IJPSR (2022), Volume 13, Issue 3 **(Research Article)**

INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES AND RESEARCH

Received on 26 May 2021; received in revised form, 14 July 2021; accepted, 17 July 2021; published 01 March 2022

PREDICTION OF HOTSPOT IN PROTEIN-PROTEIN/PROTEIN-SUBSTRATE INTERACTION: A NOVEL COMPUTATIONAL APPROACH

Kiran T. Raj [1, 2], Ankita Singh [1], Naveen Kulkarni [1] and T. S. Gopenath [* 2]

Quantumzyme LLP [1], Krishnappa Layout, Lalbagh Road, Bangalore - 560027, Karnataka, India.
Department of Biotechnology & Bioinformatics [2], Faculty of Life Sciences, JSS Academy of Higher Education & Research, Mysuru - 570015, Karnataka, India.

Keywords:

Hotspot, Protein-Protein Interaction, Sensitivity, Precision

Correspondence to Author:
Dr. T. S. Gopenath

Department of Biotechnology & Bioinformatics, Faculty of Life Sciences, JSS Academy of Higher Education & Research, Mysuru - 570015, Karnataka, India

E-mail: tgopenath@yahoo.com

ABSTRACT: Protein-protein, as well as protein-substrate interactions, play an important role in regulating specific biological functions like signal transduction, apoptosis, gene regulation, and immune response. These play a major role in drug discovery applications and can be used as molecular targets to stop disease development at the molecular level. Thus, understanding the biological mechanism of protein is an integral part of any computational study. Moreover, these interactions can be manipulated through protein engineering to address industrial applications. As part of the protein engineering strategy, the first step is the determination of hotspots. Hotspots are defined as positions in the amino acid sequence that can be targeted for mutagenesis to improve the catalytic activity or stability of an enzyme. The commonly-used method for determining hotspot residues is computational alanine scanning (CAS) mutagenesis experiments, which are computationally expensive and time-consuming. In the current study, we aim to develop a protocol to predict hotspots using a computationally less costly method. We have used two published datasets to cross-validate our method. Both the datasets belong to a different enzyme class. The key element here is that the substrate is present in the active site of the enzyme. The nearby residues present within a distance of 5 Å from the substrate were also considered during prediction. This helped in determining the exact residues responsible for the majority of performance on protein's characteristic frequency. The objective was to predict hotspots with more precision and sensitivity while demanding limited computing resources.

INTRODUCTION: Proteins are biomolecules that play an important role in cellular metabolism. They are comprised of amino acids.

QUICK RESPONSE CODE

DOI: 10.13040/IJPSR.0975-8232.13(3).1108-19

This article can be accessed online on www.ijpsr.com

DOI link: http://dx.doi.org/10.13040/IJPSR.0975-8232.13(3).1108-19

Based on the sequence of amino acids, proteins differ from one another. This usually results in protein folding into a 3-dimensional structure. The 3D structure of the protein determines its activity. Signal transduction, apoptosis, gene regulation, and immune response are biological functions mediated by protein-protein or protein-substrate interaction Proteins [1]. It acts as a biological catalyst and carries out many chemical reactions by lowering its activation energy and increasing its reaction rate. With the advent of green chemistry, enzymes have

been increasingly used as a biocatalyst in the pharmaceutical as well as different industries because of their various advantages. These include high stereoselectivity, mild reaction conditions, and environment-friendly conditions. No toxic waste is generated, which may be hazardous to human life or the environment. However, natural enzymes do not necessarily fulfill the process requirements and need further industrial-scale production refinement.

It is imperative that the natural enzyme may have some inadequacies such as substrate/product inhibition, low stability, or low activity, limiting enzymes in industries. Enzyme engineering is thus used to tailor the enzymes for specific reactions. Enzyme engineering has gained a lot of attention of late, and various protein engineering methods have emerged for the de novo designing of peptides and proteins following both experimental and computer-guided approaches. Protein engineering is an emerging technique used to design new and improved protein structures using site-directed mutagenesis or random mutagenesis. Some of the applications of protein engineering can be seen in industries like food processing, brewing, detergent, textile, cosmetics, leather, pharmaceutical, biotechnology, *etc.*

For example, in the pharmaceutical industries, enzymes are increasingly being used to manufacture active pharmaceutical ingredients to meet the market demand due to the wave of biocatalysts. The blockbuster drug sitagliptin is one such example in which protein engineering was attempted to increase the efficiency of enzymes and reduce the cost of the manufacturing process. Successful enzyme engineering helped in decreasing the number of chemical steps required for the synthesis of sitagliptin drugs. Another example is that of cellulases which are of interest in biofuel production. These are engineered for improved thermostability to scale up with industrial reaction conditions. In-plant biotechnology, transgenic plants are developed with the help of protein engineering to improve the yield. As enzymes are used in almost all these industries, they are engineered for improved catalytic activity, stability, and solubility. In recent years numerous protein engineering tools have emerged for the improvement of existing biocatalysts or their adaptation to novel substrates [2]. Conventional protein engineering techniques can be grouped into four categories: 3 (i) comparison of the protein sequence with a less homologous protein and mutation of selected amino-acid using site-directed mutagenesis. (ii) site-directed mutagenesis in which mutation is introduced at a specific position of protein sequence after studying the 3d structure of the protein. (iii) random mutagenesis in which mutations are introduced randomly in genes of an organism and (iv) SCHEMA, a structure-guided approach with recombination of stabilizing fragments. Random mutagenesis and SCHEMA are based on the expression of enzymes in microorganisms and high-throughput screening methods.

However, site-directed mutagenesis requires detailed knowledge of protein, and it is difficult to predict the effect of various mutations. On the other hand, random mutagenesis is time-consuming because a lot of experiments have to be performed to obtain the desired mutation. Another caveat of random mutagenesis is that any unwanted mutation can lead to undesirable functions in the protein. To mitigate these challenges, a new method of screening evolved, which is based on computational chemistry. The conventional methods are expensive compared to the computational method of protein engineering; computational methods are fast and accurate compared to the conventional methods.

The computational approach comprises techniques like molecular dynamics, quantum mechanics (Schrödinger equation), molecular mechanics (Newton's law), and statistical methods (Quantitative Structure-Activity Relationship). For example, with the help of computational chemistry, we can perform "*in-silico*" trials in place of experiments that are too expensive to perform in laboratories.

Computational chemistry is a robust technique used in protein engineering, drug design, protein 3D structure prediction, and modelling of transition state to understand complex reaction mechanisms. The computational protein design approach utilizes molecular modelling to understand the structure-function relationship of a given protein sequence. Mutations are carried out on a high-resolution crystal structure to optimize the physicochemical

properties of an enzyme, such as stability or activity. In protein engineering experiments, computational chemistry is mainly used to predict hotspots and engineer enzymes. Hotspots4 are defined as non-essential amino acid residues targeted for mutagenesis to improve an enzyme's catalytic activity or stability. Computer-driven strategies can screen an astronomically large number of sequences covering a wide variety of properties and functionalities compared to any existing experimental approach and thus, can be used for the creation of focused libraries for experimental validations. However, *in-silico* protein engineering efforts are conceptually intricate, difficult to grasp, and rely heavily on user's expertise to assimilate a myriad of factors that together influence the stability and uniqueness of a protein structure. Moreover, the steps involved in successfully executing various computational methods and tools to address specific protein engineering problems are tedious, cumbersome, and not free from human errors. To address these challenges, we have come up with list of methods. In this study, an attempt has been made to develop a robust protocol that is reliable, inexpensive, and requires less computational time to predict accurate hotspots for enzyme engineering.

The methods used here are evaluated based on some evaluation measures to show the efficacy of our model. If any method has to say best, it should recover the true hotspot from a number of mutations attempted. The percentage recovery was calculated based on the ratio of a number of trials to the number of sample datasets. The number of trials was determined using the ratio of True positive to the total number of mutations. Sampling was calculated based on the total score to the total number of residues present in the complex multiplied to percentage [5]. The percentage recovery should be above the random prediction random line. The recovery was measured using the formula given below based on which methods are evaluated

Recovery % = (Trial/Sampling) *100, Where, Trial = True Positive/ Total No of mutations, Sampling = [Top score (number of trials)/Total number residues] * 100

The Percentage recovery plot predicts the reliability of the method. To evaluate the accuracy of hotspot prediction methods, we adopted three evaluation measures to show the efficacy of our model. These include sensitivity (Sen), precision (Prec), and F-measure F1.

$$Sen = TP/(TP + FN),\ Prec = TP/(TP + FP),\ F1 = 2 \times Prec \times Sen / (Prec + Sen)$$

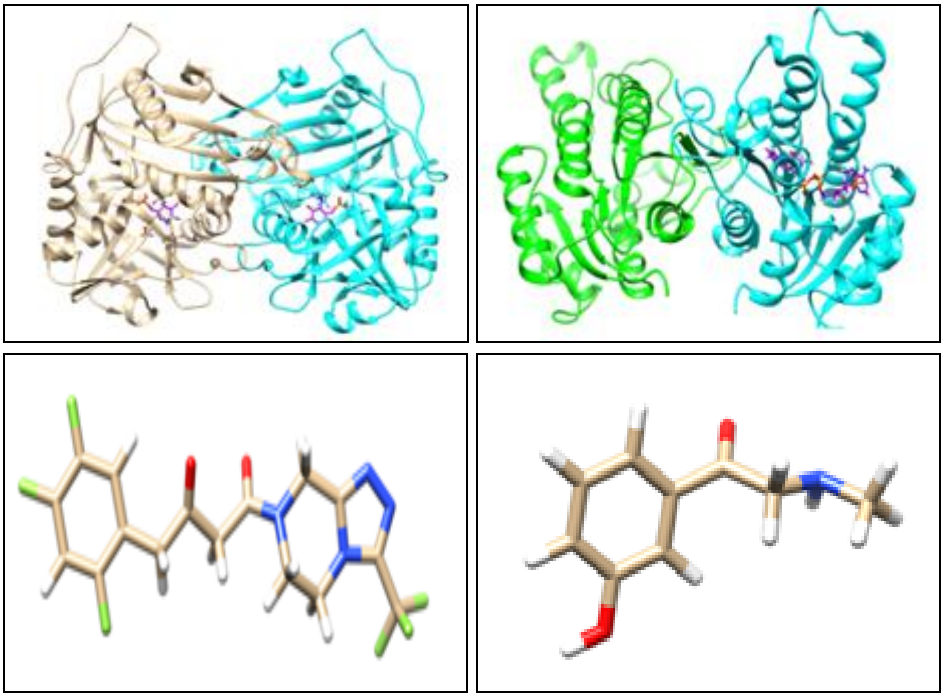

FIG. 1: MODEL SYSTEMS USED FOR THE STUDY. A SHOWS CRYSTAL STRUCTURE OF TRANSAMINASE ENZYME FROM ARTHROBACTER SP. (PDB ID: 3WWI). B SHOWS THE SUBSTRATE PRO-SITAGLIPTIN. C SHOWS THE KETOREDUCTASE ENZYME FROM LACTOBACILLUS KEFIR (PDB ID: 4RF2). D SHOWS THE SUBSTRATE 1-(3-HYDROXYPHENYL)-2-(METHYLAMINO) ETHENONE(HPMAE) USED FOR THE COMPUTATIONAL STUDY

Where, TP (true positive) is the number of correctly-predicted hotspot residues; FP (false positive) is the number of false positives (incorrectly over-predicted non-hotspot residues); TN (true negative) is the number of correctly-predicted non-hotspot residues; and FN (false negative) is false negative, *i.e.*, incorrectly under-predicted hotspot residues. The fraction of true hot spots among the set of residues predicted to be hot spots is called Precision (P).

Sensitivity (S) is the fraction of correctly identified hot spots relative to all those present in the data set. F1 score is a weighted average of the precision and sensitivity [6, 7, 8]. As most of the computational approaches focus on improvement in catalytic activity, our protein engineering study mainly focuses on the progress of the enzymatic activity. Two published datasets have been selected for the study. The validity of the method was checked across two enzyme classes; Transaminase (PDB ID: 3WWI) and the Keto reductase enzyme family (PDB ID: 4RF2), as shown in **Fig. 1.**

Materials and Methods:

System Preparation:

Protein Structures, Cofactors and Substrates: X-ray crystal structure of 3WWI and 4RF2 were obtained from the Protein Data Bank (PDB) [9]. The substrates used for the study are pro-sitagliptin and 1-(3-hydroxyphenyl)-2-(methylamino) ethenone (HPMAE), respectively. As transaminase is a Pyridoxal-5'phosphate-dependent enzyme (PLP enzymes), they contain a cofactor PLP in both chains. On the other hand, ketoreductase is a Nicotinamide-dependent adenine dinucleotide phosphate (NADPH) enzyme using NADPH as a hydride reductant. KRED is active in tetrameric form, whereas transaminase is active as a dimer. Sequences were obtained after using Basic Local Alignment Search Tool (BLAST) [10]. to obtain a series of homologous sequences for both the enzymes (Ketoreductase and Transaminase). Multiple sequence alignment was performed using Jalview software [11].

Datasets: The dataset consists of enzyme-complexes whose structures have been resolved through X-ray crystallography. Structures were obtained from the Protein Data Bank (PDB) [1]. The substrates used for the study are pro-sitagliptin and 1- (3-hydroxyphenyl)-2-(methylamino) ethenone, respectively. As transaminase is a Pyridoxal-5'phosphate-dependent enzyme (PLP enzymes), they contain a cofactor PLP in both chains. Enzyme activity data for transaminase was collected from previous works of Savile *et al*. In total 32 mutations are reported for transaminase and tabulated in **Table 1.** The enzyme is active in dimer form.

TABLE 1: TRANSAMINASE HOTSPOT POSITION LIST WITH MUTATED AMINO ACIDS

S. no.	Position	Mutation
1	S8	P
2	Y26	Y
3	Y60	F
4	L61	Y
5	H62	T
6	V65	A
7	V69	T
8	D81	G
9	M94	I
10	I96	L
11	F122	M
12	S124	T
13	S126	T
14	G136	F
15	E137	E
16	Y150	S
17	V152	C
18	A169	L
19	T178	T
20	V199	I
21	A209	L
22	G215	C
23	G217	N
24	S223	P
25	L269	P
26	L273	Y
27	T282	S
28	A284	G
29	P297	S
30	I306	V
31	S321	P
32	Q329	Q

On the other hand, ketoreductase is a Nicotinamide-dependent adenine dinucleotide phosphate (NADPH) enzyme using NADPH as a hydride reductant. Enzyme activity data for ketoreductase was obtained from Codexis Patent [3, 12].

In total 19 mutations are reported in the patent for ketoreductase and tabulated in **Table 2.** It is active in tetrameric form. The accuracy of predictions was tested for single point mutations for both the systems.

TABLE 2: KRED HOTSPOT POSITION LIST WITH MUTATED AMINO ACIDS

S. no.	Position	Mutation
1	T2	S
2	I11	L
3	A64	V
4	T76	I
5	V95	M
6	S96	L
7	V99	L
8	E145	A,
9	F147	L
10	V148	I
11	T152	A
12	L153	M
13	S159	T
14	Y190	C,G
15	D197	A
16	A202	F
17	E200	P
18	M206	C
19	Y249	F

Docking: In order to understand the enzyme-substrate interaction, substrate pro-sitagliptin was docked into the active site of the Transaminase protein (in place of PLP). The distance between N2 of PLP & C8 of the substrate is 2.83 Å. For KRED protein, the substrate is docked in the active site of protein with the presence of cofactor NADPH, and the distance of between C4N of NADPH and C=O of the substrate is 3.2Å. Docking is performed by Auto Dock 4.0 program 13 using the empirical free energy function and the Lamarckian Genetic algorithm [14].

For the ligand, Gasteiger partial charges are used and the non-polar hydrogens are conjoined. The grid map is calculated using Auto Grid and the grid box dimension was set to 50* 48 *76 for transaminase protein and 100* 86* 88 for KRED. Out of 50 docked poses, the best pose was selected based on binding energy and distance between cofactors and the substrates. After docking, the best- conformation of the substrate was considered for the next step of the molecular dynamics study.

MD Simulation: Parameters for substrates and cofactors were generated using Amber Tools [4]. Partial RESP15 charges were generated using HF/6-31G*16 methods using GAMESS-US software17. Molecular dynamics simulations were performed using GROMACS-2019.1 version [8, 9] (Abraham *et al.*, 2015); (Van Der Spoel *et al.*, 2005) using Amber99SB force field 20 and TIP3P water model [21]. The protonation state of the enzyme was predicted based on PDB2PQR server 22 at pH 7 and temperature 300K. The coordinate, restraints and topology files for ligand, protein, and protein-ligand complex were generated from the PQR file. The enzyme complex coordinates were solvated *in-silico* using the TIP3P water model, using a cubic box of dimension 1.2 with periodic boundary conditions.

Neutralizing ions were added to obtain a net zero charge on the system composed of approximately seventy thousand atoms. First, Systems were energy minimized. Subsequently, the systems were equilibrated for 500 ps in NVT and 500 ps NPT ensembles. During the equilibration period, positional restraint was applied to cofactor, substrate fatty acids, and protein backbone. Finally, 100 ns NPT simulation (production run) was performed at temperature 300 K and pressure 1 atm without positional restraint for both the systems.

Hotspot Analysis: The sequence of proteins was obtained using BLAST algorithm against the non-redundant database for multiple sequence analysis. Using the Jalview11 tool, the consensus sequence was checked for each amino acid, and based on consensus percentage hotspot residue was selected (less conserved amino acids have less consensus percentage). After sequence-based hotspot prediction, structure-based hotspot analysis was done to the protein substrate complex with cofactors in the proteins.

When the substrate is in the active site, it will make contacts with the nearby residues. All kinds of direct interactions like polar and non-polar, favourable and unfavourable contracts were assumed to be a hotspot. The contact analysis was done with the help of Chimera [23].

After structure analysis energy-based method was used for hotspot analysis. For energy analysis gmm pbsa tool 24 was used to calculate per residue binding energy. In the MM-PBSA approach, calculation of the binding free energy (ΔGbind) between a protein and a ligand can be performed as:

$$\Delta G_{bind} = \Delta H - T\Delta S \approx \Delta E_{MM} + \Delta G_{sol} - T\Delta S,\ \Delta E_{MM} = \Delta E_{internal} + \Delta E_{electrostatic} - \Delta E_{vdW},\ \Delta G_{solv} = \Delta G_{elsolv} + \Delta G_{vdWsolv}$$

Where, the total gas phase energy (sum of the Δ Einternal + Δ Eelectrostatic + Δ EvdW) on the binding of MM energy is shown as ΔEMM, the free energy of solvation as Δ Gsolv and the entropy contribution as TΔS. Electrostatic solvation energy term is computed in a continuum solvent using Poisson-Boltzmann model [25]. If the binding energy value is negative, it favours complex formation in water; positive value denotes unfavourable binding. Based on this hypothesis, hotspot residues were selected, which destabilized complex formation.

The cross-correlation between each residue was considered for predicting how each residue or group residue are interconnected for the movement of the protein. If any one of the residues is mutated, then corresponding residues that are interconnected will impact the movement of the protein. The cross-correlation between each residue was considered for predicting multiple hotspots and calculated by using Wordom tool [13]. Wordom calculates the correlations of atomic displacements along MD trajectory. It implements two different algorithms; an algorithm called dynamic cross-correlation (DCC), a well-established and straightforward method calculatinglation of the normalized covariance of atom/residue position, was used. DCC represents the extent of atom/residue displacement correlation within a range that goes from 1.0 to -1.0; where 1.0 indicates completely correlated (same period and phase) and -1.0 relates completely anti-correlated (same period and opposite phase) displacements. Linear mutual information (LMI) is a second algorithm, and it is computationally more expensive compared to DCC.

The performance of traditional hotspot prediction methods is based on precision, specificity and sensitivity. Precision is defined as the measure of consistent results obtained after a number of experiments. Specificity is defined as the measure of correctly identifying false positives. Sensitivity, on the other hand, is defined as the ability to correctly predict the maximum number of true positives. Most of the computational methods used till date have low precision and accuracy. Even sensitivity remains fairly low to overcome these challenges; we have developed a novel method of hotspot prediction, which has improved precision and sensitivity.

RESULTS AND DISCUSSION: As mentioned above we have used two data sets to validate our protocol. The first dataset corresponds to ketoreductase and the second data from Codexis patent [1]2. The transaminase data was based on the previous works of Savile *et. al*. 17 Crystal structures of keto reductase from Lactobacillus kefir 27 (pdb id: 4RF2) and transaminase from Arthro bacter sp [28]. (pdb id: 3WWI) were used for predicting hotspots. Both the enzymes have cofactors NADPH and PMP, respectively. The substrate used for the study was HPMAE and prositagliptin, respectively.

As experimental data were available for both the enzymes, we have included these datasets in our study. The KRED structure selected for our study is a dimer structure (500 amino acids) and have total 250 amino acids for monomeric structure, and NADPH as a cofactor for substrate where V95, S96, E145, T152, L153, and Y190 are in the catalytic region; F147, V148, M206, and Y249 are in the path of substrate entry/exit; A64, T76, S159, D197, V99, E200, and A202 are in the surface region of the protein. I11 and A64 are in cofactor interacting residues. T2 residue is not available in the crystal structure as the initial 2 amino acids are missing for the obtained crystal structure.

The Transaminase system has a dimeric structure with 644 amino acids (322 amino acids for monomeric) and PMP as a cofactor to the pro-sitagliptin substrate, and a total of 32 hotspots are reported for transaminase. Y60, L61, H62, V65 are found in the intermediate part of B-chain but not in active site and; Y26, D81, M94, I96 are in the surface region & not in active site; V69, S124, S126, Y150, V152 are in A-chain and near to active site; T178, G215, G217, L269, L273, T282, P297, I306, S321, Q329 far away from active site and & surface region; F122 and A284 are in the small binding pocket and V199, S223 are in the large binding pocket. G136, E137, A169, and A209 are key residues that enable substrate entry/exit. S8 residue is not available in the crystal structure as the initial 8 amino acids are missing for the obtained crystal structure. These mutations are incorporated in the structure to increase the activity towards the respective substrates. A detailed description of the data set, individual mutations, and clustering criteria has been discussed in earlier

studies [12, 29]. Some well-known hotspot prediction methods like sequence-based hotspot prediction, structure-based hotspot prediction, and energy-based hotspot prediction methods have been used to identify mutations of selected data sets. These methods are evaluated based on random prediction, selectivity, and precision. The probability of distributions and a sequence of random variables whose outcomes do not follow a deterministic pattern is called a random process.

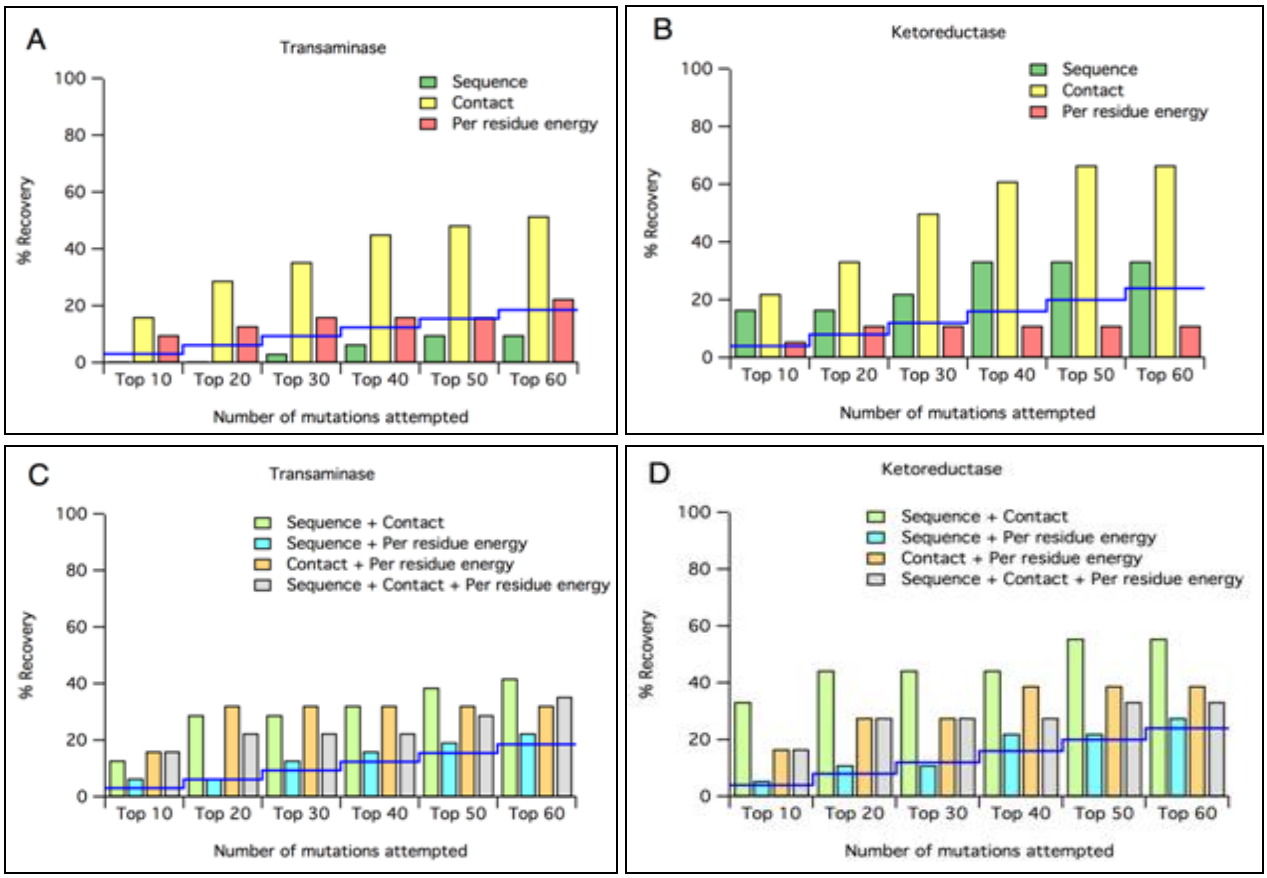

FIG. 2: HOTSPOT PREDICTIONS FOR THE TRANSAMINASE AND THE KETOREDUCTASE USING QZ-WORKBENCH. FIGURES (A) AND (B) SHOW PERCENTAGE (%) RECOVERY OF THE TOTAL TRUE POSITIONS (POSITIONS THAT WERE MUTATED TO IMPROVE ACTIVITY) FOR DIFFERENT NUMBERS OF MUTATIONAL TRIALS (TOP 10, TOP 20, TOP 30, TOP 40, AND TOP 50) BASED ON SEQUENCE SCORE, CONTACT SCORE AS WELL PER RESIDUE CONTRIBUTION IN BINDING ENERGY. FIGURES (C) AND (D) SHOW % RECOVERY OF THE TOTAL MUTATIONAL POSITIONS (POSITIONS THAT WERE MUTATED TO IMPROVE ACTIVITY) FOR DIFFERENT NUMBERS OF MUTATIONAL TRIALS WHILE SCORES ASSOCIATED WITH SEQUENCE, CONTACT AND PER RESIDUE CONTRIBUTION ARE COMBINED. THE SOLID BLUE LINE IN ALL THE FIGURES CORRESPONDS TO THE RANDOM PREDICTION

Sequence-Based Hotspot Prediction: Sequence analysis is the analytical method used to study the physicochemical characteristics, function, or evolution of protein sequences. It helps in understanding the genetic diversity of sequence and evolution of organisms through sequence alignment. In protein engineering studies, these play a vital role in knowing the functional, active site residues, their conservation, and percentage similarity/identity across different homologous sequences. Sequences were obtained and BLAST against a non-redundant database as well as Protein databank; the multiple sequence alignment was obtained for both the data sets. Based on multiple sequence alignment, the hotspots were predicted. The strategy used for selecting hotspots is (i) based on conservation of residues over the phylogeny, selecting less conserved residues as hotspots, (ii) hotspots corresponding to highly mutable residues located in the active site pocket or access tunnels, stability hotspots corresponding to flexible residues, and (iii) hotspots based on correlated residues or network residues. A focused library was created based on naturally accepted substitutions from phylogenetic analysis. Results show that already known mutations from both the datasets

(Transaminase and KRED) have mutations in highly conserved regions. These are away from active site pockets or tunnels. The percentage recovery was calculated based on the ratio of a number of trials to the number of sample datasets. The number of trials was determined using the ratio of True positive to the total number of mutations. It was found that the plot **Fig. 2** for the Transaminase dataset was varying over the number of trials, whereas for KRED it was increasing and looked better for KRED data sets. The sequence analysis method doesn't rely on the 3D structure of the protein. If the structure of protein is not available, the loops, tunnel residues or active site residues will not be known, and the method may not be able to predict accurate hotspots. The recovery plot tells us that the sequence analysis method depends on the strategy and type of data set selected.

Structure-Based Hotspot Prediction: Structure-Based hotspot prediction relies on the 3D structures of proteins where the interaction between two proteins or protein-substrate are calculated, and the amino acids which interact are considered as a hotspot. The interactions can be in the form of biochemical contacts between atom, residue, hydrogen bonds, and salt bridges. All kinds of direct interactions like polar and nonpolar, favourable and unfavourable play a role in hotspot prediction. The overlap between two atoms is defined as the sum of their Vander Waal's radii (VDW) and the difference of the distance between them as well as an allowance for potentially hydrogen-bonded pairs [23]. This can be understood by the expression given below

$$\text{overlap}_{ij} = \text{rVDW}_i + \text{rVDW}_j - d_{ij} - \text{allowance}_{ij}$$

In the structure-based method, the substrate is docked into the active site of protein and simulations are run for the enzyme-substrate complex. The best conformations or snapshots are taken from trajectory based on the active site distances (distances between active site residues to the substate) and checked for the amino acid residues that make close contact with the substrate. This is done using the tool chimera. These residues are considered as mutable residues or hotspots for designing proteins for better activity. From contact score analysis, 17 hotspot residues were recovered and the remaining 14 positions could not be recovered for the transaminase data set, whereas for KRED data set 12 hotspots were recovered among the 19 hotspots mentioned in Codexis patent. The total time taken for the contact score analysis was 3-4 h with CPU and without GPU cards, which means this method is computationally less expensive with more than 50% hotspot recovery. **Fig. 2** shows the percentage recovery plot for both data set; contact score method is reliable, which is above the random prediction line. It indicates that the true positive values increase over a number of trials. The analysis suggests that contact score analysis for the structure-based method is more reliable as it predicts true positive hotspots for the given data sets but fails to recover 100 percent of given data sets.

Energy-Based Hotspot Prediction Per Residue Energy: Binding free energies of all the complexes in the present study were calculated using Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) approach. In order to calculate the binding energy, the water molecules were removed from the system. The binding free energies were estimated using an implicit representation of water by Poisson Boltzmann (PB) approaches. In this method, various conformations or snapshots of the solute were extracted from a molecular dynamics simulation trajectory [24].

Per-residue binding energy analysis was performed in order to obtain a quantitative description of the contribution of each amino acid with the substrates considered. Per-residue binding energy analysis provides insights about the contribution of each residue based on which mutable residues are considered 30. Best snapshots were taken from trajectory based on the active distance. Decomposition of per residue was checked by calculating the binding energy. If the per residue value is negative, then it favours complex formation in water, and if the value is positive except electrically charged amino acids like Arginine, Histidine, and Lysine, it does not stabilize complex formation. The unfavourable residues which destabilize complex formation are considered as a hotspot. From the recovery plot, the method used here was reliable for transaminase data set but unreliable for KRED data set as depicted in **Fig. 2** based on random line prediction.

Further, the method is computationally inexpensive. However, the drawback of this method is that the observed values of the residues might be due to the electrochemical properties of the latter. The sequence method, structure-based method and Per residue energy-based method failed to recover 100 percent of the hotspot residues as these methods have their own limitations. To overcome these limitations, we have come up with a method where one of the above methods was used along with the cross-correlation 31 method or a combination of all methods to obtain 100 percent recovery of hotspots.

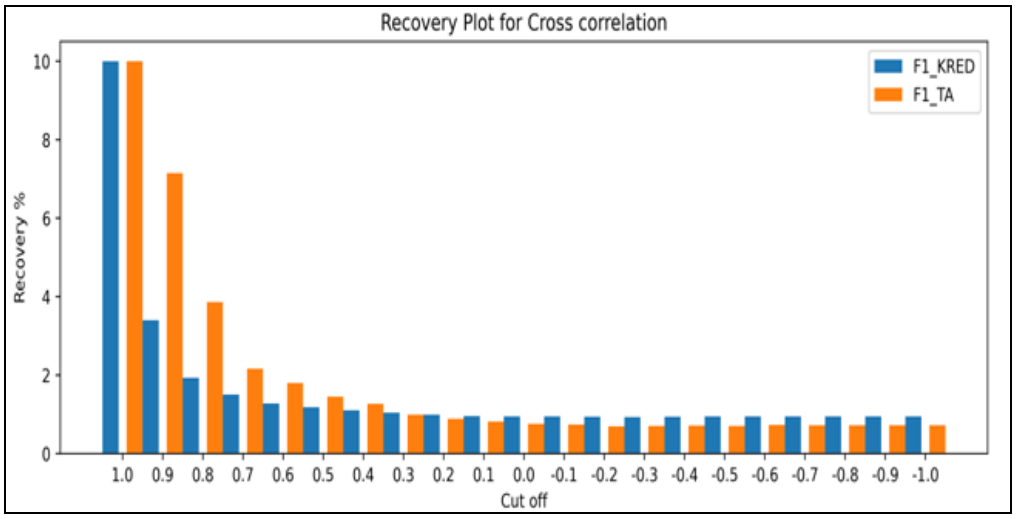

FIG. 3: GRAPH SHOWS THE RECOVERY PERCENTAGE OBTAINED FROM CROSS-CORRELATION CALCULATION. THE BAR GRAPH SHOWS THE VALUE OF CROSS-CORRELATION RANGING FROM 1 TO 0.6, HAVING MORE RECOVERY THAN OTHER VALUES. THE BAR GRAPH IN BLUE REPRESENTS TA DATASET, THE BARS IN ORANGE SHOWS THE DATA FOR KRED DATA SETS

The combination of sequence analysis, contact score and cross-correlation methods are used to recover most of the hotspot for the given data sets and can be seen by the F1 score in **Fig. 4.**

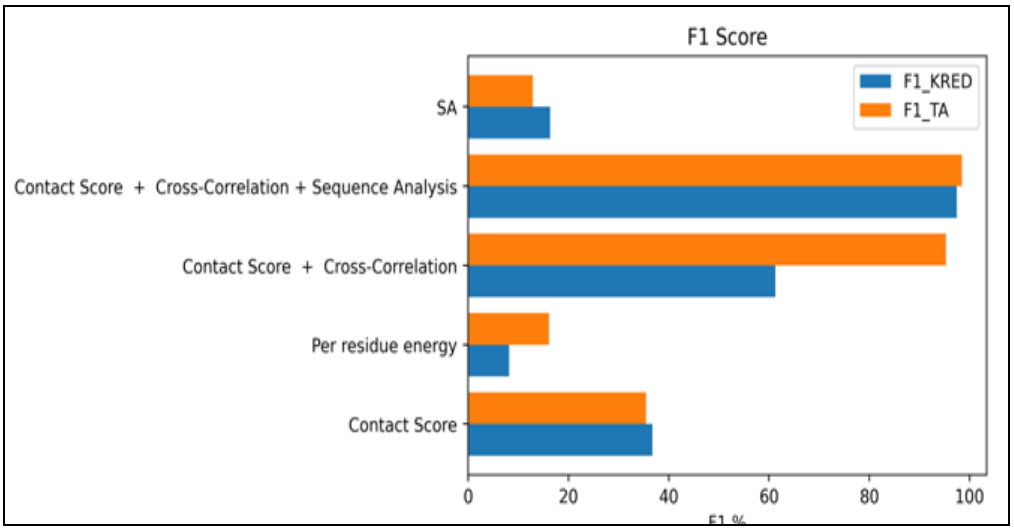

FIG. 4: F1 SCORE PREDICTION OF HOTSPOTS FOR DIFFERENT METHODS IS SHOWN IN THE ABOVE IMAGE. THE CONTACT SCORE, PER RESIDUE, AND CROSS-CORRELATION METHOD, ARE STRUCTURE-BASED METHODS. THE SEQUENCE ANALYSIS (SA) METHOD IS BASED ON CONSERVATION TO IDENTIFY FUNCTIONALLY IMPORTANT RESIDUES. FINALLY, A COMBINATION OF METHODS PREDICTING A GREATER NUMBER OF RESIDUES WITH 90-95% OF F1 SCORE WAS ACHIEVED

For the current analysis, we selected top 20 residues from the contact score as we saw their decrease in recovery of hotspots after top 20 for KRED data, whereas for transaminase, it is 30, so we took the top 20 based on comparison **Fig. 5.**

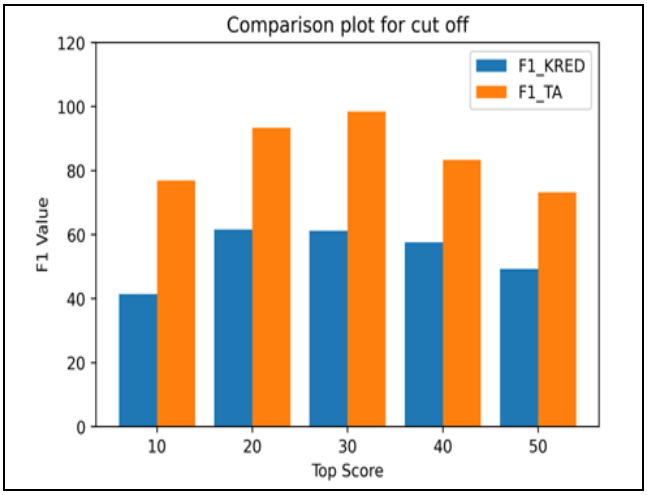

FIG. 5: COMPARISON PLOT FOR CUT-OFF PREDICTION FOR EVALUATING

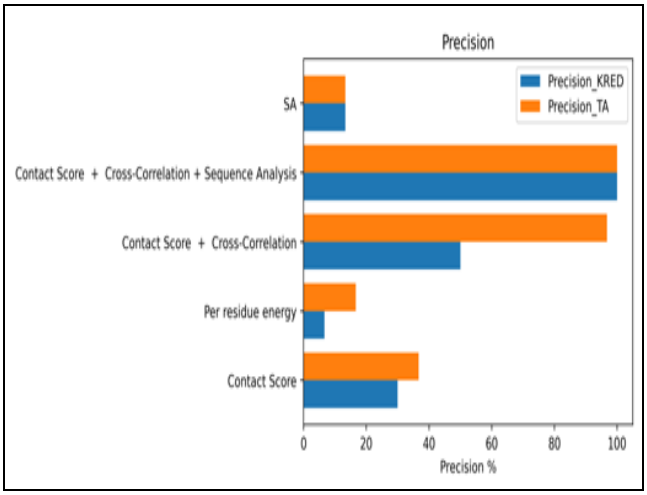

FIG. 6: PRECISION PLOT FOR EVALUATING EACH METHOD

Cross-correlation analysis of the top 10 contact score residues showed some of the residues that were not recovered from the contact score correlate (the Cross-correlation values range from 1.0 to 0.6). This combination method (contact score and cross-correlation) failed to recover most of the hotspot residues of the data set. The top 20 residues from the contact score analysis were selected, and the cross-correlation between these residues were checked.

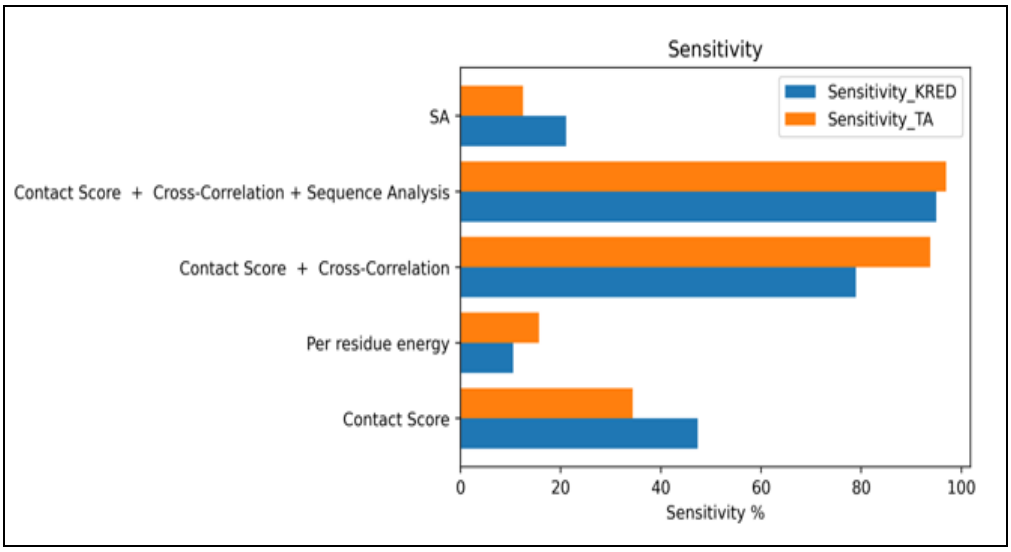

FIG. 7: SENSITIVITY PLOT FOR EVALUATING EACH METHOD

The cross-correlation analysis of top 20 residues from contact score showed that most of the residues are recovered from the given data sets (ranges between 1.0 to 0.6). Moreover, the recovered residues had a very high correlation among them.

If the cut-off value of these residues has increased to 30, it may give greater precision and sensitivity but the trial's prediction may decrease eventually. From this analysis, we found that the combination of contact score with cross-correlation method recovers a large number of hotspots with high precision and sensitivity **Fig. 6 & 7.**

Cross-Correlation (CORR): The extent of the correlation of residue-residue displacements was calculated using correlation algorithms for the simulation trajectory. A dynamics cross-correlation method has been used for a given residue or atom pair by wordom tool. The value can range from -1.0 (completely anti-correlated motion) to + 1.0 (completely correlated motion). The trajectory of the enzyme-substrate complex is chosen, and completely anti-correlated motion to completely correlated motion of residues is calculated. The recovery plot for the cross-correlation analysis showed in **Fig. 3**. As cross-correlation gives an idea

of how each residue is making cross-talking between them for a specific movement of protein, so when you mutate any one point, they have to consider the corresponding residues for mutation based on the type of cross-connected with each other's. The true positive hotspot residues obtained from the previous methods like sequence analysis and structure-based predictor methods like contact analysis and per residue energy are considered and its corresponding high cross-talking residues (the cross-correlation weight age above 0.6) are looked. From this analysis, it was observed that most of the unrecovered hotspot residues in the above methods are interconnected. In combination with the above methods, the cross-correlation method can be used for predicting the unrecovered hotspots for both the data sets. The cross-correlation between the residues decreased from 0 to -1.0 range as shown in **Fig. 3** and the time taken for the CORR calculation and analysis was computationally less expensive

CONCLUSION: The objective of this study was to develop a robust, computationally less expensive protocol and predict hotspots with high accuracy. In lieu of this, we were successfully able to develop a method that can dramatically reduce the number of variants selected for experimental validation. The efficiency of the sequence method, structure-based method, and Per residue energy-based method were analyzed, and it was observed that most of the methods failed if they were used alone. Combining methods helps achieve higher precision and sensitivity for a given set of data and overcomes limitations of the individual methods.

The technique used in the current study has the combination of contact score and cross-correlation method, which will recover most of the hotspots for the given data sets. The current methodology could be used in the future to predict hotspots in multiple data sets with higher precision and sensitivity, which leads to novel engineering strategies to design biocatalysts for specific chemical reactions with desired stereochemistry.

ACKNOWLEDGMENT: The authors would like to acknowledge Quantumzyme LLP, Krishnappa Layout, Lalbagh Road, Bangalore, and Management of JSS Academy of Higher Education & Research, Mysuru, Karnataka for supporting the basic research ideas and for the resources provid.

CONFLICTS OF INTEREST: The authors declare that there are no conflicts of interest.

REFERENCES:

1. Kaushik M: Protein engineering and de novo designing of a biocatalyst. J Mol Recognit 2016; 29: 499-03.
2. Bommarius AS, Blum JK and Abrahamson, MJ: Status of protein engineering for biocatalysts: How to design an industrially useful biocatalyst. Curr Opin Chem Biol 2011; 15: 194-00.
3. Goedegebuur F: Improving the thermal stability of cellobiohydrolase Cel7A from Hypocrea jecorina by directed evolution. J Biol Chem 2017; 292: 17418-30.
4. van der Meer JY, Biewenga L and Poelarends GJ: The Generation and Exploitation of Protein Mutability Landscapes for Enzyme engineering. Chem Bio Chem 2016; 17: 1792-99.
5. Jana, S, Ghosh S, Muk S, Levy B and Vaidehi N: Prediction of conformation speci fi c thermostabilizing mutations for class a g protein-coupled receptors. J Chem Inf Model 2019; 59: 3744-54.
6. Sciences M: Prediction of protein hotspots from whole protein sequences by a random projection ensemble system. 2017.
7. Ofran Y and Rost B: Protein - Protein Interaction Hotspots Carved into Sequences 3: 2007.
8. Lise S, Buchan D, Pontil M and Jones DT: Predictions of Hot Spot Residues at Protein-Protein Interfaces Using Support Vector Machines. 6: 2011.
9. Berman HM: The protein data bank. Acta Crystallogr Sect D Biol Crystallogr 2002; 58: 899-07.
10. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ: Basic local alignment search tool. J Mol Biol 1990; 215: 403-10.
11. Waterhouse AM, Procter JB, Martin DMA, Clamp M and Barton GJ: Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. 2009; 25: 1189-91.
12. Specification E: EP002467473B1 *. 2016; 1: 2322-28.
13. Allouche A: Software news and updates gabedit a graphical user interface for computational chemistry softwares. J Comput Chem 2012; 32: 174-82.
14. Ingersoll DW, Bronstein PM and Bonventre J: Chemical modulation of agonistic display in Betta splendens. J Comp Physiol Psychol 1976; 90: 198-02.
15. Comell WD, Cieplak P, Bayly CI and Kollman PA: Application of RESP charges. J Am Chem Soc 1931; 115: 9620-31.
16. Schmidt MW: General Atomic and Molecular Electronic Structure System 1993; 14: 1347-63.
17. Bode BM and Gordon MS: MacMolPlt A graphical user interface for GAMESS. J Mol Graph Model 16: 1988; 133-38.
18. Abraham MJ: Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. Software 2015; 2: 19-25.
19. Van Der Spoel D: GROMACS: Fast, flexible and free. J Comput Chem 2005; 26: 1701-18.
20. Lindorff-Larsen K: Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins Struct Funct Bioinforma 2020; 78: 1950-58.
21. Mark P and Nilsson L: Structure and dynamics of the tip3p spc and spc. E Water Models 2001; 9954-60.
22. Dolinsky TJ, Nielsen JE, Mc Cammon JA and Baker NA: PDB2PQR: An automated pipeline for the setup of

Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res 2004; 32: 665-67.
23. Chen JE, Huang CC and Ferrin TE: RRD ist Maps: A UCSF Chimera tool for viewing and comparing protein distance maps. Bioinformatics 2015; 31: 1484-86
24. Kumari R, Kumar R, Source O, Discovery D and Lynn A: gmmpbsa A GROMACS Tool for High-Throughput MM-PBSA Calculations 2014.
25. Im W, Beglov D and Roux B: Continuum solvation model: computation of electrostatic forces from numerical solutions to the poisson-boltzmann equation. Comput Phys Commun 1998; 111: 59-75.
26. Nursalam: metode penelitian. Amber 2019-Reference manual. J Chem In Model 2013; 53: 1689-99.
27. Noey, EL: Origins of stereoselectivity in evolved ketoreductases. Proc Natl Acad Sci 2015.
28. Guan LJ: A new target region for changing the substrate specificity of amine transaminases. Sci Rep 2015; 5: 1-8.
29. Savile CK: Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. Science 1990; 329: 305-09.
30. Gohlke H, Kiel C and Case DA: Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. J Mol Biol 2003; 330: 891-13.
31. Helleseth T: On the Cross correlation of m-Sequences and Related Sequences with Ideal Autocorrelation = L. 3: 2002.

How to cite this article:
Kiran TR, Singh A, Kulkarni N and Gopenath TS: Prediction of hotspot in protein-protein/protein-substrate interaction: a novel computational approach. Int J Pharm Sci & Res 2022; 13(3): 1108-19. doi: 10.13040/IJPSR.0975-8232.13(3).1108-19.