International Journal of Biomedical Data Mining

Arora et al.,Int J Biomed Data Min 2016, 5:1
DOI: 10.4172/2090-4924.1000119

Research article Open Access

Shikimate Kinase of *Yersinia pestis*: A Sequence, Structural and Functional Analysis

Neelima Arora[1*], Mangamoori Lakshmi Narasu[1] and Amit Kumar Banerjee[2]

[1]*Centre for Biotechnology, Institute of Science and Technology, Jawaharlal Nehru Technological University, Kukatpally, Hyderabad-500085, Telangana State, India*
[2]*Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology, Tarnaka, Hyderabad-500007, Telangana State, India*

Abstract

Yersinia pestis, the causative organism of Plague, is widely recognized as a potential bioterrorism threat. Due to the absence of homologs in human, Shikimate Kinase (SK) is considered as an excellent drug target in several bacterial and protozoan parasites. Ample literature evidences confirm the suitability of this protein as a good target. Therefore, Shikimate Kinase of Shikimate pathway in *Yersinia pestis* represents an attractive drug target. In the present study, a clustering approach was undertaken to select the proper representative for Shikimate Kinase sequences belonging to *Yersinia pestis* for structure determination. Three-dimensional models of the enzyme for KFB61218.1 (SK1), EFA47400.1 (SK2) and WP_016255950.1 (SK3) were generated using a comparative molecular modeling approach where structures were developed using the single specific template as well as multiple closely associated templates. The structures of Shikimate Kinase developed using comparative modeling were evaluated for stereochemical quality using various structural validation tools. Results from structural assessment tools indicated the reasonably good quality of models.

Keywords: Shikimate Kinase; Plague; *Yersinia pestis*; Molecular modeling; Homology modeling; Bioinformatics

Introduction

No other disease would have shaped the history of mankind as Plague with three major pandemics in past and many sporadic outbreaks thereafter. Such was the fear of Plague that it was termed as "black death", clearly reflecting the panic on its outbreaks that took a heavy toll of lives. Plague affects all the age groups and genders but the most vulnerable group comprises of young people in the age bracket of 12-45 years [1, 2]. Plague is endemic in Africa, Asia, South America, and North America with the majority of cases reported in Africa [3]. It is difficult to assess the global burden of Plague as mortality rate remains a poor reflection of disease state and endemicity due to lack of proper diagnosis and underreporting.

Plague is a zoonotic infection of wild and domestic animals with humans being the incidental host. Plague is transmitted to humans by the bite of rat flea *Xenopsylla cheopis* [4]. *Yersinia pestis*, the causative agent of bubonic, septicemic and pneumonic plague [4], is a gram-negative facultative anaerobic bacterium belonging to family *Enterobacteriaceae* [5]. The fear of the use of *Yersinia pestis* in bioterrorism, the emergence of multi-drug resistant strains, high case-fatality ratio and lack of an effective vaccine against it warrant the need to explore new drug targets and drugs for combating the threat. Shikimate pathway is present in plant, fungi and bacteria but absent in animals and represents a suitable source of drug targets [5-7]. This pathway involves seven steps which are responsible for the synthesis of chorismate [8]. Shikimate Kinase catalyzes the phosphorylation of shikimate to form shikimate 3 phosphate and ADP [9]. Shikimate Kinase is considered as a promising drug target and has been studied extensively along with other important drug targets of pathogens [10-14]. The tertiary structure of a protein determines its function and provides clues about its role in biological processes. Hence, an understanding of the tertiary structure of a protein is crucial to get a clue about its functional aspects. Despite the recent momentum gained in structure determination, the number of protein structures available in repositories still lags far behind the number of available sequences. This situation is leading to a huge gap in protein sequence and structure data. Lack of experimentally determined structure remains a major obstacle in drug designing and development process. Comparative modeling that provides a means for gaining insight into structural details of protein in the absence of experimentally derived structure is being widely used for various important targets [15-19]. In this study, Shikimate Kinase enzyme from 3 different strains of *Yersinia pestis* was selected and characterized *in silico*.

Materials and Methods

Sequence collection and physiochemical characterization

The sequences available in the NCBI protein database were extracted with the search keyword "Shikimate Kinase" combined with "*Yersinia*". The search yielded 259 sequences altogether. After the primary manual screening, 218 sequences were obtained.

Sequence Clustering and representative selection

A sequence identity-based clustering approach was adopted for the selection of a specific representative from the sequences. For this purpose, CD-HIT was employed where sequences with 90% identity were considered under the same cluster along with parametric set up of global sequence identity calculation with a bandwidth of alignment of 20. Clustering analysis of the obtained sequences showed that the shortest and longest sequences in the batch were of 167 and 214 amino acid length respectively. The observed average length of the sequences

***Corresponding author:** Dr. Neelima Arora, Centre for Biotechnology, Institute of Science and Technology, Jawaharlal Nehru Technological University, Kukatpally, Hyderabad-500085, Telangana State, India, Tel: +91-040 2315-8661; E-mail: aroraneelima123@yahoo.co.in

Received: January 29, 2016; **Accepted:** February 22, 2016; **Published:** March 15, 2016

Citation: Arora N, Narasu ML, Banerjee AK (2016) Shikimate Kinase of Yersinia pestis: A Sequence, Structural and Functional Analysis. Int J Biomed Data Min 5: 101. doi:10.4172/2090-4924.1000119

was 184 bases. The observed length distribution of the sequences is presented in Table 1.

The first cluster contained 109 sequences, whereas the second and the third comprised of 108 and 1 sequence respectively. The obtained representatives for Cluster 0, 1 and 2 comprised of 214 (gi|510938210|ref|WP_016255950.1|), 173 (gi|668665097|gb|KFB61218.1|) and 167 (gi|270336623|gb|EFA47400.1|) amino acids respectively. In this article, sequences with accession number (gi|668665097|gb|KFB61218.1), (gi|270336623|gb|EFA47400.1) and (gi|510938210|ref|WP_016255950.1) will be represented by SK1, SK2 and SK3 respectively. Obtained sequences of Shikimate Kinase were characterized *in silico* using Expasy-ProtParam tool and ProtScale [20].

Functional characterization of SK of *Yersinia pestis*

Motifs were searched using Multiple Em for Motif Elicitation available in MEME suite [21] employing default parameters. Knowledge of motifs occurring in a sequence provides a clue about its functional role in biological processes. Conservation of residues was calculated using ConSurf [22,23], a web server for the identification of biologically important residues in protein sequences.

Secondary structure prediction

Secondary structure of this protein was predicted employing NPS server [24] using different methods, viz. Double prediction method (DPM) [25], Discrimination of protein secondary structure class (DSC) [26], GOR4 [27], Hierarchical neural network (HNN) [28], PHD [20], Predator [30], SIMPA96 [31], Self-optimized prediction method with alignment (SOPMA) [32]and Sec. Cons [26] with default parameters.

Prediction of intrinsic disorder

In order to identify regions of higher flexibility, DisEMBL [33], Globplot [34], Regional order neural network (RONN) [35] and Protein disorder prediction system (PRDOS) [36] were employed to select complementary regions for the identification of consistent problematic regions in the protein.

Homology modeling of Shikimate Kinase of *Yersinia pestis*

Since the three-dimensional structure of the protein was not available in Protein Data Bank (PDB), the present task of developing the 3D model of Shikimate Kinase of *Yersinia pestis* was undertaken.

Template Selection and analysis: The template search was performed using the Modeller environment against 11079 protein structures to arrive at the best template depending on the sequence and structural profile with relation to the target sequences considered. Each target was subjected to template search individually and results were analyzed. In each case, four structures (1E6C from *Erwinia chrysanthemi*, 1KAG from *Escherichia coli*, 1L4U from *Mycobacterium tuberculosis* and 1VIA from *Campylobacter jejuni*) [37-40] were found as the most suitable templates for the target sequences Figure 5 and Figure 6. In this study, we aimed to gain knowledge about structural characteristics of the Shikimate Kinase protein of *Y.pestis* applying comparative modeling techniques. Molecular modeling is often used to gain insight into the structure in the absence of experimentally determined crystal structure. The methodology involved following steps: template selection, sequence alignment, model generation followed by refinement and model evaluation.

Sequence Length	Cluster 0(Number of sequences)	Cluster 1(Number of sequences)	Cluster 2(Number of sequences)
174	104		
214	5		
161		3	
167		1	1
173		104	

Table 1: Distribution of the sequences forming clusters based on the length.

Target-template alignment: ClustalX was used for obtaining template-target sequence alignment [41]. Setting used was: Scoring matrix =Blosum, Gap penalty =10 and gap extension penalty =0.05. The output was used for consequent generation of the models.

Model generation: The models were generated using MODELLER9v3 that implements an automatic comparative modeling approach to construct a refined three-dimensional model of a protein based on a given sequence alignment and selected template [42]. MODELLER employs probability density functions (PDFs) derived analytically using statistical mechanics and empirically using a database of known protein structures as the spatial restraints rather than energy. MODELLER uses template coordinates for deriving the spatial constraints and amalgamates the energy terms to compute the proper stereochemistry of a protein and express them through objective function as a quantitative value. In the later stage, the tool optimizes the Cartesian space using a conjugate gradient (CG) algorithm along with molecular dynamics calculations [42]. Objective function in Modeller is derived using the following formula:

$$F = F(R) = F_{symm} + \sum_i c_i(f_i, p_i)$$

where F denotes the objective function calculated with respect to Cartesian coordinates of ~10,000 atoms (3D points) that form a system containing one or more molecules, F_{symm} denotes optional symmetry term, R denotes Cartesian coordinates of all atoms, c is a restraint; i, f denote geometric features of a molecule and p denotes parameters which vary from restraint to restraint.

Structure validation: The stereochemical quality of the obtained structures were evaluated using SAVES server employing PROCHECK [43], WHATCHECK [44,45], VERIFY 3D [46] and ERRAT [47] program. PROSA was also used to evaluate the stability of generated structures [48]. These models were further subjected to identification of active sites.

Active site identification: CASTp (Computed Atlas of Surface Topology of Proteins) [49] was used for identifying and characterizing active sites, binding sites and amino acids constituting the binding site of a protein by measuring concave surface regions of three-dimensional structures of proteins.

Results and Discussion

Recent bioinformatics tools allow us to have a thorough insight about a protein from sequence to the structure with a reasonable amount of accuracy and within a stipulated time limit. The technology is progressing towards system biological analysis at a rapid pace yet there is a requirement for understanding important molecules with keen attention and specific investigation. The obtained results in this study are described in the following section in detail along with the importance of the findings.

Amino Acid composition

Amino acid composition and important properties of the enzyme are shown in Table 2-4.

It was observed that Cysteine and histidine were absent in Shikimate Kinase (KFB61218.1) of *Yersinia pestis* but present in SK1 and SK3.

Other physicochemical properties

It was found that Shikimate Kinase enzyme of WP_016255950.1 and EFA47400.1 was stable but Shikimate Kinase of KFB61218.1 was unstable as it showed a value of instability index above 40. Negative GRAVY value for the SK1, SK2 and SK3 indicated the better interaction of the enzyme with water. Theoretical pI of Shikimate Kinase sequences considered in the study indicates their acidic nature. The high value of the aliphatic index of the considered sequences also indicates their stability.

Important Motif determination

As mentioned earlier, motifs were determined using the MEME suite with default parameter settings. Shikimate kinase protein is having some unique motifs in its sequence. Motifs, such as Walker A-motif (GXXGXGKT/S) bridging the β1 (first beta strand) and α1 (first α helix) and allowing the development of signature phosphate binding loop, DT/SD, GGGXV were reported earlier [50]. The observed comparatively conserved sequence stretch considered as a motif is represented in Table 5.

The secondary structure of a protein represents repetitive geometrical conformations formed as a result of intermolecular and intermolecular hydrogen bonding. Prediction of secondary structure helps us in determining whether a given amino acid is a part of a helix, strand or coil. The results from different secondary structure prediction servers used in the analysis revealed that random coils and alpha helices were predominant among secondary structure elements followed by extended strands (Table 6).

A set of SK sequences was collected by using the SK protein sequences as a query in the Consurf server (CSI-BLAST E-value: 0.0001, the maximum number of homologs: 150, CSI-BLAST iteration: 3). The search resulted in 480 unique hits out of total 490 hits for SK1 (gi|668665097|gb|KFB61218.1). There were 495 unique sequences out of 499 hits for SK2 (gi|270336623|gb|EFA47400.1). Out of 296 CSI-BLAST hits for SK3 (gi|510938210|ref|WP_016255950.1), 291 were unique. The calculation was performed on the 150 sequences with the lowest E-value. An unrooted phylogeny was constructed in PHYLODRAW by employing multiple sequence alignment of a set of sequences using MAFT (v3.5.1). Figure 1 depicts the plot representing determined conservation scores versus residue number.

Amino acid	KFB61218.1(SK1)	EFA47400.1(SK2)	WP_016255950.1(SK3)
Ala	6.90	8.40	10.30
Arg	9.20	7.80	6.10
Asn	4.60	3.00	2.30
Asp	5.80	9.60	5.10
Cys	0.00	2.40	0.90
Gln	5.80	4.20	6.10
Glu	12.10	4.80	7.50
Gly	7.50	9.00	7.50
His	0.00	2.40	1.40
Ile	5.80	6.60	3.30
Leu	8.10	8.40	8.90
Lys	5.80	3.60	3.70
Met	2.30	4.20	4.20
Phe	2.90	3.00	3.70
Pro	2.90	4.20	3.70
Ser	4.00	6.60	5.60
Thr	5.80	2.40	7.00
Trp	0.60	0.60	0.50
Tyr	1.20	1.80	1.40
Val	8.70	7.20	10.70

Table 2: Amino acid composition of considered Shikimate Kinase sequences.

Properties	KFB61218.1(SK1)	EFA47400.1(SK2)	WP_016255950.1(SK3)
Number of amino acids	173	167	214
Molecular weight	19532	18361.9	23330.6
Theoretical pI	5.06	5.33	5.16
Total number of negatively charged residues	31	24	27
Total number of negatively charged residues	26	19	21
Ext. coefficient	8480	10220	10095
Instability index	43.77	35.73	32.56
Aliphatic index	86.18	87.6	88.83
Grand average of hydropathicity (GRAVY)	-0.621	-0.232	-0.077

Table 3: Important physicochemical properties of Shikimate Kinase of *Yersinia pestis* calculated using Protparam.

Property	KFB61218.1		EFA47400.1		WP_016255950.1	
	Min	Max	Min	Max	Min	Max
Bulkiness	0.122	0.705	0.39	0.833	0.41	0.75
Polarity(Zimmermann)	0.004	0.756	0.004	0.647	0.01	0.66
Recognition factors	0.123	0.563	0.123	0.456	0.16	0.5
Hydrophobicity(Kyte and Doolittle)	0.122	0.705	0.211	0.68	0.2	0.74
Refrtactivity	0.165	0.479	0.153	0.483	0.18	0.59
Transmembrane tendency	0.197	0.782	0.25	0.78	0.23	0.73
(%) of buried residues	0.091	0.829	0.202	0.868	0.17	0.75
(%) of accessible residues	0.356	0.7	0.284	0.69	0.37	0.7
Ratio hetero end/side	0.076	0.332	0.04	0.389	0.08	0.36
Average area buried	0.167	0.49	0.197	0.52	0.18	0.54
Average flexibility index	0.412	0.852	0.38	0.796	0.43	0.88
Relative mutability of amino acids	0.374	0.716	0.345	0.702	0.31	0.7

Table 4: Important properties of Shikimate Kinase of *Yersinia pestis* calculated using Protscale.

Identification of intrinsic disorder in protein: The intrinsic disorder profile of SK protein sequences considered in the study obtained using different servers is illustrated in Figure 2 and 3. Disordered regions predicted in SK1, SK2 and SK3 using GLOBPROT and DisEMBL are shown in table 7.

Homology modeling: Homology modeling is perceived as an alternate method for obtaining insight into the protein structure in

Motif	Width	Site count	E-value	Start	Sequence	Motif
Motif 1	32	3	1.8e+004	43	gi|510938210|ref|WP_016255950.1|	VDTKDFQVMTQTIFMVGARGAGKTTIGKA-LAQALGYRFVDTDLFMQQTSQMT
				5	gi|668665097|gb|KFB61218.1|	MAEKRNIFLVGPMGAGKSTIGRQLAQQLNMEFFDSDQE-IERRTGAD
				4	gi|270336623|gb|EFA47400.1|	MAGQSIIVMGVSGSGKTTVGEAVARQIHAKFID-GDDLHPRANIQK
Motif 2	16	2	3.2e+001	50	gi|270336623|gb|EFA47400.1|	QPLNDADRMPWLERLS
				15	gi|510938210|ref|WP_016255950.1|	QPANNNGRFFDVENLS
Motif 3	11	2	4.5e+001	2 79	gi|510938210|ref|WP_016255950.1|	ICLCGVEPRSK
					gi|270336623|gb|EFA47400.1|	IIVCSALKRCY

Table 5: Motifs identified in considered Shikimate sequences using MEME.

Accession No.	Secondary structure	DSC	GOR IV	HNNC	PHD	Predator	SIPMA96	SOPM	SOPMA	Sec. consensus
KFB61218.1	Alpha helix	21.39	46.82	48.55	44.51	39.88	54.91	46.82	6.82	46.24
	Extended strand	27.75	18.5	16.18	19.08	15.61	13.87	19.65	19.65	16.18
	Beta turn	0	0	0	0	0	0	11.56	11.56	0
	Random coil	50.87	34.68	35.26	36.42	44.51	31.21	21.97	21.97	33.53
	Ambiguous states	0	0	0	0	0	0	0	0	4.05
WP_016255950.1	Alpha helix	30.84	42.52	49.53	38.79	40.19	46.73	43.46	43.46	39.72
	Extended strand	18.69	17.29	8.41	22.43	8.41	9.35	20.09	20.09	14.49
	Beta turn	0	0	0	0	0	0	11.21	11.21	0
	Random coil	50.47	40.19	42.06	38.79	51.4	43.93	25.23	25.23	41.59
	Ambiguous states	0	0	0	0	0	0	0	0	4.21

Table 6: Secondary structure elements predicted in considered sequences using NPS server.

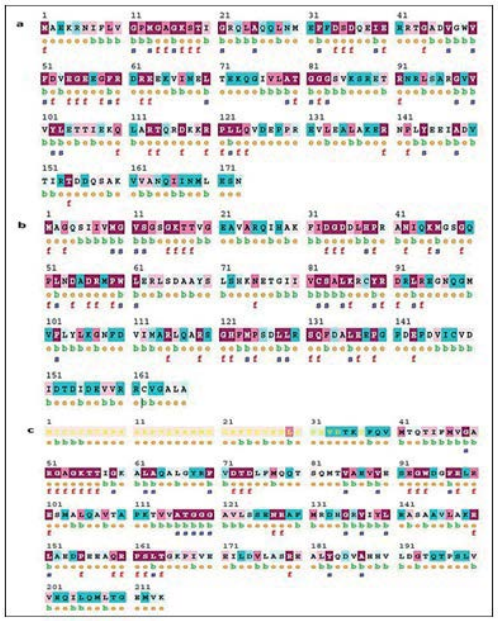

Figure 1: Conservation scores of amino acids. **A:** SK1(gi|668665097|gb|KFB61218.1), and, **B:** SK2(EFA47400.1) and **C:** SK3(gi|510938210|ref|WP_016255950.1) on a scale ranging from 0 to 9 indicating variable to conserved amino acids where e-An exposed residue according to the neural network algorithm, b-A buried residue according to the similar algorithm, f-A predicted functional residue (highly conserved and exposed), s-A predicted structural residue (highly conserved and buried), X-Insufficient data-the calculation for this site was performed on 10 of the sequences.

the absence of experimentally derived structures [42]. The modeling approach typically comprises of following steps—(i) Model generations by MODELLER9v3, (ii) selection of the best model on the basis of relative objective function values/DOPE score from the various models generated, (iii) Structure validation. Molecular modeling has been extensively used in recent past to obtain insight about drug targets for rational drug designing [16-19, 51]. All the 4 template structures (1E6C from *Erwinia chrysanthemi*, 1KAG from *Escherichia coli*, 1L4U from *Mycobacterium tuberculosis* and 1VIA from *Campylobacter jejuni*)) were compared for sequence identity and other properties before undertaking modeling exercise (Table 8) (Figure 4).

During the analysis, it was observed that for gi|668665097|gb|KFB61218.1|, single template based model number 5 with an objective function value of 772.6227 was the best model, whereas, for multiple template based modeling approach, it is model 4 with an objective function value of 6316.731. While the search for the best model based on both the approaches for gi|270336623|gb|EFA47400.1|, it was model 6 in both the cases with objective function value of 764.0565 and 6717.311. For protein gi|510938210|ref|WP_016255950.1|, it was model8 (888.4867) and model 7 (6453.276) for single and multiple template based model development exercises respectively Figure 5 and Figure 6 (Figure 8). As there is a basic difference in the range of objective function values obtained in the two approaches adopted Figure 7, therefore, analysis of the DOPE profile for the obtained best model was performed as depicted in Figure 8 and 9.

The observation suggested that for the proteins gi|668665097|gb|KFB61218.1|, gi|270336623|gb|EFA47400.1|[KIM D27] and gi|510938210|ref|WP_016255950.1| the best models were Model 4, Model6 and Model7 respectively. The GA431 value was found to be 1 for these models suggesting the good quality of the structures. All the structures were visualized and analyzed using VMD and Rasmol [52,53].

The number of helices and turns were found to be 10 and 5 in modeled structures of SK1, SK2, and SK3. SK1 showed only 13 strands

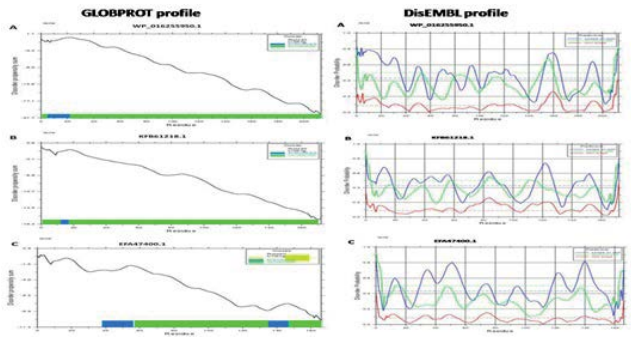

Figure 2: Disorder plot obtained using GLOBPROT and DisEMBL.

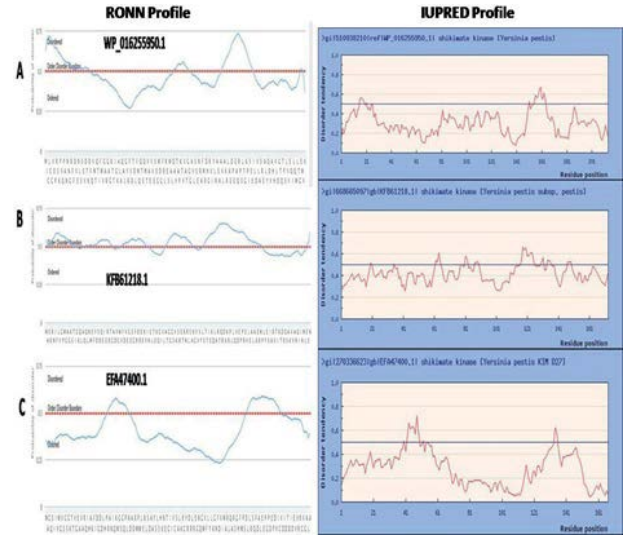

Figure 3: Disorder plot obtained using RONN and IUPRED.

	Definition	KFB61218.1	EFA47400.1	WP_016255950.1
GLOBPROT	Disordered by Russell/Linding definition	12 to 17	38-56, 135-147	5 to 22
	Potential globular domains (GlobDoms) by Russell/Linding definition	1-170	57-167	1-214
DisEMBL	Disordered by REM465	none	none	151-161
	Disordered by Loops/coil definition	1-19, 30-60, 78-95, 115-129, 142-157	30-61, 71-78, 118-152	1-38,48-55,67-80,91-98,107-130,152-167,189-197
	Disordered by HOTLOOPS definition	1-16, 69-102, 109-130, 148-158	1-16, 46-54, 70-81	1-23,151-168, 190-198

Table 7: Disordered regions predicted using GLOBPROT and DisEMBL.

while 15 strands were present in SK2 and SK3 respectively. Total H bonds present in SK1, SK2 AND SK3 were 117, 111 and 139 respectively. Shikimate kinase belongs to the Alpha and Beta proteins (a/b) category structurally. This particular type of structure majorly contains parallel

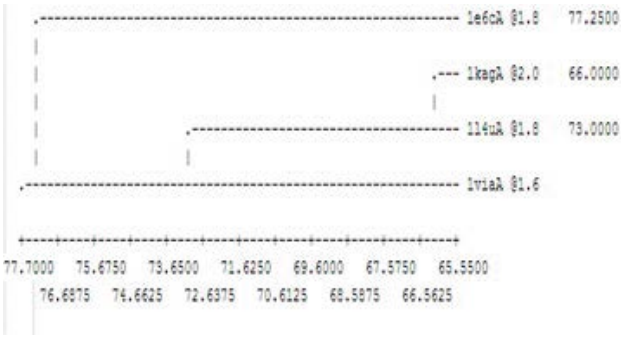

Figure 4: Cluster obtained using weighted pair-group method for the template proteins based on distance matrix.

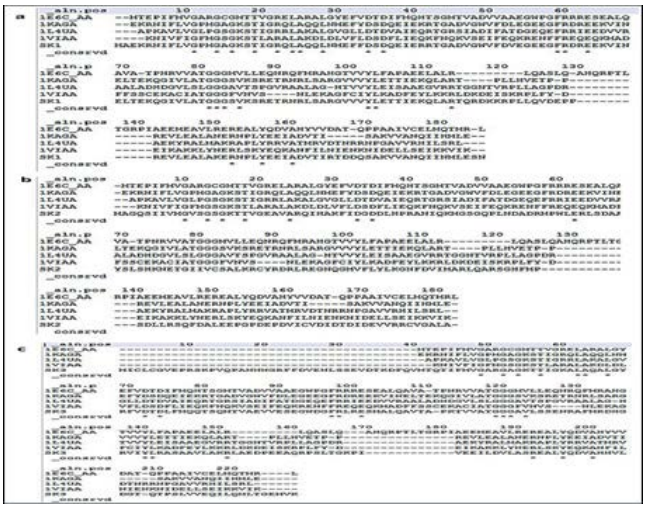

Figure 5: Target template alignment using single template 1E6C A chain for SK1 (gi|668665097|gb|KFB61218.1), SK2 (gi|270336623|gb|EFA47400.1) and SK3 (gi|510938210|ref|WP_016255950.1).

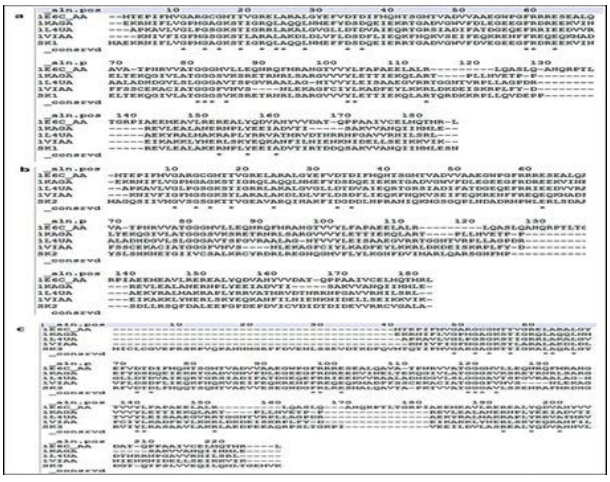

Figure 6: Target template alignment using multiple templates (1E6C A chain, 1KAG A chain, 1L4U A chain and 1V1A chain A) for SK! (gi|668665097|gb|KFB61218.1), K2(gi|270336623|gb|EFA47400.1) and SK3(gi|510938210|ref|WP_016255950.1).

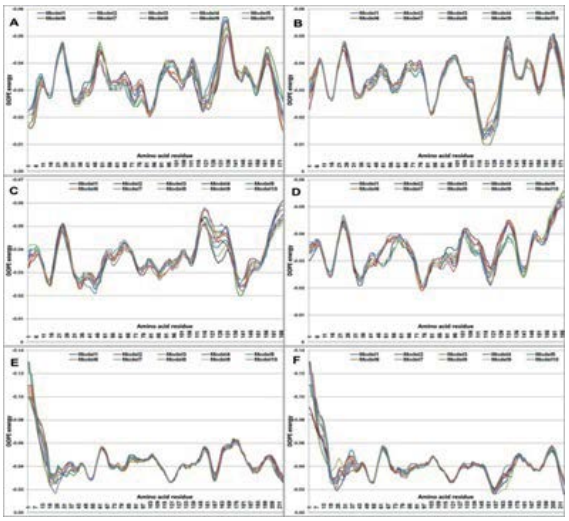

Figure 7: Distribution of the obtained Modeller objective function values. A: Representation of the single template (PDB ID: 1E6C, A chain) based models and B: Multiple template (1E6C, 1KAG, 1L4U and 1VIA "A" chains).

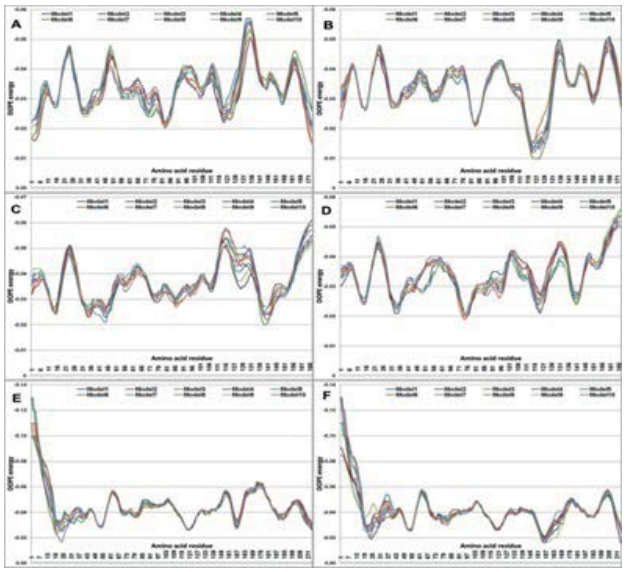

Figure 8: Distribution of the observed DOPE energy profile values for each residue in the considered protein models. The left panel of the figure (A, C, E) represents the energy profile computed for the single template (PDB ID: 1E6C, A chain) based models and the right panel (B,D,F) for the multiple template (1E6C, 1KAG, 1L4U and 1VIA "A" chains) based models. (A, B) Depiction of the DOPE profile for the protein structures for the target sequence gi|668665097|gb|KFB61218.1|, (C,D) for gi|270336623|gb|EFA47400.1|[KIM D27] and (E,F) for gi|510938210|ref|WP_016255950.1| respectively.

beta sheets following a beta-alpha-beta pattern of structural orientation (Figure 10, 11 and 12). The similar structural pattern is obtained for all the proteins in this study as depicted in Figure 11.

Structure validation: The modeled structures were evaluated using SAVES server (Figure 13-15). The geometry of model was evaluated with Ramachandran's plot calculations using PROCHECK. Stereochemical evaluation of backbone Psi and Phi dihedral angles revealed that 96.5%, 97.6% and 96.2% of residues were within the most favored regions in SK1, SK2, and SK3 respectively. Residues falling in

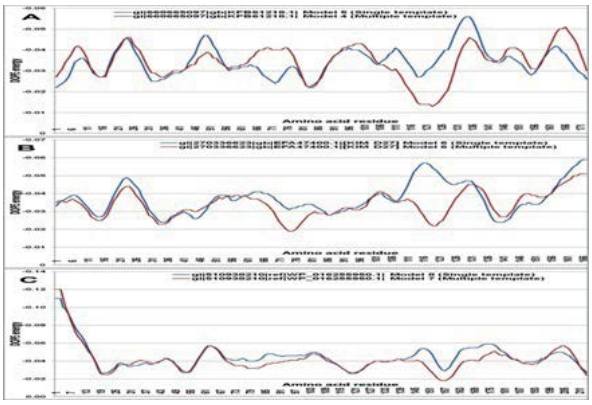

Figure 9: Comparative analysis of the DOPE profile for the best models obtained by single (Blue) and multiple (Red) template based approach.

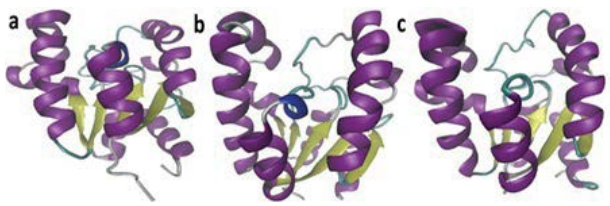

Figure 10: Models generated for **A:** SK1, **B:** SK2 and **C:** SK3.

additionally allowed regions in SK1, SK2 and SK3 were 2.9%, 0.6%, and 3.3% respectively. Residues in outlier region of SK1, SK2 and SK3 were 0.6%, 1.8% and 0.5% respectively (Figure 13). Z-Scores in PROVE server were negative which further establish the good quality of the models. These values indicate acceptable protein environment. RMSD Z-Score of backbone-backbone contacts, backbone-side chain contacts, side chain-backbone contacts and side chain-side chain contacts predicted using WHATCHECK (Table 9) were all within normal range. This also indicates the structural integrity of the models generated during the study.

PROSA was used to evaluate the quality of 3D models of protein structures. Z-score is a measure of overall model quality and denotes the deviation of the total energy of the structure compared to energy distribution derived from random conformations. The PROSA scores were negative for all the modeled protein, which indicate their correctness. PROSA profiles for the protein models were found similar to the template structures. Z-scores computed by PROSA for SK1, SK2, and SK3 were -6.67, -3.69 and -6.29 respectively which were similar to Z-scores of templates. Negative values in PROSA plot indicate stable regions of the protein. VERIFY 3D scores that indicate the compatibility of an atomic model (3D) with its own amino acid sequence were within acceptable range. ERRAT utilizes the statistics of non-bonded interactions between various types of atoms to assess the stability of protein structures. The ERRAT scores of modeled structure SK1, SK2 and SK3 were greater than 50 which reaffirm the reliability of the structure.

Active site identification: Once the final models were generated, possible binding sites of SK were searched using the CASTp server. Out of all the sites predicted, best 10 sites were selected as shown in Figure 16. Binding sites having highest surface area and volume were selected

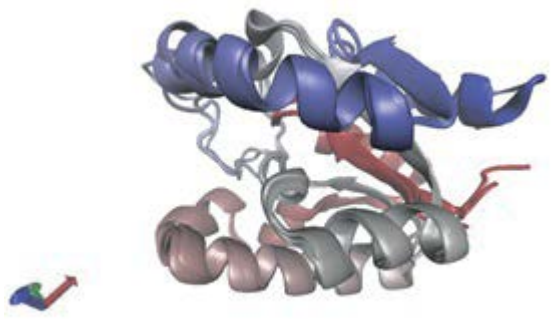

Figure 11: Superimposition of modeled structures of SK1, SK2, and SK3.

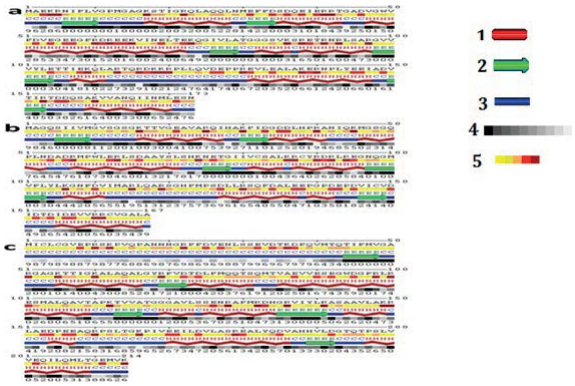

Figure 12: (a) Secondary structure, physicochemical profile and solvent accessible surface area as predicted by POLYVIEW where (1) H-α and other helices (view 1), (2) E-β-strand or bridge, (3) C–coil, (4) Relative solvent accessibility (RSA) where 0-completely buried (0-9% RSA), 9-fully exposed (90-100% RSA), (5) where H-hydrophobic: A,C,F,G,I,L,M,P,V; A-amphipathic: H,W,Y; P-polar: N,Q,S,T and N/C-charged: D,E–negative, R,K–positive for SK1(a), S2(b) and SK3(c).

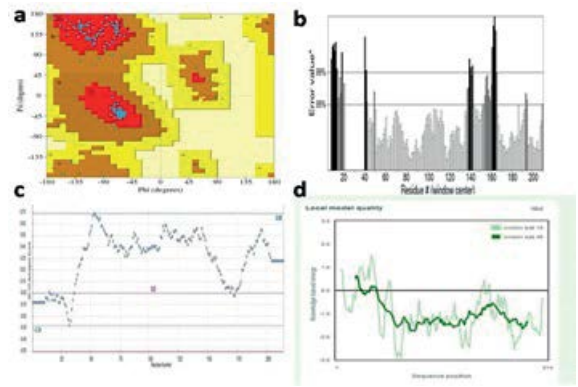

Figure13: A: Ramachandran plot analysis **B:** ERRAT2 score **C:** Verify 3D plot **D:** PROSA energy plot of the modeled structure Sk1.

as the most probable active sites for each modeled protein.

Out of 21 pockets selected for SK1 , the pocket having highest volume and area comprised of PRO12, MET13, ASP36, VAL47, PHE51, PHE59, ARG60, GLU63, GLY81, Gly82, GLy83, SEr84, TYR102,

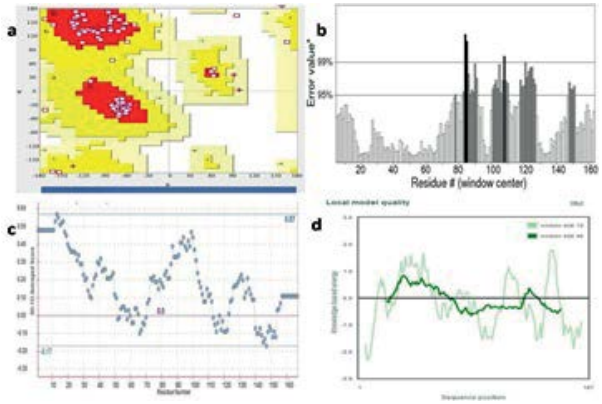

Figure 14: A: Ramachandran plot analysis **B:** ERRAT2 score **C:** Verify 3D plot **D:** PROSA energy plot of the modeled structure SK2.

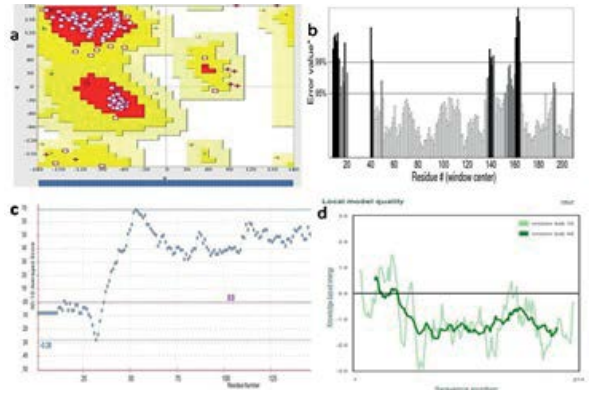

Figure 15: A: Ramachandran plot analysis **B:** ERRAT2 score **C:** Verify 3D plot **D:** PROSA energy plot of the modeled structure SK3.

THR105, GLN110, LYS118, ARG120, PRO121, LEU122, VA125, ASP126, PRO129, LEU133, LEU136, ALA137, ARG140, ASN 141 and TYR144.

Out of 30 pockets predicted for SK2 using CASTP, a site showing surface area and volume of 755.1 and 1018.9 was selected. The residues forming the pocket of SK2 were VAL11, SER12, GLY13, GLY15, LYS 16, THR17, ASP33, ASP35, ASP36, PRO39, LYS45, MET46, GLN50, MET58, ILE80, VAL81, CYS82, PHE109, ASP110, ILE112, MET113, ALA114, LEU116, GLN117, ARG119, SER120, GLY121, HIS122, PHE123, MET124, PRO125, SER126, LEU128, LEU129, GLN132 and PHE132.

A total of 28 pockets were predicted for SK3 and pocket having the highest surface area (871.5) and volume (1286.4) was selected. It comprised of LEU123, ALA144, LEU147, ALA148, LYS149, ARG150, LEU151, ASP154, PRO155, GLU156, GLU157, ALA158, GLN159, ARG160, PRO161, SER162, LEU163, ILE168, VAL169, GLU171, ILE172, LEU173, VAL175, LEU176 and ARG179.

Shikimate Kinase family of enzymes are one of the most important enzyme family which is considered under the P-loop containing nucleoside triphosphate superfamily specifically having a parallel beta sheet containing structures with an important domain catalyzing phosphorylation having highly conserved motifs. This enzyme is a unique drug target due to its absence in human and presence in harmful microorganisms.

Sequence identity comparison:				
Diagonal= Number of residues; Upper triangle = Number of identical residues; Lower triangle = Percentage of sequence identity.				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	170	42	43	31
1kagA	27	158	54	46
1l4uA	26	34	165	40
1viaA	19	29	25	161
Position comparison (FIT_ATOMS):				
Cutoff for RMS calculation: 3.5, Upper=RMS,Lower = Number of equivalent positions.				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	0	1.429	1.534	1.479
1kagA	136	0	1.412	1.25
1l4uA	130	146	0	1.315
1viaA	128	144	142	0
Distance comparison (FIT_ATOMS):				
Cutoff for RMS calculation:3.5, Upper=Distance RMS, Lower = Number equivalent distances				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	0	0.992	1.114	1.142
1kagA	9132	0	0.985	0.949
1l4uA	8695	10547	0	1.026
1viaA	8418	10278	10542	0
Sequence Comparison				
Diagonal= Number of residues, Upper= Number of equivalent residues, Lower= percent sequence identity				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	170	42	43	31
1kagA	27	158	54	46
1l4uA	26	34	165	40
1viaA	19	29	25	161
Dihedral angle (Alpha) comparison:				
Cutoff for RMS calculation:60, Upper= RMS Alpha, Lower = Number equivalent angles				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	0	11.455	14.871	14.939
1kagA	125	0	10.742	13.355
1l4uA	121	130	0	13.656
1viaA	119	130	134	0
Dihedral angle (Phi) comparison:				
Cutoff for RMS calculation:60, Upper= RMS Phi, Lower = Number equivalent angles				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	0	13.577	13.91	15.118
1kagA	130	0	14.896	13.326
1l4uA	126	141	0	16.213
1viaA	124	140	140	0
Dihedral angle (Psi) comparison:				
Cutoff for RMS calculation:60, Upper= RMS Psi, Lower = Number equivalent angles				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	0	7.247	7.154	7.248
1kagA	114	0	6.531	7.182
1l4uA	96	108	0	7.113
1viaA	98	116	106	0
Dihedral angle (Omega) comparison:				
Cutoff for RMS calculation:60, Upper= RMS Omega, Lower = Number equivalent angles				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	0	4.095	3.974	4.343
1kagA	134	0	5.029	4.189
1l4uA	132	143	0	3.107
1viaA	130	142	145	0
Dihedral angle (Chi1) comparison:				
Cutoff for RMS calculation:60, Upper= RMS Chi1, Lower = Number equivalent angles				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	0	16.501	15.555	11.795
1kagA	62	0	15.765	11.129

1l4uA	53	54	0	10.011
1viaA	51	52	64	0
Dihedral angle (Chi2) comparison:				
Cutoff for RMS calculation:60, Upper= RMS Chi2, Lower = Number equivalent angles				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	0	16.75	16.113	19.006
1kagA	35	0	15.674	17.66
1l4uA	32	35	0	19.045
1viaA	36	40	41	0
Dihedral angle (Chi3) comparison:				
Cutoff for RMS calculation:60, Upper= RMS Chi3, Lower = Number equivalent angles				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	0	24.802	30.063	27.978
1kagA	12	0	21.919	19.114
1l4uA	10	7	0	18.665
1viaA	11	15	10	0
Dihedral angle (Chi4) comparison:				
Cutoff for RMS calculation:60, Upper= RMS Chi4, Lower = Number equivalent angles				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	0	13.73	36.668	3.267
1kagA	4	0	23.668	10.418
1l4uA	5	4	0	31.644
1viaA	3	5	6	0
Dihedral angle (Chi5) comparison:				
Cutoff for RMS calculation:60, Upper= RMS Chi5, Lower = Number equivalent angles				
	1e6cA	**1kagA**	**1l4uA**	**1viaA**
1e6cA	0	0	0	0
1kagA	0	0	0	0
1l4uA	0	0	0	0
1viaA	0	0	0	0

Table 8: Representation of the template comparison with respect to sequence and structural properties.

Interactions	SK1		SK2		SK3	
	Average	Z score	Average	Z score	Average	Z score
All contacts	-0.167	-0.97	-0.545	-3.12	-0.466	-2.67
Backbone-Backbone	-0.038	-0.32	-0.181	-1.28	-0.205	-1.44
Backbone-Side Chain	-0.03	-0.33	-0.474	-3.54	-0.327	-2.48
Side Chain-Backbone	-0.274	-1.66	-0.492	-3.03	-0.444	-2.73
Side Chain-Side Chain	-0.235	-1.24	-0.787	-4.66	-0.506	-2.92

Table 9: Predicted Z-Scores of modeled structures using Whatcheck.

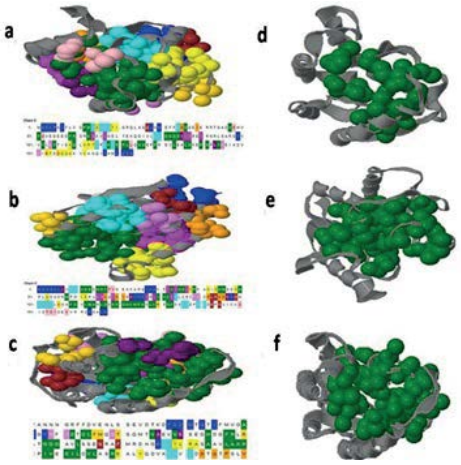

Figure 16: Top 10 possible binding sites as predicted by CASTP where Pocket 1=green, Pocket 2=blue, Pocket 3=cyan, Pocket 4=yellow, Pocket 5=magenta, Pocket 6=pink, Pocket 7=orange, Pocket 8=purple, Pocket 9=brown, Pocket 10=gold for SK1(**A**), SK2(**B**) and SK3(**C**) and active sites of the modeled proteins SK1(**D**), SK2(**E**) and SK3(**F**).

Conclusion

A sudden rush for sequencing has led to a huge gap in the number of protein sequences available and experimentally derived structures. In such cases, comparative modeling plays a pivotal role in providing insight about protein structure in the absence of crystal structures. Development of resistance in Yersinia towards the available drugs underscores the need for exploring and exploiting novel drug targets and drugs. In this study, bioinformatics tools were applied for determining important features and properties of the SK of *Yersinia pestis*. The extent of similarity between target and template protein is the deciding factor for the accuracy of predicted structure. We have obtained models of reasonably good quality which was affirmed by various structure validation tools. Results of this study will aid in understanding this enzyme and will pave a way for effective inhibitor design. These models will aid by providing valuable insights about the structure and binding pockets of Shikimate Kinase in *Yersinia pestis* and will aid in rational drug designing.

Acknowledgement

Authors are thankful to Director, Institute of Science and Technology, JawaharLal Nehru Technological University for her constant support and encouragement throughout the study. Dr. Neelima Arora thanks University Grants Commission (UGC) for Dr. D.S. Kothari postdoctoral fellowship.

References

1. DeWitte SN (2009) The effect of sex on risk of mortality during the Black Death in London, A.D. 1349-1350. Am J Phys Anthropol 139: 222-234.
2. Dewitte SN (2010) Age Patterns of Mortality During the Black Death in London, A.D. 1349-1350. J Archaeol Sci 37: 3394-3400.
3. Boire NA, Riedel VAA, Parrish NM, Riedel S (2013) Lessons Learned from Historic Plague Epidemics: The Relevance of an Ancient Disease in Modern Times. Anc Dis Prev R 2: 1-17.
4. Oyston PC, Williamson D (2011) Plague: Infections of Companion Animals and Opportunities for Intervention. Animals (Basel) 1: 242-255.
5. Inglesby TV, Dennis DT, Henderson DA, Bartlett JG, Ascher MS, et al. (2000) Plague as a biological weapon: medical and public health management. Working Group on Civilian Biodefense. JAMA 283: 2281-2290.
6. Abell C, Coggins JR, Evans LB, Frederickson M, Robinson DA, et al. (2003) Experiences with the shikimate-pathway enzymes as targets for rational drug design. Biochem Soc Trans 31: 548-552.
7. Roberts CW, Roberts F, Lyons RE, Kirisits MJ, Mui EJ, et al. (2002) The shikimate pathway and its branches in apicomplexan parasites. J Infect Dis 185 Suppl 1: S25-36.
8. Herrmann KM (1995) The shikimate pathway as an entry to aromatic secondary metabolism. Plant Physiol 107: 7-12.
9. Maeda H, Dudareva N (2012) The shikimate pathway and aromatic amino Acid biosynthesis in plants. Annu Rev Plant Biol 63: 73-105.
10. Filgueira de AWJ, Canduri F, Simões de Oliveira J, Basso LA, Palma MS, et al. (2002) Molecular model of shikimate kinase from Mycobacterium tuberculosis. Biochem Biophys Res Commun 295: 142-148.
11. Paulia I, Caceresa RA, de Azevedo WFJ (2008) Molecular modeling and dynamics studies of Shikimate Kinase from Bacillus anthracis. Bioorg Med Chem 16: 8098–8108.
12. Arora N, Banerjee AK, Murty USN (2010) In silico characterization of Shikimate Kinase of Shigella flexneri: A potential drug target. Interdisciplinary Sciences. Computational Life Sciences. 2: 280-290.
13. Arora N, Banerjee AK, Murty USN (2010) Homology model of 2C-methyl-d-erythritol 2, 4-cyclodiphosphate (MECP) synthase of Plasmodium falciparum 3D7. EJB 6: 52-57.
14. Arcuri HA, Canduri F, Pereira JH, da Silveira NJ, Camera Júnior JC, et al. (2004) Molecular models for shikimate pathway enzymes of Xylella fastidiosa. Biochem Biophys Res Commun 320: 979-91.
15. Krell T, Coggins JR, Lapthorn AJ (1998) The three-dimensional structure of shikimate kinase. J Mol Biol 278: 983-997.
16. Arora N, Chari UVNM, Banerjee AK, Murty USN (2007) A computational approach to explore Plasmodium falciparum 3D7 chorismate synthase. Int J Genomics Proteomics 3: 1-11.
17. Banerjee AK, Arora N, Murty USN (2009) Structural model of the Plasmodium falciparum Thioredoxin reductase: a novel target for antimalarial drugs. J Vector Borne Dis 46: 171–183.
18. Banerjee AK, Arora N, Murty U (2012) Analyzing a Potential Drug Target N-Myristoyltransferase of Plasmodium falciparum Through In Silico Approaches. J Glob Infect Dis 4: 43-54.
19. Srikanth Duddela, P. Nataraj Sekhar, G. V. Padmavati, Amit Kumar Banerjee, U.S.N. Murty. Probing the structure of human glucose transporter 2 and analysis of protein ligand interactions. Med. Chem. Res., 19: 836-853.
20. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, et al. (2005) Protein identification and analysis tools on the ExPASy server. In: John M. Walker, The Proteomics Protocols Handbook. (edn), Humana Press, United States, pp: 571-607.
21. Bailey TL1, Boden M, Buske FA, Frith M, Grant CE, et al. (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37: W202-208.
22. Celniker G, Nimrod G, Ashkenazy H, Glaser F, Martz E, et al. (2013) ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. IJC 53: 195-206.
23. Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, et al. (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics 20: 1322-1324.
24. Combet C, Blanchet C, Geourjon C, Deléage G (2000) NPS@: network protein sequence analysis. Trends Biochem Sci 25: 147-150.
25. Deléage G, Roux B (1987) An algorithm for protein secondary structure prediction based on class prediction. Protein Eng 1: 289-294.
26. King RD, Sternberg MJ (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci 5: 2298-2310.
27. Garnier J, Gibrat JF, Robson B (1996) GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 266: 540-553.
28. Guermeur Y, Geourjon C, Gallinari P, Deleage G (1999) Improved performance in protein secondary structure prediction by inhomogeneous score combination. Bioinformatics 15: 413–421.
29. Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232: 584-599.
30. Frishman D, Argos P (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Eng 9: 133-142.
31. Levin JM, Robson B, Garnier J (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. FEBS Lett 205: 303-308.
32. Geourjon C, Deleage G (1995) SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosc 11: 681–684.
33. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, et al. (2003) Protein disorder prediction: implications for structural proteomics. Structure 11: 1453-1459.
34. Linding R, Russell RB, Neduva V, Gibson TJ (2003) GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res. 31(13):3701-3708.
35. Yang ZR, Thomson R, McNeil P, Esnouf RM (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21:3369-3376.
36. Ishida T, Kinoshita K (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res 35: W460-464.
37. Krell T, Maclean J, Boam DJ, Cooper A, Resmini M, Brocklehurst K, et al., (2001) Biochemical and X-ray crystallographic studies on shikimate kinase: the

important structural role of the P-loop lysine. Protein Sci. 10(6):1137-1149.

38. Romanowski MJ, Burley SK (2002) Crystal structure of the Escherichia coli shikimate kinase I (AroK) that confers sensitivity to mecillinam. Proteins 47(4):558-562.

39. Gu Y, Reshetnikova L, Li Y, Wu Y, Yan H, Singh S, Ji X (2002) Crystal structure of shikimate kinase from Mycobacterium tuberculosis reveals the dynamic role of the LID domain in catalysis. J Mol Biol. 319(3):779-789.

40. Badger J, Sauder JM, Adams JM, Antonysamy S, Bain K, Bergseid MG, et al. (2005) Structural analysis of a set of proteins resulting from a bacterial genomics project. Proteins. 60(4):787-796.

41. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876-4882.

42. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779-815.

43. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: A program to check the stereochemical quality of protein structures. J App Cryst 26: 283–291.

44. Vriend G (1990) WHAT IF: a molecular modeling and drug design program. J Mol Graph 8: 52-56, 29.

45. Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381: 272.

46. Eisenberg D, Lüthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol 277: 396-404.

47. Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 2: 1511-1519.

48. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35: W407-410.

49. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J (2006) CASTp: Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34: W116–W118.

50. Cheng WC, Chang YN, Wang WC (2005) Structural basis for shikimate-binding specificity of Helicobacter pylori shikimate kinase. J Bacteriol 187: 8156-8163.

51. Banerjee AK, Arora N, Murty USN (2012) Aspartate carbamoyltransferase of Plasmodium falciparum as a potential drug target for designing anti-malarial chemotherapeutic agents. Med Chem Res 21: 2480-2493.

52. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14: 33-38, 27-8.

53. Sayle RA, Milner-White EJ (1995) RASMOL: biomolecular graphics for all. Trends Biochem Sci 20: 374.