BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 9(9)

Insight from βC1 protein model for implication in cotton leaf curl disease

Khuram Shahzad[1], Abdul Hai[2], Nadeem Kizilbash[2]*, Jawaria Ambreen[3] & Jamal Alruwaili[2]

[1]Illinois Informatics Institute, University of Illinois, Urbana-Champaign, Illinois, U.S.A; [2]Department of Biochemistry, Faculty of Medicine & Applied Medical Sciences, Northern Border University, Arar-91431, Saudi Arabia; [3]Department of Chemistry, Quaid-i-Azam University, Islamabad-45320, Pakistan; Nadeem Kizilbash - Email: nadeem_kizilbash@yahoo.com; *Corresponding author

Received March 03, 2013; Accepted April 01, 2013; Published May 25, 2013

Abstract:

DNA β is approximately half of the size of Begomovirus DNA. It encodes a βC1 gene that is conserved in position and size. This gene has the capacity to encode a 13 to 14 kDa protein comprising 118 amino acid residues. It has been shown earlier that βC1 protein is necessary for inducing symptoms of cotton leaf curl disease. The structure for βC1 (CLCuDβ01-Pakistan) is still unknown. Therefore, a model of βC1 (CLCuDβ01-Pakistan) was developed using DoBo and I-TASSER servers followed by validation by PROCHECK and VERIFY 3D servers. The developed model provides an insight in a role for this multifunctional protein in causing Cotton Leaf Curl Disease (CLCuD). A possible function of this protein might be the suppression of RNA-silencing in cotton plants.

Keywords: Cotton Leaf Curl Disease, DNA β, βC1 gene, threading, protein structure prediction, RNA silencing.

Background:

The members of family Gemini viridae consist of single-stranded DNA (ssDNA) viruses that infect a wide range of plants and cause serious crop damage. One of the four genera of Gemini viruses is *Begomovirus* **[1]**. Most members of *Begomovirus* have bipartite DNA, called DNA A and DNA B. DNA A encodes the proteins required for viral DNA replication and encapsidation, whereas, DNA B encodes two proteins that are essential for systemic movement. Recently some whitefly-transmitted *Begomovirus* have been shown to require the presence of single-stranded DNA satellite (known as DNA β) to induce characteristic symptoms of Cotton Leaf Curl Disease (CLCuD) in some hosts **[2, 3]**.

Approximately half of the *Begomovirus* DNA consists of DNA β (1.3 to 1.4 kb) **[2, 4]**. Sequence analyses have shown that the complementary-sense strand of all DNA β molecules encode a βC1 gene which is conserved in sequence, position and length. This gene has the capacity to encode 13 to 14 kDa protein comprising of 118 amino acids **[5, 6]**. The precise function of DNA β and the βC1 protein in the pathogenesis of CLCuD is still not fully understood. It has been proposed that DNA β may play a direct or an indirect role in viral DNA replication, facilitation of the movement of viruses or countering the host defense response **[4]**.

The functions of DNA β have been shown to be mediated by complementary-sense gene, *βC1*. The protein product of *βC1* gene has been shown to act as a suppressor of post-transcriptional gene silencing **[5-9]**. The DNA β-encoded protein, βC1, is the cause of both pathogenicity and suppression of gene silencing **[10]**. In this study, the 3D structure of CLCuDβ01-Pakistani protein has been predicted by use of the online structure prediction server, called I-TASSER. DoBo server was used to predict the domains of the protein. PROCHECK and VERIFY 3D servers were used for evaluation of the predicted βC1 structure. The final 3D model was evaluated in terms of functional capability of the protein.

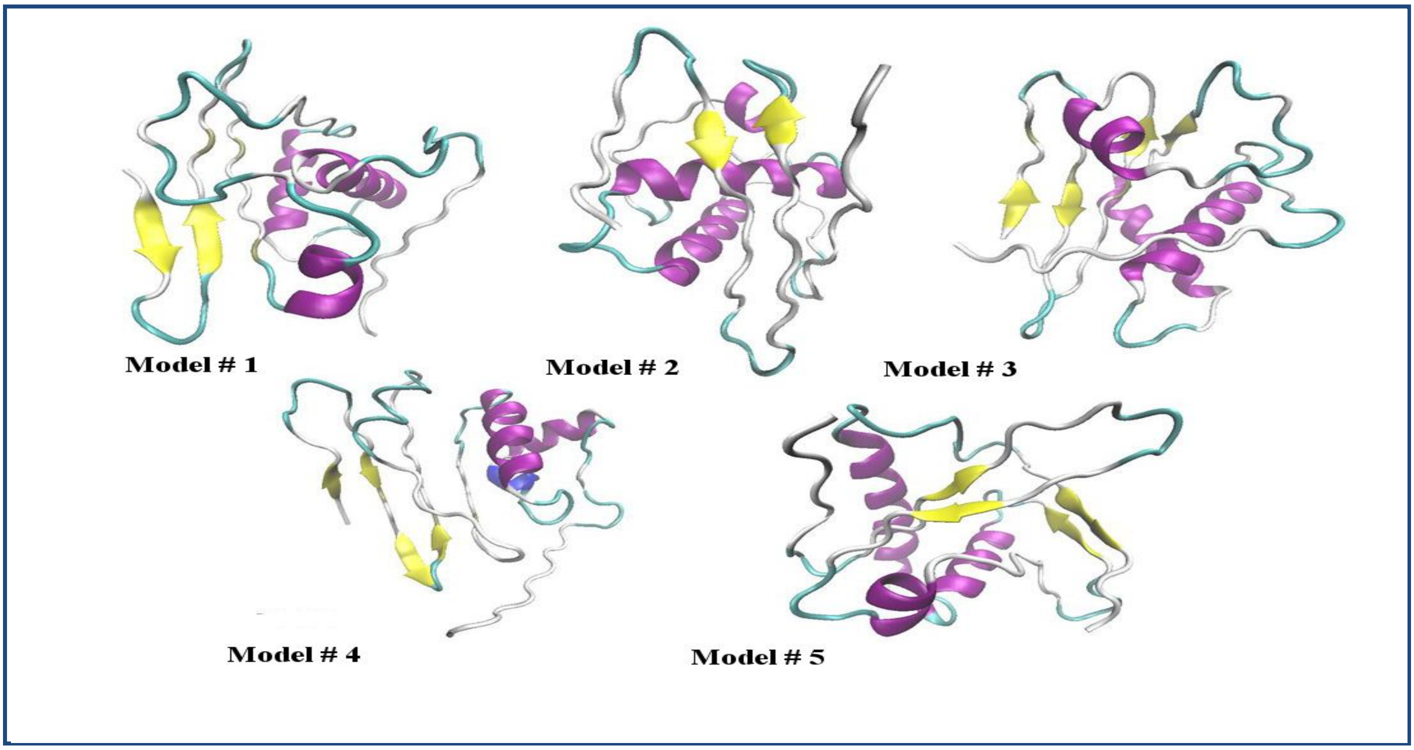

Figure 1: The five models for tertiary structure of CLCuDβ01-Pakistani protein predicted by the I-TASSER server. Each model is represented by α-helices (*purple* colored) and β-strands (*yellow* colored).

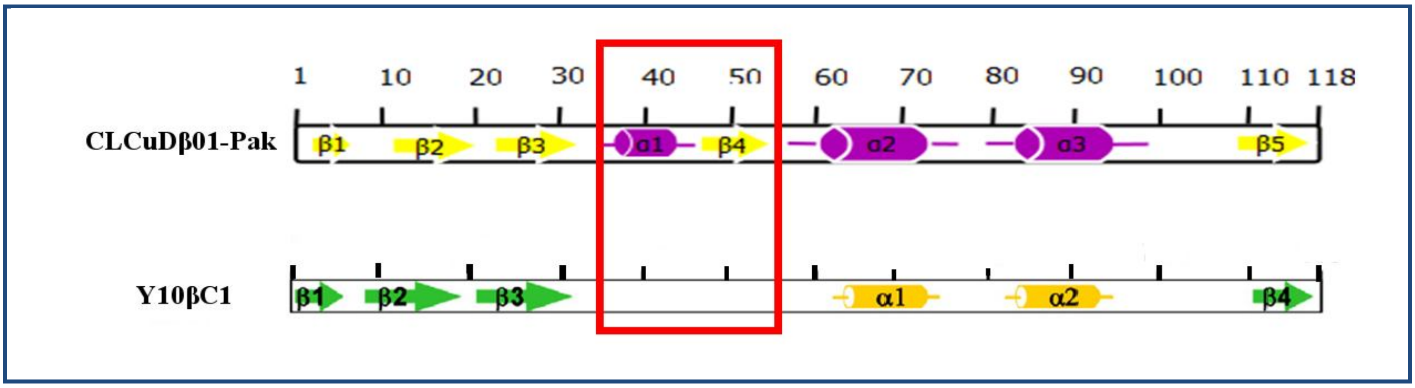

Figure 2: Schematic comparison of predicted structural elements of CLCuDβ01-Pakistani protein and Chinese viral protein Y10βC1 **[18]**. The β-strands and α helices are shown by *yellow* arrows and *purple* colored bars respectively for βC1 (Pakistan). While in Chinese Y10βC1 sequence, the β-strands and α-helices are shown by *green* arrows and *yellow* bars respectively. The part of the amino acid sequence where structural differences are found between the two proteins is outlined by a *box*.

Methodology:

The amino acid sequence of CLCuDβ01-Pakistani protein was retrieved from ExPASy Bioinformatics resource portal (http://expasy.org) with accession number (UniProtKB = Q911I3). The nucleotide sequence was accessed through EMBL nucleotide sequence database (Accession number: AJ292769). The nucleotide sequence is 357 base pairs long, which encodes a protein of 118 amino acid residues. The primary sequence of 118 amino acids was used to unravel the structural aspects of this protein. Both structure prediction and evaluation tools (data not shown) were used, but the following tools were relied upon more for greater accuracy. The secondary structure elements were determined using the DoBo **[11]** and PredictProtein (results not shown) **[12]** severs. For protein homology modeling, we used online available tool called I-TASSER **[13]**. This server uses the threading technique to predict the 3D models. The server generated 5 best models based on multiple-threading alignments and iterative template fragment assembly simulations along with their confidence scores **(Figure 1).** The 5 models were visualized by the Visual Molecular Dynamics (VMD) software **[14]**. To evaluate these models, different validation techniques were used. In a similar fashion, PROCHECK **[15]** and VERIFY 3D **[16]** servers were used to validate the predicted protein structures. The PROCHECK software generates ramachandran plot which nicely explains the steriochemical configuration of amino acid residues. The VERIFY 3D analyzes the compatibility of an atomic model with its amino acid sequence. Each amino acid residue is assigned a structural class. A collection of structures is used as a reference to obtain a score for each of the 20 amino acids in any structural class. The scores are then plotted for individual residues **[17, 18]**. Finally, the better model was evaluated based on the aforementioned tools.

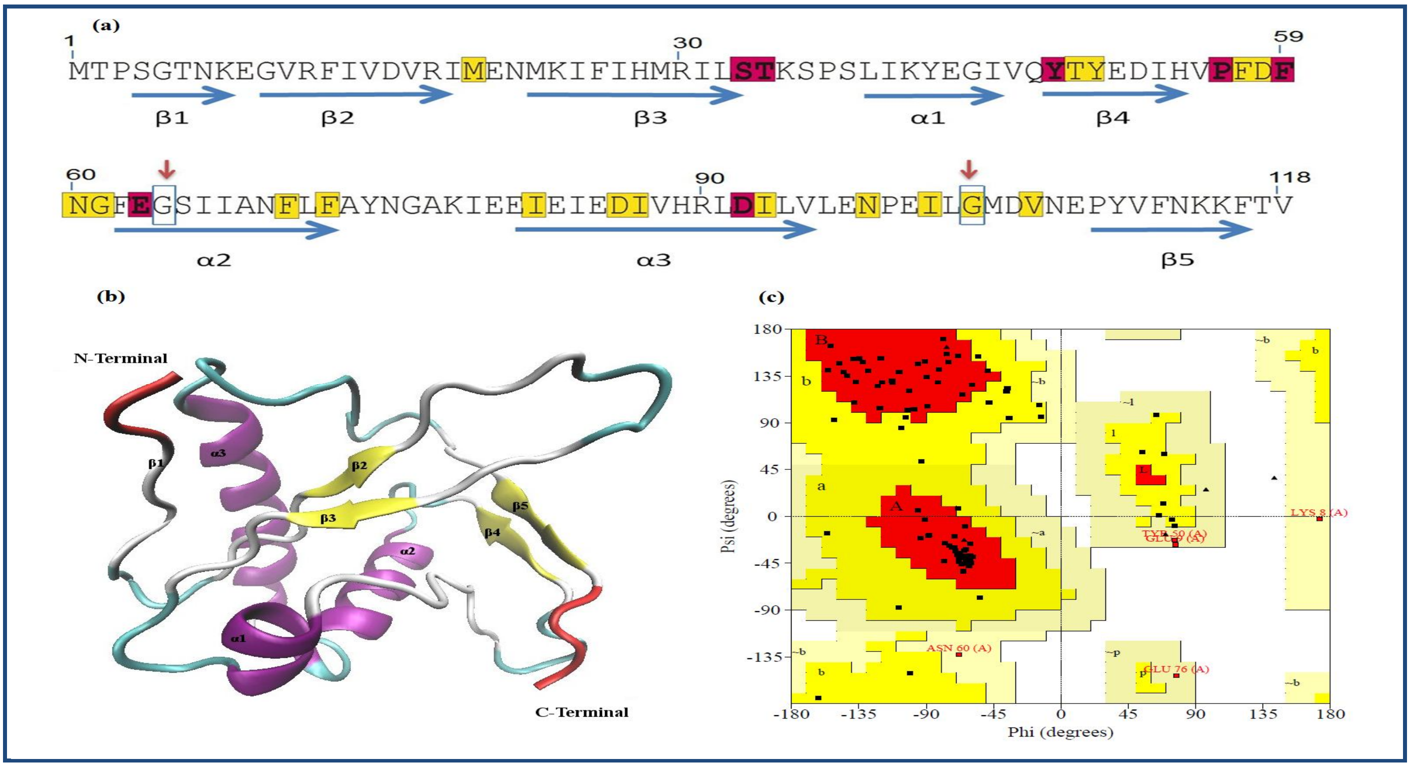

Figure 3: (a) Amino acid sequence and secondary structure elements of the CLCuDβ01-Pakistani protein. The consensus secondary structure elements were determined by use of DoBo, PredictProtein and I-TASSER servers. Highlighted in *red* are amino acids that are highly conserved between Malvaceous β satellites; those in *yellow* are the amino acids that are highly conserved between all the β satellites; *arrows* indicate the position of Glycine and Proline residues in the secondary structure elements. **(b)** Model number 5 for the tertiary structure of βC1 predicted protein by I-TASSER server. The α-helices are represented by *purple* color, while β-strands are represented by *yellow* color. N- and C-terminus residues are colored *red*. **(c)** Ramachandran plot of the predicted protein model of βC1 showing the values of Psi and Phi angles.

Results & Discussion:

The CLCuDβ01-Pakistani protein and Chinese Y10βC1 are distantly related from each other with respect to amino acid sequence. However, when the secondary structure elements of CLCuDβ01-Pak were compared to Chinese Y10βC1, the structural conservation was observed between the two species **[19] (Figure 2)**. Since no template was found for homology modeling of CLCuDβ01-Pakistani protein, it was decided that threading technique can be useful for tertiary structure prediction. This guided us towards the implementation of the threading-based I-TASSER server. Five models of the tertiary structure were received from the server **(Figure 1).** These models were verified using PROCHECK and VERIFY 3D online available servers. The overall percentage representation of results is shown in **Table 1 (see supplementary material).** The threading-based method was successful in determining the 3D structure of the CLCuDβ01-Pakistani protein and it provided five complete models. Model numbers 3 and 5 had good scores in terms of Ramachandran plots and 3D-1D amino acid distributions. Both the models passed the percentage representation and averaged 3D-1D scores. Model number 5 was selected as the better model because of the presence of larger percentage of amino acid residues in the core and allowed regions in the Ramachandran plot. None of its amino acid residues were present in the forbidden or disallowed region of the Ramachandran plot as shown by **(Table 1)** and **(Figure 1c)**. The accuracy of the model was also verified by use of the criteria developed by VERIFY 3D server. This server verifies the 3D structural distribution of amino acids as compared to the 1D distribution of amino acid residues [**16**].

The selected model number 5 showed structural conservation between the CLCuDβ01-Pakistani protein and Chinese Y10βC1 for two α-helices located near the C-terminus **(Figure 2)**. These α-helices have been implicated in the multimerization of the protein **[19]**. From the amino acid sequence analysis it was observed that CLCuDβ01-Pak lacks cysteine residues. Therefore, a zinc finger DNA-binding domain (Cys-His motif) is missing from the protein structure **[20, 21, & 22]**. However, a Histidine-based DNA-binding domain might still be present in the protein that may allow it to bind single stranded or double stranded DNA without size or sequence specificity and to be able to suppress the host RNA silencing activity as observed in Chinese strain **[8]**. The exact sequence location in CLCuD β01 protein still needs experimental verification.

The secondary structure elements of CLCuDβ01-Pakistani protein **(Figure 3a)** comprise three α-helices and five β-strands. Three of the β-strands (β_1, β_2 and β_3) are at the N-terminus, the fourth β-strand (β_4) is located almost in the middle of the of sructure and the fifth β-strand (β_5) is present near the C-terminus. All of the three α-helices (α_1, α_2 and α_3) are located in the middle of the protein structure. Five Glycine residues present in the secondary strcuture elements. Glycine 5 is present in the middle part of β_1; Glycine 10 is at the beginning of β_2; and Glycine 44 and Glycine 64 are present in α_1 and α_2 respectively. Only one Proline residue is present in the secondary structure elements (Proline 109 in β_5). Glycines and Prolines can both produce a "kink" in α-helices and β-strands. The amino acids residues: Alanine, Aspartic Acid, Glutamic acid, Isoleucine, Leucine and Methionine favor the formation of α-helix. The three α-helices of CLCuDβ01-Pakistani protein contain a total of 31 amino acid residues out of which 18 belong to the type that favor formation of α-helices. The secondary structure elements were further compared with those of the Chinese viral protein Y10βC1 **[19]** **(Figure 2).** Both the proteins were found to be comparable. The only region showing any difference between the two proteins is the region between amino acid 35-55 where an extra α-helix and a β-strand are present in CLCuDβ01-Pakistani strain.

The predicted tertiary structure of CLCuDβ01-Pakistani protein shows three α-helices and five β-strands. The four β-strands are arranged antiparallel to each other **(Figure 3b).** According to the results provided by the DoBo server, CLCuDβ01-Pakistani protein contains two domains: an N-terminal domain which stretches from amino acid 1-51 and contains three β-strands and one α-helix and a C-terminal domain which stretches from amino acid 55-118 and one β-strand and two α-helices. Both the domains are critical for functioning of the protein [**10**].

Figure 3 shows the secondary elements distribution across the primary sequence (a), the predicted 3D model (b) and Ramachandran plot (c) for the amino acids distribution. To analyze the stereochemical quality of the predicted structure PROCHECK software was used. According to PROCHECK results, the fifth model (model#5) **(Figure 3b)** seems the most appropriate one because it has most of the amino acid residues present in the core and allowed regions (92.2%) while only 4.7 % of the total amino acids are found in the generous region as indicated by Ramachandran plot **(Figure 3c).** Further, the quality was assessed by the VERIFY 3D software. Both the PROCHECK and VERIFY 3D results are shown in the **Table 1 (see supplementary material)**. The **Figure 3** shows the secondary elements distribution across the primary sequence (a), the predicted 3D model (b) and Ramachandran plot for the amino acids distribution (c).

The molecular basis of pathogenicity of βC1 can be explained by the suppression of RNA silencing activity. RNA silencing is a surveillance system that exists in many species under different names but with same phenomena e.g., in animals (RNA interference), fungi (quelling) and plants (post-transcriptional gene silencing) **[22, 23]**. RNA silencing, either initiated or inhibited by different viruses, plays an important antiviral role in eukaryotes (e.g. animals and plants etc.). Some viruses have evolved or acquired functional proteins that suppress RNA silencing by targeting different steps of silencing pathways **[7, 24, 25]**. This role has only recently been elucidated in βC1 protein of tomato yellow leaf curl China betasatellite (TYLCCNB) which forms a multimeric complex with the help of cysteine residues for its proper functioning **[18]**. However, this type of role in CLCuDβ01 has not been experimentally elucidated yet. βC1 is also involved in other functions also such as pathogen (virus) movement in host plants, DNA-binding and post-transcriptional gene silencing **[26]**.

It is known that expression of βC1 interferes with local silencing in transient Agrobacterium-based assays. βC1 protein targets different stages in silencing process by overlapping the miRNA. It binds both single stranded and double stranded DNA in a non-specific manner without the presence of a zinc finger domain **[8]**. The βC1 fusion proteins have been shown to be primarily localized in the nucleus in both insect and plant cells. They require a nuclear localization sequence (NLS) for entering the nucleus of a cell **[8]**. Although comparable with other Begomovirus proteins such as AL2/AC2, with respect to size, DNA-binding properties, and nuclear localization, βC1 lacks the zinc finger motif and shares little or no sequence homology with these proteins **[27]**. The predicted structure can help virologists in pinpointing the major regions of interaction between plant hosts and viruses to uncover the interaction mechanism. Future research in the direction of predicting the binding sites will also help uncover the possible mechanisms of RNA-silencing and several other functions. By knowing the structural aspects of this protein, it will be easier to target the disease causing viruses by introducing the novel drugs inside the plants or using the gene therapy techniques in plant to eliminate the disease effects and to increase the potential economic growth at industrial level.

Conclusion:

The secondary and tertiary structures of CLCuDβ01-Pakistani protein associated with Cotton Leaf Curl Disease CLCuD have been predicted using *in silico* methodologies. The novel aspects of the protein structure have been highlighted using already available literature. The secondary structure elements were compared with the Chinese viral protein Y10βC1 which revealed that both the proteins are structurally somewhat similar despite sequence dissimilarities. The only difference is in the region between amino acid 35-55 where an extra α-helix and a β-strand are present in CLCuDβ01-Pakistani protein. The novel structural aspects can be further used to highlight the potential role of this protein inside the host to unravel the potential role of the disease causing protein inside the host. The future advancement such as protein-protein interactions, role in DNA binding and RNA interference need expert level approaches to investigate the structural and functional aspects of this protein and to control its role in spreading of Cotton Leaf Curl Disease (CLCuD).

References:

[1] Borah BK & Dasgupta I, *J Biosci.* 2012 **37**: 791 [PMID: 22922204]

[2] Sattar MN *et al. J Gen Virol.* 2013 **94**: 695 [PMID: 23324471]

[3] Jose J & Usha R, *Virol.* 2003 **305**: 310 [PMID: 12573576]

[4] Saunders K *et al. Proc Natl Acad Sci. USA* 2000 **97**: 6890 [PMID: 10841581]

[5] Saunders K *et al. Virol.* 2004 **324**: 37 [PMID: 15183051]

[6] Zhou X *et al. J Gen Virol.* 2003 **84**: 237 [PMID: 12533720]

[7] Voinnet O, *Trends Genet.* 2001 **17**: 449 [PMID: 11485817]

[8] Cui X *et al. J Virol.* 2005 **79**: 10764 [PMID: 16051868]

[9] Gopal P *et al*. *Virus Res*. 2007 **123**: 9 [PMID: 16949698]

[10] Briddon RW & Stanley J, *Virol*. 2006 **344**: 198 [PMID: 16364750]

[11] Eickholt J. *et al*. *BMC Bioinformatics*. 2011 **12**:43 [PMID: 21284866]

[12] Rost B et al. *Nucl Acids Res*. 2004 **32**:W321 [PMID: 15215403]

[13] Zhang Y, *BMC Bioinform*. 2008 **9**: 40 [PMID: 18215316]

[14] Humphrey W *et al*. *J Molec Graph*. 1996 **14**: 33 [PMID: 8744570]

[15] Laskowski RA *et al*. *J Biomol NMR*. 1996 **8**: 477 [PMID: 9008363]

[16] Eisenberg D *et al*. *Methods Enzymol*. 1997 **277**: 396 [PMID: 9379925]

[17] Bowie JU *et al*. *Science*. 1991 **253**: 164 [PMID: 1853201]

[18] Lüthy R *et al*. *Nature*. 1992 **356**: 83 [PMID: 1538787]

[19] Cheng X *et al*. *Virol*. 2011 **409**: 156 [PMID: 21035158]

[20] Hartitz MD *et al*. *Virol*. 1999 **263**: 1 [PMID: 10544077]

[21] Kirthi N & Savithri HS, *Arch Virol*. 2003 **148**: 2369 [PMID: 14648292]

[22] Bisaro DM, *Virol*. 2006 **344**: 158 [PMID: 16364747]

[23] Li HW *et al*. *Science*. 2002 **296**: 1319 [PMID: 12016316]

[24] Baulcombe DC, *Curr Biol*. 1999 **9**: R599 [PMID: 10469584]

[25] Roth BM et al. *Virus Res*. 2004 **102**: 97 [PMID: 15068885]

[26] Voinnet O *et al*. *Proc Natl Acad Sci USA*. 1999 **96**: 14147 [PMID: 10570213]

[27] Nawaz-ul-Rehman MS *et al*. *Virol*. 2010 **405**: 300 [PMID: 20598726]

[28] Sunter G *et al*. *Virol*. 2001 **285**: 59 [PMID: 11414806]

Edited by P Kangueane

Citation: **Shahzad *et al.*** Bioinformation 9(9): 471-476 (2013)

Supplementary material:

Table 1: Evaluation results of the I-TASSER models of the tertiary structure by PROCHECK and VERIFY 3D

		Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
PROCHECK	**Core Region**	70.8%	76.4%	76.4%	74.5%	72.6%
	Allowed Region	18.9%	17.0%	17.0%	17.9%	22.6%
	Generous Region	7.5	4.7%	2.8%	2.8%	4.7%
	Disallowed Region	2.8	1.9%	3.8%	4.7%	0.0%
***VERIFY 3D**		54.62%	78.99%	95.80%	50.42%	83.19%

* These models had an averaged 3D-1D score > 0.2.