J Mol Model (2005) 11: 248–255
DOI 10.1007/s00894-005-0272-7

ORIGINAL PAPER

Santhanam Vijayasri · Shipra Agrawal

Domain-based homology modeling and mapping of the conformational epitopes of envelope glycoprotein of west nile virus

Received: 20 May 2004 / Accepted: 15 March 2005 / Published online: 3 May 2005

Abstract Knowledge-based modeling has proved significantly accurate for generating the quality models for proteins whose sequence identity with the structurally known targets is greater than or equal to 40%. On the other hand, models obtained for low sequence identities are not reliable. Hence, a reliable and alternative strategy that uses knowledge of domains in the protein can be used to improve the quality of the model generated by the homology method. Here, we report a method for developing a 3D-model for the envelope glycoprotein (Egp) of west nile virus (WNV), using knowledge of structurally conserved functional domains amongst the target sequence (Egp of WNV) and its homologous templates belonging to the same protein family, flaviviridae. This strategy is found to be highly effective in reducing the root mean square deviation (RMSD) value at the Cα positions of the target and its experimental homologues. The 3D structure of a protein is a prerequisite for structure-based drug design as well as for identifying the conformational epitopes that are essential for the designing vaccines. The conformational epitopes are mapped from the 3D structure of Egp of WNV modeled using the concept of an antigenic domain. A total of five such epitope regions/sites have been identified. They have been found distributed in the loop regions (surface) of the whole protein model composed of dimerization, central and immunological domains. These sites are proposed as the binding sites for HLA proteins/B-cell receptors. Binding is required to activate the immune response against WNV.

S. Vijayasri · S. Agrawal (✉)
Indian Institute of Information Technology,
Allahabad, Deoghat Jhalwa Campus, Allahabad,
211011, India
E-mail: shipra_ag@yahoo.co.uk
Tel.: +91-532-2431684
Fax: +91-532-2430006

Keywords West nile virus · Flaviviridae · Comparative modeling · Envelope glycoprotein · Domains · Genome polyprotein · Template structure

Introduction

Today, genome-sequencing projects of various organisms have led to the accumulation of voluminous data of gene and protein sequences. Proteins are the ultimate building blocks and play key roles in different physiological and metabolic processes of an organism. Determining the functionality of a protein can be greatly aided by knowledge of its 3D structure in space. However, structural information for 25,251 proteins has been determined and made available in the PDB format. A large number of proteins are yet to be worked out. Hence, the structural information of these large numbers of protein sequences, which can be analyzed using computational approaches of structure determination, is of great biological significance and can reduce the knowledge gap between sequence and structure information. In addition, the 2D/3D structures have been used in designing potential drug targets and in protein-folding studies. Also, knowledge-based protein modeling (homology) has been proved to produce remarkably good results [1]. Hence, to explore these structural aspects here, model building is applied to a structurally unknown protein, whose sequence homology with the proteins of known experimental 3D structures is taken as the basis.

The accuracy of the above-mentioned models depends on the structure of the template, which acts as a backbone in the construction of the protein model. As the model coming out of the homology process should emphasize the functionality of the protein, it is advisable to consider the functional status of the target sequence and the template structures before modeling. The active sites of a protein often reside in the structurally conserved regions (SCRs). The SCRs can be constructed

well with knowledge of the conserved motifs, folds and domains of the template protein structure and the target sequence. It has already been established that lower sequence homology between the target and template leads to an unreliable model. On the other hand, researchers are exploring strategies for obtaining high quality models in the aforementioned situation.

Proteins of the same superfamilies have similar structures and functions, despite low sequence homology [2]. A recent report suggests that a reliable protein model can be developed using common conserved motifs in distantly related homologues of cytochrome P450 [3]. Similarly, the quality of the model can also be improved by considering a domain strategy while selecting the templates. The folding of domains takes place individually and they also contribute to the independent specific function(s).

The present study is also an effort in the same direction. We emphasize the application of knowledge of conserved domains imparting characteristic functionality to both target and template. Besides, the model has been refined by loop modeling using a local loop database in Swissmodel. Consequently, the energy of the model and root mean square deviation (RMSD) at Cα positions of target and the template sequence are reduced significantly.

Here, we will describe a 3D structure model of the coat polyprotein (Egp) of west nile virus (WNV), which causes encephalitis in mammals, using two templates, i.e., dengue virus envelope protein (1OAN) and Egp of tick-borne encephalitis virus (1SVB) [4]. The structure has been developed on the concept that all the members of flaviviridae family contain the domains, namely Flavi_glycoprot and Flavi_glycopC, that have antigenicity property and carrying characteristic 12 cysteine residues, which make six disulfide bridges [5].

Further, the model has been utilized to characterize the conformational epitopes, as they are essentially needed to design potent vaccine candidates against this pathogen. The residues of conformational epitopes may serve as target sites for structure-based drug design.

Materials and methods

The Egp of WNV of length 3,430 amino acids codes of the genome polyprotein was extracted from the SWISSPROT database (accession number-P06935) [6].

The templates, i.e., dengue 2 virus Egp (1OAN) and Egp of (1SVB) were selected on the basis of the identity score, i.e., 45 and 38%, respectively, and also the "e" value from the BLASTP results [7]. The target and the templates were studied for their domain classification using Pfam [8].

The three dimensional structure of the target protein was modeled using the SWISSMODEL program [9]. The secondary structure of the target protein was predicted using the DSSP program [10]. The RMSD between the protein structures were obtained by SWISSMODEL. The model was refined on the basis of energy minimization by GROMOS96 [11] and loop building with the help of the ProMod tool by scanning through the loop database in SWISSMODEL [12]. The model was validated for the 3D–1D profile with VERIFY3D, [13] non-bonded interactions with ERRAT [14] and stereochemical qualities with PROCHECK [15] and WHATCHECK [16].

Results and discussion

Model building and refinement

A recent report describes comparative modeling of the proteins using a conserved motif strategy that can be applicable to proteins forming superfamilies. Domains that are conserved in the proteins of the same family will have more similar functions. Thus, the model built by selecting templates on the basis of a domain strategy will be more reliable, not only structurally but also functionally.

Considering the aforementioned logic and also the large size of our target polyprotein (Egp of WNV), which does not have any homologues as a whole, we have focused our strategy to model the domains of Egp having antigenic character. The antigenic domains were identified from the Pfam report (Table 1). The target and the templates were found to share most of the domains. On the basis of this analysis, the sequence stretch (291–685 amino acids) that makes specific domains and imparts antigenicity to the Egp was considered for modeling.

The Egp domain is made of a central domain, dimerization domain and immunological domain (Fig. 1). Based on this domain concept, we have selected

Table 1 Domain classification of the WNV genome polyprotein. The highlighted domains are taken to model the protein since they are the regions of antigenicity

Source	Domain	Start residue	End residue
Pfam	Flavi_capsid	6	123
Pfam	Flavi_propep	128	214
Pfam	Flavi_M	216	290
Pfam	**Flavi_glycoprot**	**291**	**585**
Pfam	**Flavi_glycop_C**	**587**	**685**
Pfam	**Pfam-B 130**	686	787
Pfam	Flavi_NS1	789	1,143
Pfam	Flavi_NS2A	1,151	1,370
Pfam	Flavi_NS2B	1,371	1,501
Pfam	Pepitdase_S7	1,508	1,679
Pfam	**Pfam-B_31**	1,695	1,829
Pfam	Pfam-B_187	1,830	1,860
Pfam	Helicase_C	1,876	1,967
Pfam	Flavi_NS4A	2,123	2,267
Pfam	Flavi_NS4B	2,270	2,519
Pfam	FtsJ	2,579	2,753
Pfam	Flavi_NS5	2,778	3,426
Smart	DEXDc	1,668	1,856
Smart	HELICc	1,871	1,967

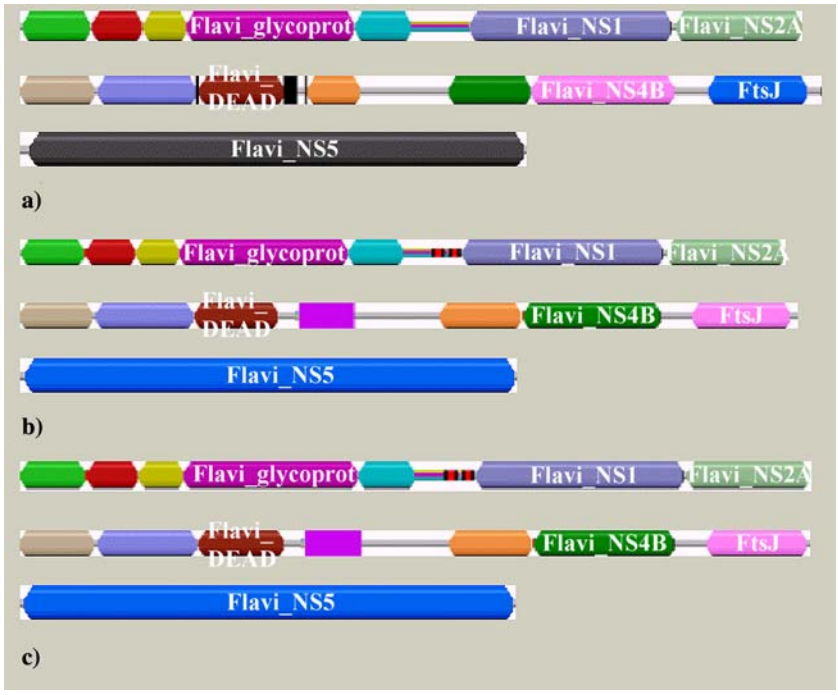

Fig. 1 Pfam domain classification and description in target and templates sequences. **(a)** Genome polyprotein of WNV. **(b)** Genome polyprotein of Dengue virus. **(c)** Genome polyprotein of tick borne encephalitis virus. It describes the uniqueness among the *flaviviridae* family viruses. Most of the domains are common in all three proteins (14) except presence of an extra domain namely transmembrane protein in the template sequences (**b**) and (**c**). The Flavi_glycoprot domain denotes central and dimerization domains and its adjacent domain i.e., Flavi_glycop_C is immunological like domain. They cumulatively form the envelope protein. These are the key targets of the immune response to flaviviruses. The central domain contains the glycosylation site in addition to the epitopes with serological or biological activities. The dimerization domain is the site of neutralization and hemagglutination. The third domain is the immunological like domain (http://www.sanger.ac.uk/Software/Pfam/)

dengue 2 virus envelope protein (1OAN) and Egp of (1SVB) as the templates to model the Egp of WNV (P06935).

The template sequences were modified according to their structural data in the PDB. WNV is closely related to the Kunjin virus, for which no structural data are available in PDB, but is related to dengue 2 virus envelope protein [17]. The target is also related antigenically to the 1SVB [5]. Hence, these two proteins were taken as the templates. More effective use of multiple templates and better alignments promise to improve the generated model significantly (Fig. 2).

Multiple sequence alignment was performed to map the SCRs and a phylogenetic tree was drawn with PHYLODRAW. It showed the closer evolutionary relationship between the target and the templates (Table 2) [18]. The superimposition of the two templates with RMSD 1.38 Å emphasizes the good structural alignment between them (Fig. 3). The target protein was modeled with SWISSMODEL (Fig. 4) by fitting the raw sequence to the superimposed structure of the templates. Both the templates have several common structural features like the presence of more β-sheets with only two short α-helices and overall structural pattern (Fig. 5). The alignment between the target sequence and the template structures was optimized by manual alignment such that the initial energy was reduced drastically from 279.0 to 160.0 kJ mol^{-1} . Care was taken not to allow any gap between the SCRs.

The loop that corresponds to the sterically disallowed region in the models obtained before was replaced with the most plausible loop by scanning through the loop database in the Swiss PDB viewer. The energy was minimized with 200 cycles of steepest descent. The final model thus obtained after substitution with the most plausible loop was found to be satisfactory in all aspects. It is a better 3D structure with the minimized energy ($-15{,}286.964$ kJ mol^{-1}) (Fig. 6).

It was found that 91.94% of the residues in the final model have an averaged 3D-1D score greater than 0.2 (Fig. 7), which is better than that of previous models. The non-bonded interactions were also found to be favorable (quality factor = 91.099) (Fig. 8). A Ramachandran plot shows that the most of the residues are in the sterically allowed region. The few amino acids (approximately ten amino acids) that are in the disallowed region belong to the loop regions corresponding to the structurally variable region. The validation results done in terms of RC plot, chi1–chi2 plots, main chain and side chain parameters, G–factors, M/c bond length, M/c bond angles (by PROCHECK) show that the model is good (Table 3). All the amino acid nomenclature was found to be normal as per WHATCHECK report.

The protein modeled here has been found to be satisfactory on the grounds of energy, steric features,

Fig. 2 Multiple sequence alignment of the target with the template sequences showing the great degree of SCRs in the sequences (CLUSTAL W). The first sequence is of WNV (P06935), the second sequence is of DEN 2P (P12823) and the third sequence is of TBEVW (P14336)

```
FNCLGMSNRDFLEGVSGATWVDLVLEGDSCVTIMSKDKPTIDVKMMNMEAANLADVRSYCYLAS
MRCIGISNRDFVEGVSGGSWVDIVLEHGSCVTTMAKNKPTLDFELIKTEAKQPATLRKYCIEAK
SRCTHLENRDFVTGTQGTTRVTLVLELGGCVTITAEGKPSMDVWLDAIYQENPAKTREYCLHAK
 .*  :.****: *..* : * .*** ..***  ::.**::*. :     :*  *.**  *.

VSDLSTRAACPTMGEAHNEKRADPAFVCKQGVVDRGWGNGCGLFGKGSIDTCAKFACTTK—-AT
LTNTTTDSRCPTQGEPTLNEEQDKRFVCKHSMVDRGWGNGCGLFGKGGIVTCAMFTCKKN—-ME
LSDTKVAARCPTMGPATLAEEHQGGTVCKRDQSDRGWGNHCGLFGKGSIVACVKAACEAKKKAT
::: .. :***  *  .  :. :  ***:.  ****** *******. * :*. :* :

GWIIQKENIKYEVAIFVHGPTTVESHGKIGATQAGRFSITPSAPSYTLKLGEYGEVTVDCEPRS
GKIVQPENLEYTVVITPHS-GEEHAVGNDTGKHGKEVKITPQSSITEAELTGYGTVTMECSPRT
GHVYDANKIVYTVKVEPHTGDYVAANETHSGRKTASFTIS—-SEKTILTMGEYGDVSLLCRVAS
* : : ::: * * :  *     :  . . :   ..*:..:     : ** *:: *   :

GIDTSAYYVMSVGE------KSFLVHREWFMDLNLPWSSAGST---TWRNRETLMEFEEPHATK
GLDFNEMVLLQMKD------KAWLVHRQWFLDLPLPWLPGADTQGSNWIQKETLVTFKNPHAKK
GVDLAQTVILELDKTVEHLPTAWQVHRDWFNDLALPWKHEGAQ---NWNNAERLVEFGAPHAVK
*:*    :::: .    .:: ***:** ** ***   .   .* : * *: *  *** *

QSVVALGSQEGALHQALAGAIPVEFSSNTVKLTSGHLKCRVKMEKLQLKGTTYGVCSK-AFKFA
QDVVVLGSQEGAMHTALTGATEIQMSSGNLLFT-GHLKCRLRMDKLQLKGMSYSMCTG-KFKVV
MDVYNLGDQTGVLLKALAGVPVAHIEGTKYHLKSGHVTCEVGLEKLKMKGLTYTMCDKTKFTWK
 .*  **.* *.:  **:*.   .::. .  :.:**:.*.: ::**::** :* :*   *.

RTPADTGHGTVVLELQYTGTDGPCKVPISSVASLNDLTPVGRLVTVNPFVSVATANSKVLIELE
KEIAETQHGTIVIRVQYEGDGSPCKTPF-EIMDLEKRHVLGRLTTVNPIVTEK—-DSPVNIEAE
RAPTDSGHDTVVMEVTFSGT-KPCRIPVRAVAHGSPDVNVAMLITPNPTIENN---GGGFIEMQ
:  ::: * *:*:*.:: : * . **: *.  :   .   :. * * ** :   ::...   ** :

PPFGDSYIVVGRGEQQINHHWHKS
PPFGDSYIIIGVEPGQLKLDWFKK
LPPGDNIIYVG----ELSHQWFQK
 * **. * :*    ::. .*::.
```

Table 2 Pair distances between the template and the target sequences as given by the PHYLODRAW

S. no.	Gene	Gene	Distance (Å)
1	sp|P06935|POLG_WNV	sp|P12823|DEN2P	0.530460
2	sp|P06935|POLG_WNV	sp|P14336|TBEVW	0.600000
3	sp|P12823|DEN2P	sp|P14336|TBEVW	0.624360
Average distance			0.584940
Variation in distance			0.022971

non-bonded interactions and 3D–1D profile and stereochemical parameters and also on the basis of RMSD. The secondary structural data generated by DSSP also show the reliability of the model to be good on the basis of their secondary structural content.

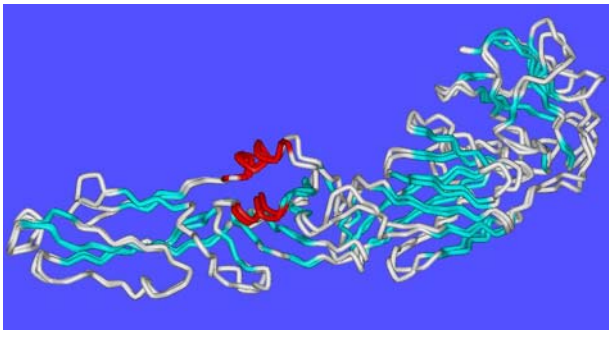

Fig. 3 Superimposed image of the two templates with the RMSD 1.38 Å shows the closest structural similarity between the templates

Structural description of the model

The protein model has more β-sheets with only three short α-helices. On comparison with the model of the Egp of Japanese Encephalitis virus [19], it is found that there is an additional α–helix that may play an important role and may elicit certain function. The same observation is made with the two template structures. The model and templates are found to have similar folds.

In addition, the model and templates show the presence of three functional domains; central domain (domain I), dimerization domain (domain II), immunological like domain (domain III) utilizing Pfam domain library. The domains I and II are corresponding to the Flavi_glycoprot Pfam domain whereas

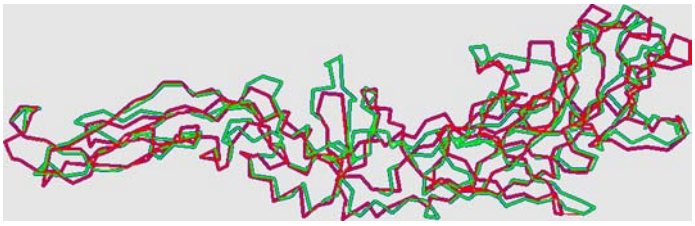

Fig. 4 Threading of the template sequence with the target structures

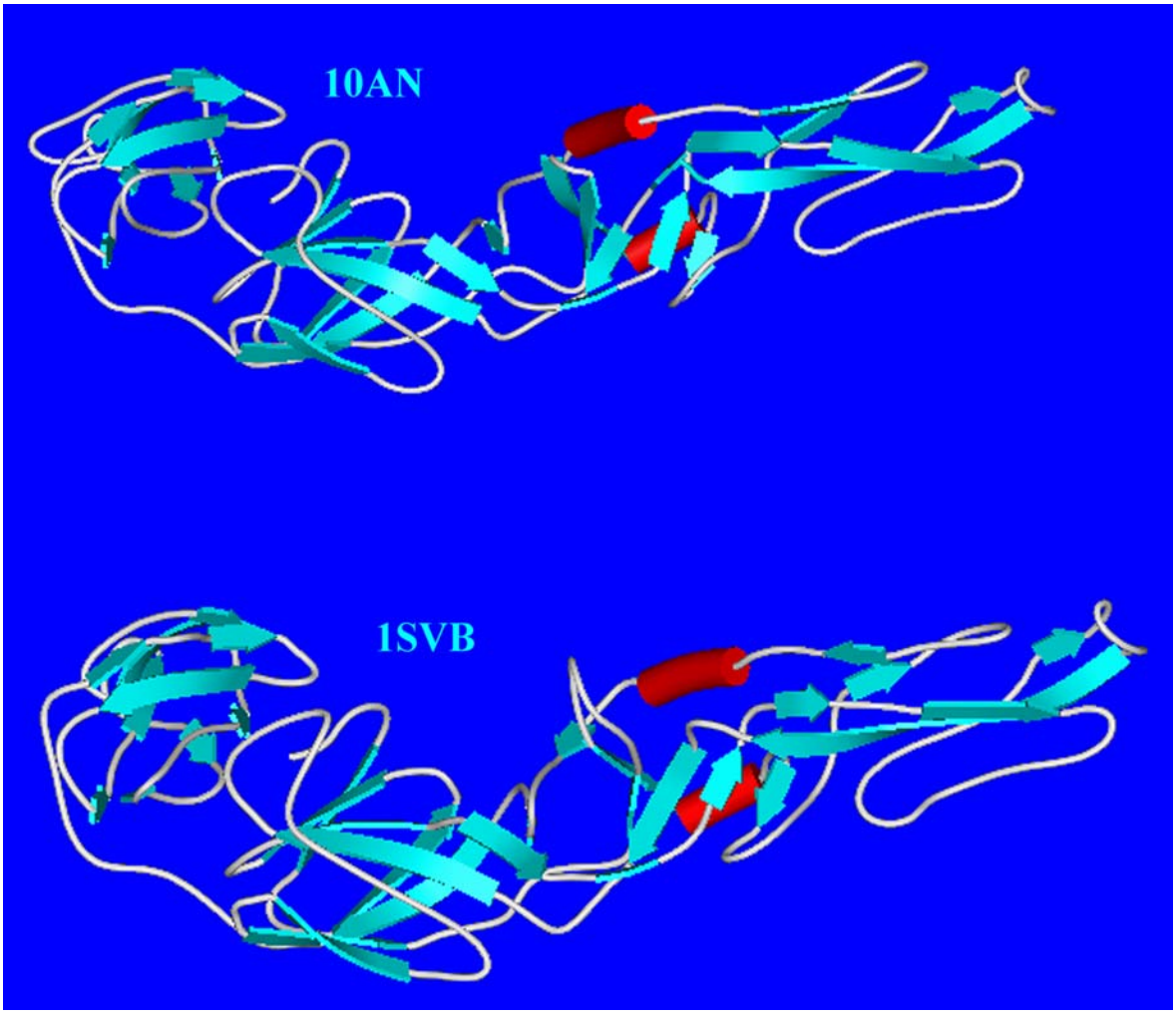

Fig. 5 The template structures that show common structural features between them

Flavi_glycop_C is corresponding to domain III (immunological domain) [20] (Fig. 9).

Domain I contains the glycosylation site in addition to the epitopes with serological or biological activities. Domain II is a hydrophobic region and the site of neutralization and hemagglutination. Domains I and II consist of 294 contiguous residues from 1 to 294 (Fig. 9).

The 12 cysteine residues, conserved in the flaviviridae envelope protein, are observed here in all the domains that add to the reliability of the model. There are five disulphide bridges (Cys^{3} –Cys^{30} ; Cys^{60} –Cys^{121} ; Cys^{74} –Cys^{105} ; Cys^{92} –Cys^{116} ; Cys^{186} –Cys^{284}) observed in the former two consecutive domains. The region also consists of 29 β-sheets with three short α-helices. The region

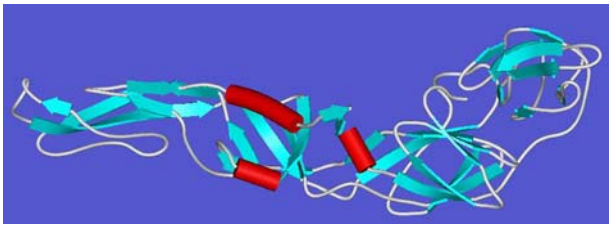

Fig. 6 The final model that is good in structural and functional aspects

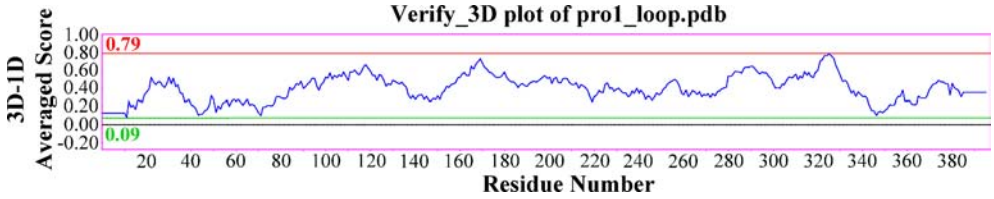

Fig. 7 3D–1D graph for the final model. For final model, it is 91.94% of the residues that has an averaged 3D–1D score >0.2. The VERIFY3D scores above the threshold of 0.2 indicate good local structural environments to individual residues

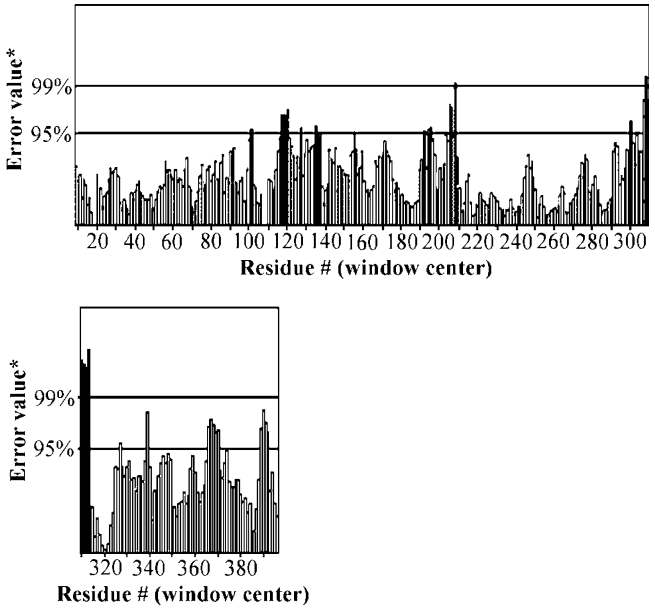

Fig. 8 ERRAT graphs for the model generated graph for the final model shows the great improvement in the quality of the model that has been brought by the loop modeling by scanning through loop database

96–112 lying in domain II is a highly conserved region among flaviviruses, has a β-hairpin motif and has been suggested to be involved in fusion activity [21].

Domain III is the immunological like domain with a continuous stretch of one hundred and two residues (295–396). One disulphide bridge (Cys^{301} –Cys^{332}) is observed in this domain. Domain III has also been suggested to be involved in receptor-binding activities. The above-described three domains can be related to antigenic domains of the EGP of Japanese encephalitis virus (JEV) [19].

Finally, the RMSD between the Egp of WNV and DEP (1OAN) is found to be 0.59 Å. The RMSD between the Egp of WNV and Egp of TBE (1SVB) is found to be 0.98 Å (Fig. 10). A favorable RMSD between the model and the templates should be around 1 Å [22]. The RMSD is calculated by considering all atoms of all the amino acids. The protein modeled shows a favorable RMSD with its targets.

The above results and observations have proved that the function of a protein can be chosen as criteria to determine its 3D-coordinates. The model is found to be an unusual extended structure with large number of β sheets. Moreover, the presence of 12 cysteine residues in the model, which is the characteristic of the flaviviridae family, has proved the structural alignment to be good. The model has been developed fully on the basis of domain strategy that is found to be fruitful in modeling a protein with low sequence similarity.

Prediction of conformational epitopes

The antigenic determinants (conformation epitopes) in Egp of WNV were predicted using its 3D-structure on the basis of an algorithm by Kolaskar and Kulkarni-Kale [19]. The promiscuous residues were obtained using a 5 Å distance criteria. According to this method, 80%

Table 3 ROCHECK report for the final model

Ramachandran plot	79.8% core	17.9% allowed	2.1% generous	0.3% disallowed
Chi1-chi2 plots	2 labeled residues (out of 211)			
Main-chain parameters	6 better	0 inside	0 worse	
Side-chain parameters	5 better	0 inside	0 worse	
G-factors	Dihedrals: −0.41	Covalent: 0.10	Overall: −0.20	
M/c bond lengths	99.9% within limits	0.1% highlighted		
M/c bond angles	94.4% within limits	5.6% highlighted		

The Ramachandran plot shows the satisfactory results, few amino acids in disallowed regions correspond to the loop regions and hence can be neglected. The chi1–chi2 plots show the chi1–chi2 torsion angle combinations for all residue types that have both these angles. The PROCHECK G factors for dihedral angles are −0.41. The overall average G factor is also closer to 0.0; the G factor is essentially a log-odd score based on the observed distributions of various stereo chemical parameters and indicates that the overall structure is stereo-chemically correct and satisfactory. The other parameters have also shown the model to be satisfactory on all grounds.

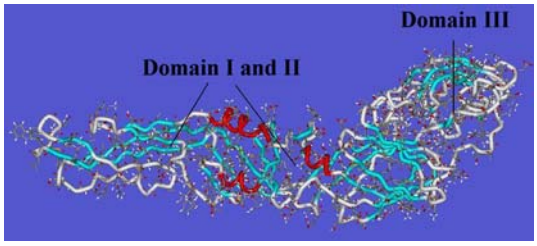

Fig. 9 The 3D structure of the model showing the three domains. They are antigenic in nature

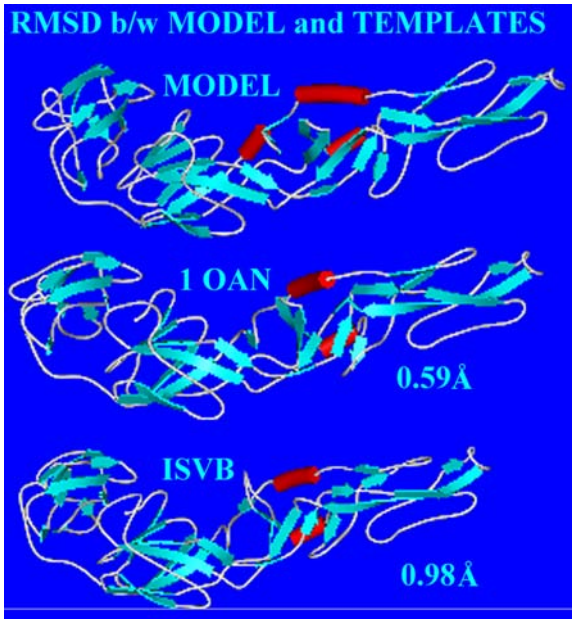

Fig. 10 The closer structural similarity between the templates and the target as shown by their RMSD value

of the residues that interact with the antibody have an accessible surface area (ASA) of ≥30%. The remaining 20% residues interact with the antibody even though their ASA is <30% (Table 4). A sum of five conformational epitope sites, namely sites I, II, III, IV and V, is observed on the surface (loop regions) of the coat protein of WNV. The sites I and II, i.e., ^{102}DRGWNGCGLF^{112}, ^{271}PVEFSSNTVT^{282} are located at a significant mutual distance and can be described as lying in domain I and at the hinge of domains II and III. On the other hand, the conformational location of sites III, IV and V almost overlaps, whereas their corresponding sequences and locations are ^{289}RVKMEK^{294}, ^{23}TWVDL^{27} and ^{183}YGEVTVDC^{190}, respectively (Fig. 11).

Table 4 Predicted epitopes on Egp of WNV

Conformational epitope sites on 3 D protein surface of Egp of WNV	Determinants that are part of conformational epitopes	Accessible residues that are within 5 Å from conformational epitopes
Site I	^{102}DRGWNGCGLF^{112}	C^{78}
Site II	^{271}PVEFSSNTVT^{282}	I^{134}, I^{139}, Y^{201}
Site III	^{289}RVKMEK^{294}	V^{203}
Site IV	^{23}TWVDL^{27}	I^{37}, S^{39}, K^{42}, P^{43}, I^{45}
Site V	^{183}YGEVTVDC^{190}	P^{43}, I^{45}, I^{145}

Conclusions

A similar strategy can be applied to model any target protein that has a low percentage of similarity with its homologue but has common functional domains. The protein thus modeled plays a vital role in the virulent activity of this virus and hence the predicted feasible 3D–model can very well act as a target in designing an antiviral drug against the structure of this pathogenic protein. In addition, vaccine-design strategies can be executed using knowledge of the conformational epitopes, which can be extracted with this 3D-model. The model can be studied further for its interaction with the antibody. This work can be well regarded as a means of improving the theoretical model through homology modeling with the help of domain strategy.

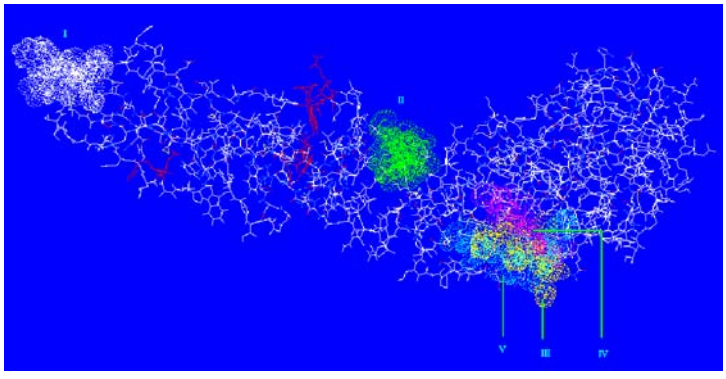

Fig. 11 The conformational epitopes that are distributed in all the domains. They are found out by the algorithm by Kolaskar and Kulkarni-Kale [19]

References

1. Xiong B, Gui C-S, Xu X-Y, Luo C, Chen J, Luo H-B, Chen L-L, Li G-W, Sun T, Yu C-Y, Yue L-D, Duan W-H, Shen J-K, Qin L, Shi T-L, Li Y-X, Chen K-X, Luo X-M, Shen X, Shen J-H, Jiang H-L (2003) Acta Pharmacol Sin 24:497–504
2. Rufino SD, Blundell TL (1994) J Comput Aided Mol Des 8:5–27
3. Chakrabarti S, John J, Sowdhamini R (2004) J Mol Model 10:69–75
4. Sampson BA, Armbrustmacher V (2001) Ann NY Acad Sci 951:172–178
5. Mandl CW, Guirakhoo F, Holzmann H, Heinz FX, Kunz C (1989) J Virol 63:564–571
6. Bairoch A, Apweiler R (1997) J Mol Med 75:312–316
7. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) J Mol Biol 215:403–410
8. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) Nucleic Acids Res 30:276–280
9. Guex N, Peitsch MC (1997) Electrophoresis 18:2714–2723
10. Kabsch W, Sander C (1983) Biopolymers 22:2577–2637
11. van Gunsteren WF, Berendsen HJC (1990) Angew Chem Int Ed Engl 29:992–1023
12. Peitsch MC (1995) PDB Q Newsl 72:4
13. Luthy R, Bowie JU, Eisenberg D (1992) Nature 356:83–85
14. Colovos C, Yeates TO (1993) Protein Sci 2:1511–1519
15. Laskowski RA, Moss DS, Thornton JM (1993) J Mol Biol 231:1049–1067
16. Hooft RWW, Vriend G, Sander C, Abola EE (1996) Nature 381:272–272
17. Scherret JH, Poidinger M, Mackenzie JS, Broom AK, Deubel V, Lipkin WI, Briese T, Gould EA, Hall RA (2001) Emerg Infect Dis 7(4):697–705
18. Choi J-H, Jung H-Y, Kim H-S, Cho H-G (2000) Bioinformatics 16:1056–1058
19. Kolaskar AS, Kulkarni-Kale U (1999) Virology 261:31–42
20. Rey FA, Heinz FX, Mandl C, Kunz C, Harrison SC (1995) Nature 375:291–298
21. Roehrig JT, Johnson AJ, Hunt AR, Bolin RA, Chu MC (1990) Virology 177:668–675
22. Chothia C, Lesk AM (1986) EMBO J 5:823–826