP) superfamily helps to maintain tissue RA concentrations with appropiate bounds. The CYP26 family – CYP26A1, CYP26B1 and CYP26C1 is distinguished by being both regulated by and active toward all-trans-retinoic acid (At-RA) while being expressed in different tissue specific patterns. The CYP26A1 gene is regulated by multiple RA response elements. CYP26A1 is essential for postnatal survival as well as germ cell embryonic development. Study of enzyme kinetics have demonstrate that several CYP proteins are capable of metabolizing at-RA, however its likely that CYP26A1 play a major role in RA clearence.

Vitamin A through metabolism to all-trans-retinoic acid (at-RA), plays a crucial role in cellular proliferation and differentiation. The physiological actions of RA begins early in development and continue throughout life[6,4,8]. At-RA functions as an endogenous ligand for nuclear retinoic acid receptors (RAR α, β, and γ), which dimerize with retinoid X receptors (RXR α, β and γ) and bind to specific DNA sites, known as retinoic acid response elements (RAREs), usually located in the promoter regions of the genes that are direct transcriptional targets of RA[2,6,4,9].

The CYP26 family is now recognized as a major contributor to the oxidative metabolism of RA under nutritional and pharmacological conditions. A schematic of the Phase I(oxidation) and Phase II(conjugation) metabolism of RA,as a mediated sequentially by CYP26 and uridine 5'-diphospho(UDP)-glucuronosyl transferase enzyme, is illustrated in **Figure 2.**

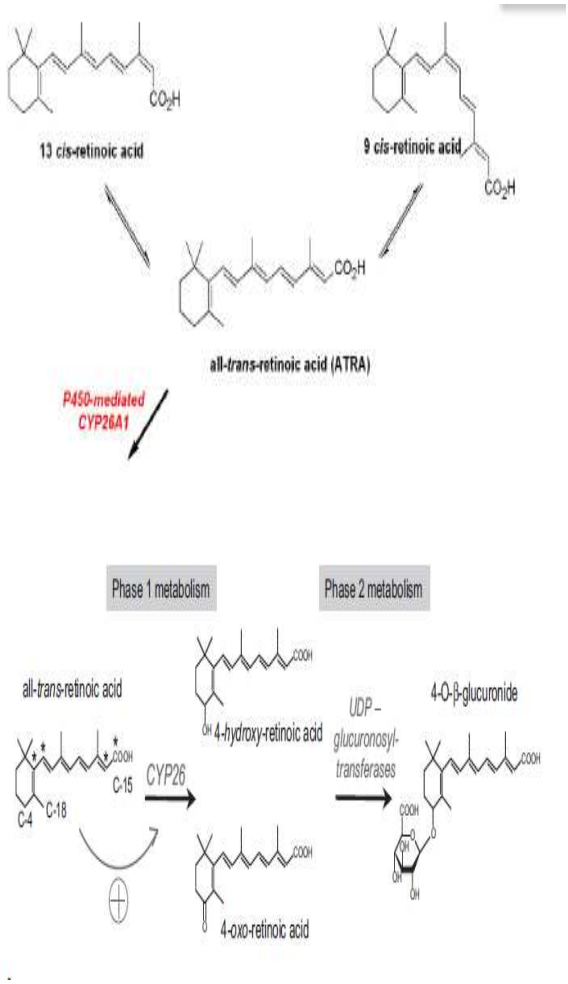


Figure 2 : Metabolism of All-trans- retinoic acid to 4-hydroxy and 4-oxo metabolites (phase I metabolism) followed by glucuronidation to from water soluble more polar metabolisms.

Retinoids in general have been used for some time in the treatment of psoriasis, cystic acne, cutaneous malignancies due to hyperkeratinisation as well as in the treatment of photo-damaged skin[10,11]. Retinoic acid has been used in a number of clinical situations, especially oncology and dermatology, atRA has also been shown to improve the efficacy of other treatments such as radiation, cisplatin and interferon therapies[12,13].

2. METHODS

Amino acid sequences of CYP26A1 protein of human cytochrome P450 was retrieved from the database of National Centre for Biological Information (NCBI). Template was searched using NCBI-BLAST and FASTA program in Protein Data Bank (PDB) and the simillar sequence was taken as a template for the construction of model of CYP26A1. Three dimensional modelling of CYP26A1 protein is done using Modeller9v9. The required input information for this software (from Clustal X), template file from PDB.

Output model is visualised by pymol and structure validation done by using web server SAVS (Structural Analysis and Validation Server). In this we validate the stereochemical properties of model by Ramchandran Plot, PROCHECK, What if and verify 3D. 3D model refinement done by used GROMACS. Energy minimization was performed using molecular dynamics program CHARMM. The constructed 3D models were energy minimized in GROMACS force field using steepest descent minimization Algorithm.

2.1 **Comparative Protein Structure Modeling**

A useful three dimensional (3D) model for a protein of unknown structure (the target) can frequently be built based on one or more related proteins of known structure (the templates). This is the aim of comparative or homology modelling. Homology modelling, also known as comparative modelling is a class of methods for constructing an atomic-resolution model of a protein from its techniques rely on the identification of one or more known proteamino acid sequence[38]. Comparative modeling remains the only method that can be reliably predict the 3D structure of protein with an accuracy comparable to that of low resolution experimental structures. The most basic use of MODELLER in comparative modelling, in which the input are protein data bank (pdb) atom files of known protein structures their alignment with the target sequence to be modelled, and the output is a model for the target that include all non hydrogen atoms. Although MODELLER can find template structures as well as calculate sequence and structure alignments[20,24].

2.1.1 Steps in Comparative Modeling

Comparative modeling usually consists of the following five steps: search for templates, selection of one or more templates, target–

template alignment, model building, and model evaluation (**Figure.3**). Each of these steps is described as follows.

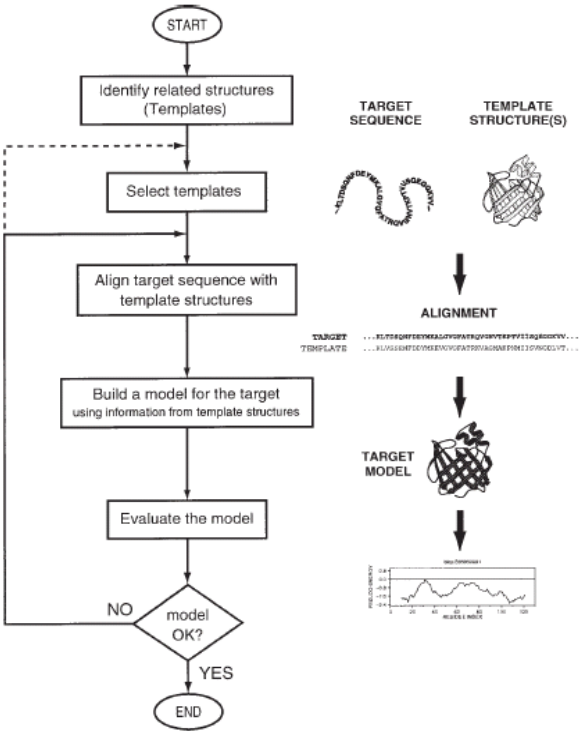


Figure 3 : Steps in comparative protein structure modeling.

2.1.1.1 Search for Templates

Comparative modeling usually starts by searching the database of known protein structures (Protein Data bank, PDB) **[43]** using the target sequence as the query. This is generally done by comparing the target sequence with the sequence of each of the structures in the database. A variety of sequence – sequence comparison methods can be used **[43-44]**. Sometimes, the availability of many sequences related to t he target makes it possible to do more sensitive searching with p rofile methods and Hidden Markov Models (HMM) **[45].** It is also possible to search for templates by evaluating directly the compatibility between the target sequence and each of the structures in the database. This is achieved by fold-recognition methods also known as "threading"**[19]**. Threading uses sequence–structure fitness functions, such as low resolution, knowledge-based force-fields, to evaluate potential target-template matches. The first step is sequence similarity search of CytochromeP450 26A1 retinoic acid metabolizing protein sequence taken from NCBI(**Figure.4**).

```
>P1;CYP26A1
sequence:CYP26A1:::::::0.00: 0.00
MGLPALLASALCTFVLPLLLFLAAIKLWDLYCVSGRDRSCALPLPPGTMGFPFFGETLQM
VLQRRKFLQMKRRKYGFIYKTHLFGRPTVRVMGADNVRRILLGEHRLVSVHWPASVRTIL
GSGCLSNLHDSSHKQRKKVIMRAFSREALECYVPVITEEVGSSLEQWLSCGERGLLVYPE
VKRLMFRIAMRILLGCEPQLAGDGDSEQQLVEAFEEMTRNLFSLPIDVPFSGLYRGMKAR
NLIHARIEQNIRAKICGLRASEAGQGCKDALQLLIEHSWERGERLDMQALKQSSTELLFG
GHETTASAATSLITYLGLYPHVLQKVREELKSKGLLCKSNQDNKLDMEILEQLKYIGCVI
KETLRLNPPVPGGFRVALKTFELNGYQIPKGWNVIYSICDTHDVAEIFTNKEEFNPDRFM
LPHPEDASRFSFIPFGGGLRSCVGKEFAKILLKIFTVELARHCDWQLLNGPPTMKTSPTV
YPVDNLPARFTHFHGEI*
```

Figure 4: Protein sequence of CYP26A1 in PIR format.

2.1.1.2 Template Selection

Once a list of potential templates has been obtained using one or more template searching methods. Usually, the higher the overall sequence similarity (i.e., higher percentage of identical residues, and lower number and shorter length of gaps in the alignment) between the target and the template sequences, the better the template is likely to be. Other factors should also be taken into account when selecting a template:

1. The family of proteins that includes the target and the templates frequently can be organized in subfamilies. The construction of a multiple alignment and a phylogenetic tree[46] can help in selecting the template from the subfamily that is closest to the target sequence **Figure 5**.

2. The similarity between the "environment" of the template and the environment in which the target needs to be modeled should also be considered. The word "environment" is used here in a broad sense, including everything that is not the protein itself: solvent, pH, ligands, quaternary interactions. In particular, the template(s) bound to the same or similar

ligand(s) as the model should be used whenever possible.

3. The quality of the experimental template structure is another important factor in template selection. The resolution and R-factor of a crystallographic structure and the number of restraints per residue for a nuclear magnetic resonance (NMR) structure are indicative of the accuracy of the structure. This information can generally be obtained from the template PDB files or from the articles describing structure determination.

A search for potentially related sequences of known structure can be p erformed using the profile.build() command of MODELLER . The command uses the local dynamic programming algorithm to identify related sequences (Smith and Waterman, 1981; Eswar, 2005). In the simplest case, the command takes as input the target sequence and a database of sequences of known structure and returns a set of statistically significant alignments. Accordingly, the parameters matrix offset and gap penalties 1d are set to the appropriate values for the BLOSUM62 matrix. Used pdb-blast server and Modeller set to use the BLOSUM62 similarity matrix, out of many results selected out most similarity from 1TQN,1WOE,2NNH,1PQ2,1OG5, on the basis of sequence similarity % index and lowest E-value[**17]** is illustrated in **Table 1.**

```
Sequence identity comparison (ID_TABLE):

   Diagonal      ... number of residues;
   Upper triangle ... number of identical residues;
   Lower triangle ... % sequence identity, id/min(length).

        1tqnA @2 1w0eA @2 1pq2A @2 1og5A @2 2nnhA @2
1tqnA @2     468      439       86       89       86
1w0eA @2      97      453       86       89       86
1pq2A @2      19       19      463      351      462
1og5A @2      19       20       76      461      351
2nnhA @2      19       19      100       76      463

Weighted pair-group average clustering based on a distance matrix:

                                                    .----- 1tqnA @2.0     3.0000
                                                    |
                  .---------------------------------------------------- 1w0eA @2.8    80.7500
                  |
                  |                                                 .--- 1pq2A @2.7     0.0000
                  |                                                 |
                  |                                  .------------------ 2nnhA @2.6    24.0000
                  |                                  |
                .---------------------------------------------------- 1og5A @2.5

               +----+----+----+----+----+----+----+----+----+----+----+----+
             83.9800   69.4450   54.9100   40.3750   25.8400   11.3050   -3.2300
                76.7125   62.1775   47.6425   33.1075   18.5725    4.0375
```

Figure 5: Comparison with pairwise sequence distances that can be used as input to the dendogram() calculates a clustering tree.

Table 1: Selected templates shows sequence identity and E-value

Sequence name	PDB Id	Sequence identity %	E-value
CYP3A4	1TQN	24.52	0.0
CYP3A4	1WOE	25.12	0.0
CYP2C8	1PQ2	24.84	0.0
CYP2C8	2NNH	24.84	0.0
CYP2C9	1OG5	23.75	0.0

2.1.1.3 Target-Template Alignment

To build a model, all comparative modeling programs depend on a list that establishes structural equivalences between the target and template residues. This is defined by the alignment of the target and template sequences. The alignment is relatively simple to obtain when the target–template sequence identity is above 40%. In most such cases, an accurate alignment can be obtained automatically using standard sequence– sequence alignment methods. If the target–template sequence identity is lower than 40%, the alignment generally has gaps and needs manual intervention to minimize the number of misaligned residues. In these low-sequence identity cases, the alignment accuracy is the most important factor affecting the quality of the resulting model. However, it is always important to check and edit the alignment by inspecting the template structure visually, especially if the target–template sequence identity is low. A misalignment by only one residue position will result in an error of approximately 4 Å in the model because the current modeling methods cannot recover from errors in the alignment.

A good way of aligning a target sequence and structure is Salign command is based on a dynamic programming algorithm, it is different from standard sequence-sequence alignment methods because it takes into account structural information from the template when constructing an alignment. This task is achieved through a variable gap penalty function that tends to place gaps in solvent exposed and curved regions, outside secondary structure segments, and between two positions that are close in space. As a result, the alignment errors are reduced by approximately

one third relative to those that occur with standard sequence alignment techniques. This improvement becomes more important as the similarity between the sequences decreases and the number of gaps increases.Multiple sequence alignment of CYP26A1 with template sequence whose structure are available in PDB database is illustrated in **Figure 6**.

```
 _aln.pos        10        20        30        40        50        60
1tqnA     ---HSHGLFKKLGIPGPTPLPFLGNIL--SYHKGFCMFDMECHKKYGKVWGFYDGQQPVLAITDPDMI
1w0eA     YGTHSHGLFKKLGIPGPTPLPFLGNIL--SYHKGFCMFDMECHKKYGKVWGFYDGQQPVLAITDPDMI
2nnhA     -----------KLPPGPTPLPIIGNML-QIDVKDICKSFTNFSKVYGPVFTVYFGMNPIVVFHGYEAV
1pq2A     -----------KLPPGPTPLPIIGNML-QIDVKDICKSFTNFSKVYGPVFTVYFGMNPIVVFHGYEAV
1og5A     -------------PPGPTPLPVIGNIL-QIGIKDISKSLTNLSKVYGPVFTLYFGLKPIVVLHGYEAV
CYP26A1   MGLPALLASALCTFVLPLLLFLAAIKLWDLYCVSGRDRSCALPLPPGTMGFPFFGETLQMVLQRRKFL
 _consrvd                *  *     *                 *      *

 _aln.p   70        80        90       100       110       120       130
1tqnA     KTVLVKECYSVFTNRRPFG--PVGFMKSA--ISIAEDEEWKRLRSLLSPTFTSGKL--KEMVPIIAQY
1w0eA     KTVLVKECYSVFTNRRPFG--PVGFMKSA--ISIAEDEEWKRLRSLLSPTFTSGKL--KEMVPIIAQY
2nnhA     KEALIDNGE-EFSGRGNSPISQRITKGLG--IISSNGKRWKEIRRFSLTTLRNFGMGKRSIEDRVQEE
1pq2A     KEALIDNGE-EFSGRGNSPISQRITKGLG--IISSNGKRWKEIRRFSLTTLRNFGMGKRSIEDRVQEE
1og5A     KEALIDLGE-EFSGRGIFPLAERANRGFG--IVFSNGKKWKEIRRFSLMTLRNFGMGKRSIEDRVQEE
CYP26A1   QMKRRKYGFIYKTHLFGRPTVRVMGADNVRRILLGEHRLVSVHWPASVRTILGSGCLSNLHDSSHKQR
 _consrvd                                  *                 *

 _aln.pos  140       150       160       170       180       190       200
1tqnA     GDVLVRNLRREAETGKPVTLKDVFGAYSMDVITSTSFGVNIDSLNNPQDPFVENTKKLL--RFDFL--
1w0eA     GDVLVRNLRREAETGKPVTLKDVFGAYSMDVITSTSFGVNIDSLNNPQDPFVENTKKLL--RFDFL--
2nnhA     AHCLVEELRKT--KASPCDPTFILGCAPCNVICSVVFQKRFDY---KDQNFLTLMKRFNENFRILN--
1pq2A     AHCLVEELRKT--KASPCDPTFILGCAPCNVICSVVFQKRFDY---KDQNFLTLMKRFNENFRILN--
1og5A     ARCLVEELRKT--KASPCDPTFILGCAPCNVICSIIFHKRFDY---KDQQFLNLMEKLNENIEILS--
CYP26A1   KKVIMRAFSREALECYVPVITEEVG-SSLEQWLSCGERGLLVYPEVKRLMFRIAMRILLGCEPQLAGD
 _consrvd                          *       *             *

 _aln.pos    210       220       230       240       250       260       270
1tqnA     -DPFFLSITVFPFLI----------PILEVLNICVFPREVTNFLRKSVKRMKESRLEDTQKHRVDFLQ
1w0eA     -DPFFLSITVFPFLI----------PILEVLNICVFPREVTNFLRKSVKRMKESRL---------FLQ
2nnhA     -SPWIQVCNNFPLLIDCFPGTHNKV--------LKNVALTRSYIREKVKEHQA---SLDVNNPRDFID
1pq2A     -SPWIQVCNNFPLLIDCFPGTHNKV--------LKNVALTRSYIREKVKEHQA---SLDVNNPRDFID
1og5A     -SPWIQVVNNFPALLDYFPGTHNKL--------LKNVAFMKSYILEKVKEHQE---SMDMNNPQDFID
CYP26A1   GDSEQQLVEAFEEMTRNLFSLPIDVPFSGLYRGMKARNLIHARIEQNIRAKICGLRASEAGQGCKDAL
 _consrvd           *

 _aln.pos      280       290       300       310       320       330       340
1tqnA     LMIDSQNSS----HKALSDLELVAQSIIFIFAGYETTSSVLSFIMYELATHPDVQQKLQEEIDAVLPN
1w0eA     LMI-------------DSDLELVAQSIIFIFAGYETTSSVLSFIMYELATHPDVQQKLQEEIDAVLPN
2nnhA     CFLIKMEQEKDNQKSEFNIENLVGTVADLFVAGTETTSTTLRYGLLLLLKHPEVTAKVQEEIDHVIGR
1pq2A     CFLIKMEQEKDNQKSEFNIENLVGTVADLFVAGTETTSTTLRYGLLLLLKHPEVTAKVQEEIDHVIGR
1og5A     CFLMKMEKEKHNQPSEFTIESLENTAVDLFGAGTETTSTTLRYALLLLLKHPEVTAKVQEEIERVIGR
CYP26A1   QLLIEHSWERG---ERLDMQALKQSSTELLFGGHETTASAATSLITYLGLYPHVLQKVREELKSKGLL
 _consrvd                        *          * ***          *   * *  *  **

 _aln.pos       350       360       370       380       390       400
1tqnA     -----KAPPTYDTVLQMEYLDMVVNETLRLFPIAMRL-ERVCKKDVEINGMFIPKGVVVMIPSYALHR
1w0eA     -----KAPPTYDTVLQMEYLDMVVNETLRLFPIAMRL-ERVCKKDVEINGMFIPKGVVVMIPSYALHR
2nnhA     -----HRSPCMQDRSHMPYTDAVVHEIQRYSDLVPTGVPHAVTTDTKFRNYLIPKGTTIMALLTSVLH
1pq2A     -----HRSPCMQDRSHMPYTDAVVHEIQRYSDLVPTGVPHAVTTDTKFRNYLIPKGTTIMALLTSVLH
1og5A     -----NRSPCMQDRSHMPYTDAVVHEVQRYIDLLPTSLPHAVTCDIKFRNYLIPKGTTILISLTSVLH
CYP26A1   CKSNQDNKLDMEILEQLKYIGCVIKETLRLNPPVPGGF-RVALKTFELNGYQIPKGWNVIYSICDTHD
 _consrvd                    *   *  *  *                          ****

 _aln.p  410       420       430       440       450       460       470
1tqnA     DPKYWTEPEKFLPERFSKK--NKDNIDPYIYTPFGSGPRNCIGMRFALMNMKLALIRVLQNFSFKPCK
1w0eA     DPKYWTEPEKFLPERFSKK--NKDNIDPYIYTPFGSGPRNCIGMRFALMNMKLALIRVLQNFSFKPCK
2nnhA     DDKEFPNPNIFDPGHFLDKNGNFKKSD--YFMPFSAGKRICAGEGLARMELFLFLTTILQNFNLKSVD
1pq2A     DDKEFPNPNIFDPGHFLDKNGNFKKSD--YFMPFSAGKRICAGEGLARMELFLFLTTILQNFNLKSVD
1og5A     DNKEFPNPEMFDPHHFLDEGGNFKKSK--YFMPFSAGKRICVGEALAGMELFLFLTSILQNFNLKSLV
CYP26A1   VAEIFTNKEEFNPDRFMLP--HPEDASRFSFIPFGGGLRSCVGKEFAKILLKIFTVELARHCDWQLLN
 _consrvd          * *  *                  **  * * * *   *

 _aln.pos  480       490       500
1tqnA     ET-QIPLKLSLGGLLQPEKPVVLKVESRDGT
1w0eA     ET-QIPLKLSLGGLLQPEKPVVLKVESRDGT
2nnhA     DLKNLNTTAVTKGIVSLPPSYQICFIPV---
1pq2A     DLKNLNTTAVTKGIVSLPPSYQICFIPV---
1og5A     DPKNLDTTPVVNGFASVPPFYQLCFIPV---
CYP26A1   GPPTMKTSPTVYPVDNLPARFTHFHGEI---
 _consrvd
```

Figure 6: The alignment between target sequence and templates sequence in MODELLER PAP format.

2.1.1.4 Model Building

Once an initial target–template alignment has been built, a variety of methods can be used to construct a 3D model for the target protein. The original and still most widely used method is modeling by rigid-body assembly [41,47].This method constructs the model from a few core regions and from loops and side chains, that are obtained from dissecting related structures. Another family of methods, modeling by segment matching, relies on the approximate positions of conserved atoms from the templates to calculate the coordinates of other atoms [48]. The third group of methods, modeling by satisfaction of spatial restraints, uses either distance geometry or optimization techniques to satisfy spatial restraints obtained from the alignment of the target sequence with the template structures[49] .Once a target-template alignment has been constructed, MODELLER calculates a 3D model of the target in a completely automated way.

2.1.1.5 **Model Evaluation**

After a model has been built, it is important to check it for possible errors. Two types of evaluation should be carried out: (1) “internal” evaluation of self consistency that checks whether or not the model satisfies the restraints used to calculate it and (2) “external” evaluation that relies on information that was not used in calculating the model [19,50].

If several models are calculated for the same target. The “best” model can be selected by picking the model with lowest value of the Modeller objective function, which is reported in the second line of the model PDB file. The value of the objective function in the modeller is not an absolute measure in the sense that it can only be used to rank the models calculated from the same alignment. Once a final model is selected, there are many ways to assess it. The DOPE potential in MODELLER is used to evaluated the model fold. Both DOPE and PROCHECK confirm that a reasonable model was obtained with an energy comparable to that of the template. Dope score profile shows the clear differences between the two profile for the long active site loop and the long helices at the C-terminal end of the target sequences.

Produced models MOLPDF and Dope score value shown in **Table (3)**.Model can be selected on the basis of lowest value of MOLPdf and DOPE Score.

Table 3:- summary of successfully produced models:

FILENAME	MOLPDF	DOPE SCORE
CYP26A1.B99990001.PDB	25905.32422	-51026.015625
CYP26A1.B99990002.PDB	24445.68945	-51420.4882
CYP26A1.B99990003.PDB	26125.72461	-50936.35935
CYP26A1.B99990004.PDB	24536.35742	-52355.984315
CYP26A1.B99990005.PDB	25010.58008	-50553.015625

2.2 Molecular Dynamics Simulation

Molecular dynamics is a computer simulation of physical movements of atoms and molecules. The atoms and molecules are allowed to interact for a period of time given a view of the motion of the atoms. The interaction between the particles is either described by a " force field (classical MD)", a quantum chemical model, or a mix between the two. MD serves as an important tool in protein structure determination and also applied with limited success as a method of refining protein structure prediction.

Molecular dynamics simulation of Modelled 3D structure of the CYP26A1 was performed by using GROMACSv4.3(URL:http:\\www.gromacs.org) [33,39] software to tract the motion of individual atoms. Two methods of energy minimization available in GROMACS, the steepest descent method, it simply takes a step in the direction of the negative gradient (hence in the direction of the force), without any consideration of the history built up in previous steps. The step size is adjusted such that the search is fast, but the motion is always downhill. This is a simple and sturdy, but somewhat its convergence can be quite slow, especially in the vicinity of the local minimum!. The faster-converging conjugate gradient method uses gradient information from previous steps. In general, steepest descents will bring you close to the nearest local minimum very quickly, while conjugate gradients brings you very close to the local minimum, but performs worse far away from the minimum. Energy minimization method used to optimize the modeled structure and molecular dynamics used to simulate the natural motion of atoms. Before starting simulation, GROMACS environment was set up and input files for the simulation were prepared, the structure was solvated in water , minimized and equilibrated. Firstly to prepare the topology from the pdb file, 'pdb2gmx' command was used and the remainder of the file involves defining a few other useful topologies, starting with position restraints. The 'prose.itp' file was generated by pdb2gmx, it defines a force constant used to keep atoms in place during equilibration. The solvent water was added around the protein to generate a simulation box used 'genbox' command, Specifying a solute-box distance of 1.0 nm, mean that there are atleast 2.0 nm between any two periodic images of a protein. This distance will sufficient for just about ant cutoff scheme commonly used in simulation. The energy minimization was performed in 1000 steps using the steepest descent minimization algorithm. GROMOS 96 force field was chosen for the calculation of potential energy of the structure. A standard cutoff of the 1.0 nm, both for the neighbor generation and the Coulomb & Lennard-Jones interactions was employed. An equilibrium run of water around the protein was performed to avoid the unnecessary distortion of the protein using the 100 pico second (ps) time scale and 50000 steps (iterations). Finally molecular dynamics simulation was performed using the 2000 ps (2ns) time scale and 1000000 steps at 300 °K temperature and 1atm. pressure. The simulation results were analyzed using the 'grace' program.

2.3 DLigandsite web server : predicting ligand-binding-site

Protein often perform their function on ligands(e.g. enzyme substrates) are regulated by them. Therefore the identification of ligand binding sites is important. The explosion of protein sequences from genome sequencing projects

makes it essential for automated methods to predict ligand binding sites. Further, protein structures are often solved in the absence of ligands, making it important that was able to identify binding sites for such protein [28,29,30,31,32].

3D ligand site is a web serever for the prediction of ligand-binding sites. It is based upon successful manual methods used in the eighth round of the Critical Assessment of techniques for Protein Structure Prediction to provide structural model (CASP8). Ligands bound to structures simillar to the query are superimposed onto the model and used to predict the binding site [33,34]. In benchmarking against the CASP8 target 3Dligandsites obtains a Mathew's correlation coefficient (MCC)[35] of 0.64 and coverage and accuracy of 71 and 60% respectively, simillar results to our manual performance in CASP8 **(Figure 7)** and (T**able 4).** 3DLigandsite is available for use at http:// www.sbg.blo.ic.ac.uk/3dligandsite.

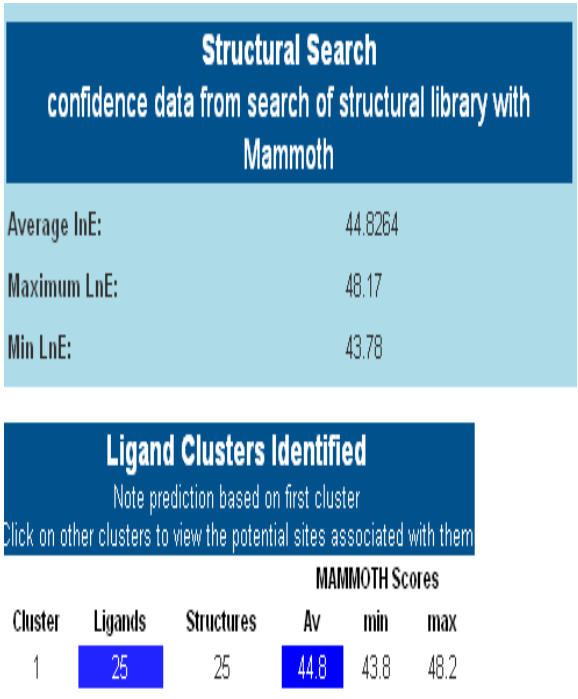


Figure 7: Ligand clusters produced by web server.

Table 4 :-Predicted Binding Site of ligand

Residue	Amino	Contact	Average distance	JS divergance
110	VAL	12	0.61	0.00
112	TRP	25	0.03	0.00
125	LEU	25	0.05	0.42
126	SER	17	0.29	0.00
133	HIS	24	0.04	0.76
144	PHE	24	0.12	0.69
193	LEU	7	0.66	0.38
296	GLU	14	0.33	0.32
297	LEU	24	0.18	0.53
300	GLY	23	0.24	0.61
301	GLY	24	0.16	0.63
304	THR	25	0.01	0.76
305	THR	13	0.57	0.64
308	ALA	20	0.60	0.34
364	LEU	9	0.54	0.66
369	PRO	25	0.06	0.54
370	VAL	25	0.06	0.47
373	GLY	16	0.42	0.26
375	ARG	25	0.07	0.80
396	TYR	25	0.29	0.40
398	ILE	12	0.11	0.33
434	PRO	25	0.20	0.78
435	PHE	18	0.44	0.86
436	GLY	12	0.46	0.64
439	LEU	7	0.13	0.43
440	ARG	25	0.12	0.81
441	SER	9	0.45	0.38
442	CYS	25	0.16	0.92
443	VAL	25	0.00	0.56
444	GLY	17	0.27	0.79
447	PHE	24	0.17	0.69
448	ALA	25	0.34	0.71

The identified ligand clusters identified containing the greatest number of ligands is automatically selected for prediction by Ligandsite[34]. The **Table 4** provides details of the other clusters and allows the user to view the potential sites associated with these clusters. A table lists all of the predicted binding sites residue conservation score (JSD)[37].

3. RESULTS

Comparative modelling methods use structural templates hat have the highest sequence homology with the target protein. Homologous proteins were identified by scanning the protein sequence of CYP26A1[21] obtained from the NCBI against 3D structures deposited in the Protein data bank (PDB)[18] using PSI-BLAST[22]. The search returned aminoacid sequences of different P450 isolated from different species including three human P450 recently deposited in the PDB. The percent sequence identity , chain length and E-value for the homologous P450 sequences are shown in table 1.

CYP3A4 (1TQN), CYP2C8 (1PQ2,2NNH) and CYP2C9 (1OG5) were promising templates with percent sequence identity ranging from 24-26% and good E-values[14,15,16].

To evaluate the quality of the modelled structures the lowest energy model generated from each template was subjected to number of checks. Stereochemical quality was assessed using Ramachandran plots[23] used SAVS server (**Figure 8**), the residue in most favoured region 351 (89.5%) and disallowed regions 1(0.3%). and amino acid enviornment was assessed using verify3D[19] value is 83.52 % in **(Figure 9)** and Errat [20] value is 68.61% shown in **Figure(10)** . PROCHECK is to assess how normal or conversely how unusal the geometry of the residues in a given protein structure is as compared with stererochemical parameters derived from well derived from well refined, high resolution structure. For molecular dynamics used GROMACS software and minimize the energy of modelled structure and the energy minimization parameters are Polak-Ribiere Conjugate Gradients: Tolerance (Fmax) = 1.00000e+03 , Number of steps = 50000 , F-max = 1.23729e+03 on atom 2821,
F-Norm = 7.38221e+01 and the value of Potential energy = **-1.4770752e+06,** because less energy is good for the best model and also molecular dynamics optimized the protein most stable structure shown in **Figure 11** and **Figure 12** shown cluster of lignad binding site .

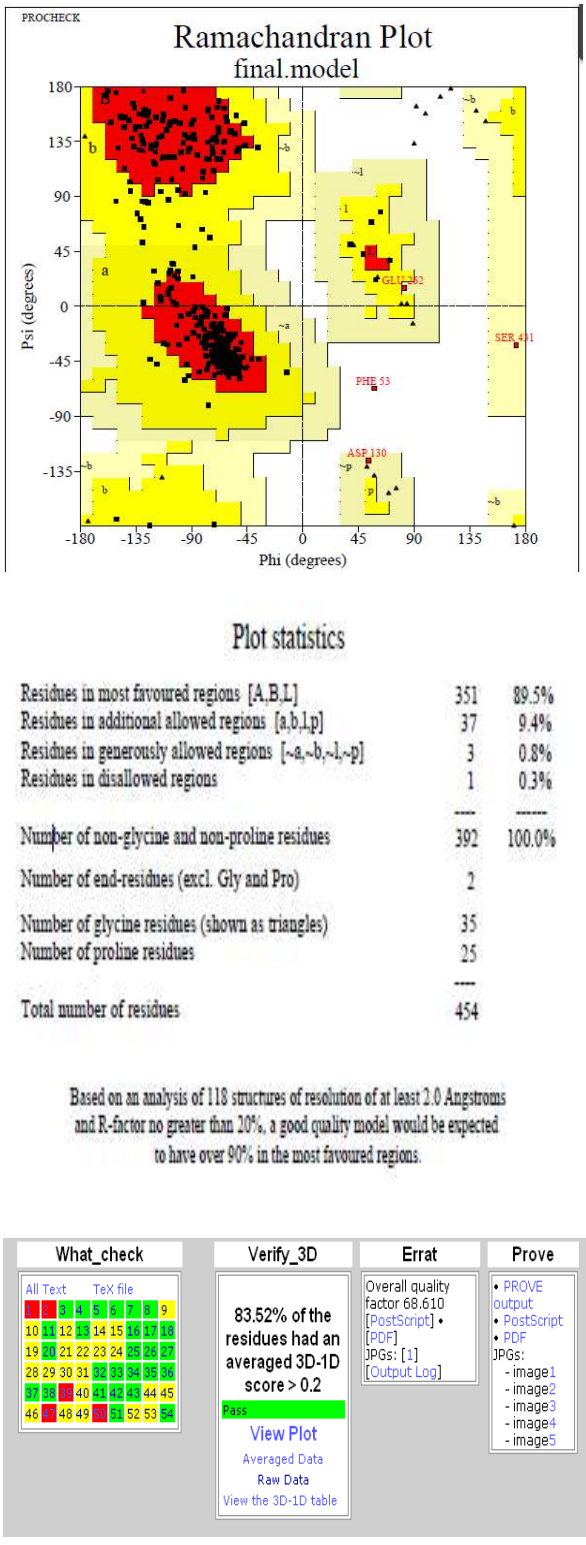


Figure 8 : Shows SAVS (Structural Analysis and Validation Sever) Results.

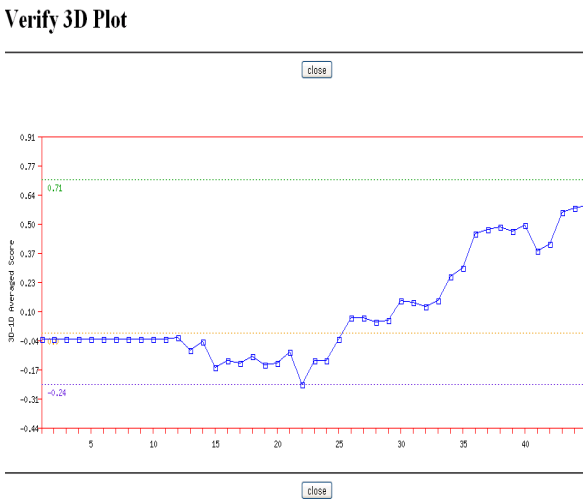
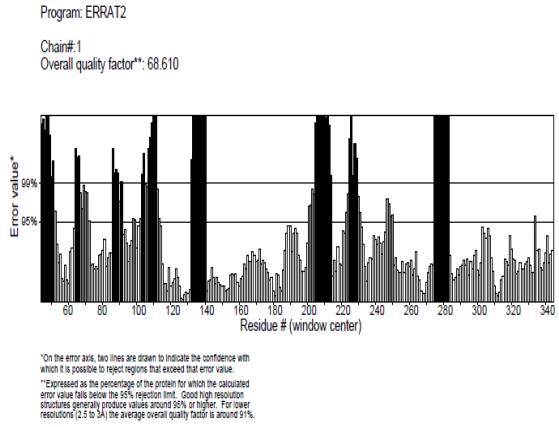


Figure 9 : Stereochemical property validation result through Verify 3D Plot



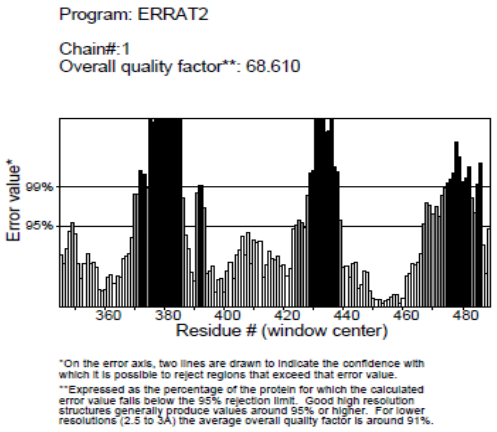


Figure 10 : Stereochemical property validation result through Errat Program

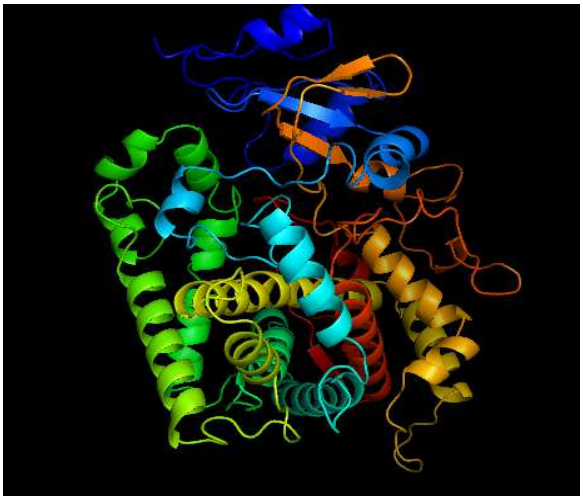


Figure 11 .

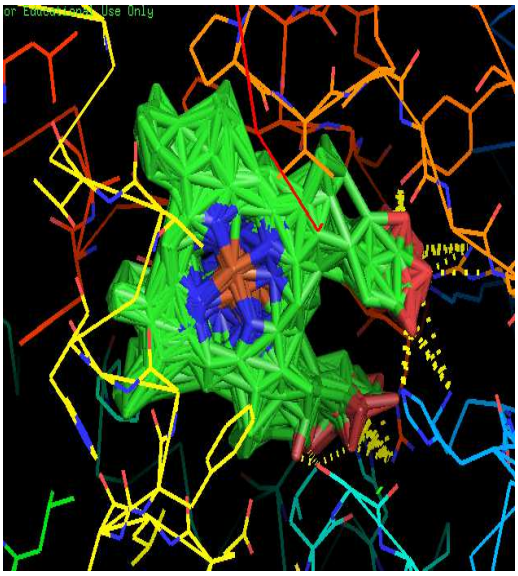


Figure 12.

Figure 11 & 12: Final model of CYP26A1.pdb & Model.pdb show ligand binding site cluster respectively.

4. DISSCUSSION

The 3D structural models of CYP26A1 were predicted by Modeller9v9 software from those protein target-templates alignments (http://www.salilab.org/modeller) and homology detected. To evaluate and identify any anomalies in the predicted model of CYP26A1 it was submitted to the Structural Analysis and Verification server

(http://nihserver.mbi.ucla.edu/SAVES/). SAVES is a metaserver for analyzing and validating protein structures which integrates five modules i.e. Procheck, What_check, Errat, Verify_3D and Prove. To improve the quality of predicted model of CYP26A1, energy minimization was performed with the GROMOS 96 forcefield implementation. This force field permits to evaluate the energy of the modeled structure as well as overhaul distorted geometries through energy minimization.

Molecular dynamics simulation were run using GROMACSv4.3 (URL:http:\\www.gromacs.org)[33]. An electrostatic twin range cut off of 1.8 nm was employed. The vander waals cut off was 1.0 nm. The time step for integration was 2fs , number of steps integration 2ns, using the LINCS algorithm to constrain bond lengths[34]. Pressure and temperature were controllred using the weak coupling method[35].Water (and counter ion) octane, protein and K+ were each coupled separately to a temperature bath at 300K. The pressure in the z direction was close to 1.0 bar using time constant tau_t=0.1 ps , and compressibility Kz=4.5*10-5 bar^-1. The x and y dimensions of the simulation box were fixed. Atomic co-ordinates and velocities were saved every 2ps, and were kept for subsequent analysis of simulation using the 'Grace' program.

The evaluation of 3DLigandSite performance has been benchmarked on the set of structures that were used for the assessment of FINDSITE and on the targets assessed for the ligand-binding category in CASP8 . The FINDSITE data set was filtered using Our list of accepted ligands, which resulted in a set of test structures. 3DLigandSite performance was assessed using a range of distance cut offs between 0.2A˚ and 2.0A˚ at 0.2A˚ intervals and with m in Equation : Threshold = m * cluster size+1set (where m is a constant that determines the proportion of the ligands that need to be within the distance cut off to be predicted as part of the binding site). The threshold needs to account for variation between the modeled and real structure and between the ligands in the cluster. We assessed the predictions using the MCC , and coverage and accuracy, all of which have been used for assessment in recent CASP experiments.

5. CONCLUSION

Homology modelling is a comparative modelling method by which CYP26A1 model created by using following template CYP2C8, CYP2C9 and CYP3A4 because it show good structural alignment having sequence simillarity index 24 – 26 %. Final selected CYP26A1.B99990002.pdb having 25 ligand cluster at different position in protein sequence as shown in Table I. Molecular dynamics used for optimization of model of protein and we get final optimized structure of CYP26A1 retinoic acid metabolizing protein after taken time 5000ps (5ns) number of steps integrator and at temperature 300k. 3DLigandSite was developed to automate our manual approach for predicting ligand-binding sites used in CASP8 . We have demonstrated that 3DLigandSite is able to obtain performance comparable to ours in CASP8 and that this performance is also retained for a much larger test set.

6. REFRENCES:

[1] Ashla AA, Hoshikawa Y, TsuchiyaH,Hashiguchi K, Enjoji M, et al. 2010. Genetic analysis of expression profile involved in retinoid metabolism in non-alcoholic fatty liver disease. *Hepatol. Res*. 40:594–604

[2] Bastien J, Rochette-Egly C. 2004. Nuclear retinoid receptors and the transcription of retinoid-target genes. *Gene* 328:1–16

[3] de Th´e H, Chen Z. 2010. Acute promyelocytic leukaemia: novel insights into the mechanisms of cure.*Nat. Rev. Cancer* 10:775–83 18. DeLuca HF, Roberts AB. 1969. Pathways of retinoic acid and retinol metabolism. *Am. J. Clin. Nutr*. 22:945–52

[4] Fisher GJ, Voorhees JJ. 1996. Molecular mechanisms of retinoid actions in skin. *FASEB J.* 10:1002–13

[5] Gaemers IC, van Pelt AM, van der Saag PT, de Rooij DG. 1996. All-trans-4-oxo-retinoic acid: a potent inducer of in vivo proliferation of growth-arrested A spermatogonia in the vitamin A-deficient mouse testis. *Endocrinology* 137:479–85

[6] Pijnappel WW, Hendriks HF, Folkers GE, van den Brink CE, Dekker EJ, et al. 1993. The retinoid ligand 4-oxo-retinoic acid is a highly active modulator of positional specification. *Nature* 366:340–44

[7] Roberts AB, DeLuca HF. 1967. Pathways of retinol and retinoic acid metabolism in the rat. *Biochem. J.* 102:600–5

[8] Ting W. 2010. Tretinoin for the treatment of photo damaged skin. *Cutis* 86:47–52

[9] Wei LN. 2003. Retinoid receptors and their coregulators. *Annu. Rev. Pharmacol. Toxicol.* 43:47–72

[10] Ahmad N, Mukhtar H. Cytochrome P450: A target for drug development for skin diseases. J Investig Dermatol2004;123:417–425.

[11] Brecher AR, Orlow SJ. Oral retinoid therapy for dermatologic conditions in children and adolescents. J Am Acad Dermatol 2003;49:171–182.

[12] Weiss GR, Liu PY, Alberts DS, Peng YM, Fisher E, Xu MJ,Scudder SA, Baker LH, Moore DF, Lippman SM. 13-cis-Retinoic acid or all-trans-retinoic acid plus inteferon-alpha in recurrent cervical cancer: A Southwest Oncology Group phase II randomized trial. Gynecol Oncol 1998;71:386–390.

[13] Pettersson F, Colston KW, Dalgleish AG. Retinoic acid enhances the cytotoxic effects of gemcitabine and cisplatin in pancreatic adenocarcinoma cells. Pancreas 2001;23:273–279.

[14] Schoch GA, Yano JK, Wester MR, Griffin KJ, Stout CD,Johnson EF. Structure of human microsomal cytochrome P450 2C8 Evidence for a peripheral fatty acid binding site. J Biol Chem 2004; 279:9497–9503.

[15] Williams PA, Cosme J, Ward A, Angove HC, Matak-Vinkovic D, Jhoti H. Crystal structure of human cytochrome P450 2C9 with bound warfarin. Nature 2003;424:464–468.

[16] Yano JK, Wester MR, Schoch GA, Griffin KJ, Stout D,Johnson EF. The structure of human microsomal cytochrome P450 3A4 determined by X-ray crystallography to 2.05-A° resolution. J Biol Chem 2004;279:38091–38094.

[17] The ExPASy (Expert Protein Analysis System)., proteomics server of the Swiss Institute of Bioinformatics (SIB) http://ca.expasy.org

[18] R CSB Protein Data Bank (PDB)., http://www.rcsb.org/pdb

[19] Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure.Science 1991;253:164–170.

[20] Colovos C, Yeates TO. Verification of protein structures:Patterns of nonbonded atomic interactions. Protein Science 1993;2:1511–1519.

[21] White JA, Beckett-Jones B, Guo Y-D, Dilworth FJ, Bonasoro J,Jones G, Petkovich M. cDNA cloning of human retinoic acid-metabolizing enzyme (hP450RAI) identifies a novel family of cytochromes P450. J Biol Chem 1997;272:18538–18541.

[22] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z,Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST:A new generation of protein database search programs.Nucleic Acids Res 1997;25:3389–3402.

[23] Lovell SS, Davis IW, Arendall III WB, e Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Ca geometry: F, C and Cb deviation. Proteins:

[24] Structure Function & Genetics 2002;50:437–450.

[25] Ray, W. J.; Bain, G.; Yao, M.; Gottlieb, D. I. J. Biol.Chem. 1997, 272, 18702–18708.

[26] White, J. A.; Beckett, B.; Scherer, S. W.; Herbrick, J. A.Genomics 1998, 48, 270–272.

[27] Njar, V. C. O. Mini-Rev. Med. Chem. 2002, 2, 261–269.

[28] Gherardini,P.F. and Helmer-Citterich,M. (2008) Structure-based function prediction: approaches and applications. Brief. Funct. Genomic Proteomic, 7, 291–302.

[29] Berezin,C., Glaser,F., Rosenberg,J., Paz,I., Pupko,T., Fariselli,P.Casadio,R. and Ben-Tal,N. (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics, 20, 1322–1324.

[30] Fischer,J.D., Mayer,C.E. and Soding,J. (2008) Prediction of Protein Functional Residues from Sequence by Probability Density Estimation. Bioinformatics, 24, 613–620.

[31] Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol., 257, 342–358.

[32] Aloy,P., Querol,E., Aviles,F.X. and Sternberg,M.J. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function

from homology in genome annotation and to protein docking. J. Mol. Biol., 311, 395–408.

[33] Capra,J.A., Laskowski,R.A., Thornton,J.M., Singh,M. and Funkhouser,T.A. (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput. Biol., 5, e1000585.

[34] Kelley,L.A. and Sternberg,M.J. (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protoc., 4, 363–371.

[35] Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci., 11, 2606–2621.

[36] Capra,J.A. and Singh,M. **(2008)** Characterization and prediction of residues determining protein functional. Bioinformatics, 24,1473–1480.

[37] Marill, J.; Idres, N.; Capron, C. C.; Nguyen, E.; Chabot,G. G. Curr. Drug Metab. 2003, 4, 1–10.

[38] http://www.salilab.org/modeller/tutorial/advanced.html

[39] http://www.gromacs.org .

[40] Ray WJ, Bain G, Yao M, Gottlieb DI. CYP26 a novel mammalian cytochrome P450 is induced by retinoic acid and defines a new family. J Biol Chem 1997;272:18702–18708.

[41] Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thornton, J. M. (1987) Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326,** 347–352.

[42] Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T., and Weng, J. (1987) Protein data bank, in *Crystallographic Databases — Information, Content, Software Systems,Scientific Applications* (Allen, F. H., Bergerhoff, G., and Sievers, R., eds.), Data Commission of the International Union of Crystallography, Cambridge, pp. 107–132.

[43] Doolittle, R. F. (1990) Searching through sequence databases. *Methods Enzymol*.**183,** 99–110.

[44] Pearson, W. R. (1996) Effective protein sequence comparison. *Methods Enzymol*.**266,** 227–258.

[45] Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **84,** 4355–4358.

[46] Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39,** 783–791.

[47] Greer, J. (1981) Comparative model-building of the mammalian serine proteases. *J. Mol. Biol.* **153,** 1027–1042.

[48] Levitt, M. (1992) Accurate modeling of protein conformation by automatic segment matching.*J. Mol. Biol*. **226,** 507–533.

[49] Havel, T. F. and Snow, M. E. (1991) A new method for building protein conformations from sequence alignments with homologues of known structure. *J. Mol. Biol*. **217,** 1–7.

[50] Sippl, M. J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins* **17,** 355–362.