Bangladesh J Microbiol, Volume 31, Number 1&2, June-December 2014, pp 53-58

Original Article

In Silico Analysis: Annotations about Structural and Functional Features of DUF 2726 Family Member Proteins

SM Sabbir Alam, M Ruhul Amin, M Anwar Hossain*

Department of Microbiology, University of Dhaka, Dhaka 1000, Bangladesh

Domains of unknown functions (DUFs) are a big set of protein families within the Pfam database that includes proteins of unknown function. In the absence of functional information, proteins are classified into different families based on conserved amino acid sequences and are potentially functionally important. In Pfam database, the numbers of families of DUFs are rapidly increasing and in current the fraction of DUF families had increased to about twenty two percent of all protein families. In this study we targeted DUF2726 member proteins which are mainly present in different bacterial species of Gamma-proteobacteria and have a particular domain organization. We analyzed the protein sequences of domain DUF2726 using different computational tools and databases. We found that this domain contains a nuclear localization signal peptide, which is conserved in *Escherichia* spp. and *Shigella* spp. It were also predicted that it has nucleic acid binding properties. Analyzing protein-protein interactions functional partners associated with DUF 2726 were revealed. Protein secondary structure, transmembrane helices structure were predicted. We have found that it has gene neighbourhood and co-occurrences with protein RepA and RepB. RepA and RepB are functionally associated with replication. RepA is a replication protein and RepB is a replication regulatory protein. Presence of a nucleic acid binding properties, a nuclear localization signal (NLS) signalling peptide, and possible interaction pattern with replication proteins, conjectures its possible role as a NLS like signalling peptide.

Keywords: Domain of unknown function, DUF2726, Protein function prediction, NLS-like protein, Signalling peptide.

Introduction

Pfam is a conserved protein family database and is widely used to annotate and classify proteins[1-2]. Pfam has a database of curated multiple sequence alignments for each of its all families. It also contains structural and functional annotations, cross-database links and citation references for each group. For each Pfam family there have two multiple alignments: (1) the seed alignments which contains a little number of representative members, and (2) the full alignment containing all members in the database[2-3].

The current version of Pfam contains about 14,831 protein families. Compared with the previous version there have an increase of 1,159 families and 16 new clans. Entire sequence database (Pfam 23.0) contains about 5.3 million amino acid sequences. And this number is increasing rapidly as growing efficiency of genomic and metagenomic sequencing projects[2-3].

DUFs, are a large set of families within the Pfam database that includes only proteins of unknown function. The Pfam database contains over 3,000 such families and the number of families of DUFs is quickly increasing. It represents about 20% of known protein families. Some DUF families are predicted to contain more than one protein domains[1,4-6]. Using different database information and alignment tools it was found that, 25% of the DUF families are actually linked to some previously characterized domain groups. Structural comparisons confirm these distant relationships. Automated structure comparisons with manual structure analysis it was found that these families can be classified into known protein folds[5,7].

To gain the ultimate goal of systems biology it is important to know the function of all con-stituent parts of an organism. It was found that even extensively studied organisms can have many proteins unidentified or functionally unclassified. In yeast *Saccharomyces cerevisiae* have around 1,000 proteins (~17% of the genome) are still uncharacterized[8].

In practical there have three ways to determine the function of an uncharacterized domain – (1) identifying similarity to a known function domain, either by sequence comparison or by structural analysis, (2) using contextual information such as genomic context to computationally identify function (*e.g.*, STRING[9] and PROLINKS[10]), and (3) using old-fashioned molecular biology or biochemical techniques[4]. Various tools and databases are now available to identify the relationships between DUFs and other known families. Profile-HMM comparison tools, *e.g.*, *HHsearch*[11],

Corresponding author:
M Anwar Hossain, Department of Microbiology, University of Dhaka, Dhaka 1000, Bangladesh
Tel: +880 (02) 9661920-73, Ext 7735/7730; E-mail: hossaina@du.ac.bd

PRC[12], *SIMPRO* [13] and *SCOOP* [14], have proved to be very helpful in this perspect [4].

DUF 2726 family member proteins are mainly present on many different bacterial species, especially of class Gamma-proteobacteria and have a particular domain organization. This domain is introduced in Pfam Release 24.0 in 2009[2] and is still uncharacterized. A large portion of sequences of this domain have came from organism *Escherichia* spp. and *Shigella* spp. Finding relevant to our work we studied this group to predict its function. We analyzed DUF2726 family member proteins *in silico*, using different computational tools and databases to predict its putative function.

Materials and Methods

Data

Domain information of DUF 2726 was found from Pfam database. For analysis accession Q99QC5 was used as a sample sequence. It is an uncharacterized protein of *Shigella flexneri* encoded in gene *yihA*. Sequence data was downloaded from database UniprotKB[15].

Function prediction of protein using PFP and SVM prot

Elementary functional prediction of protein was done by using PFP[16] and SVM prot[17]. PFP (Protein Function Prediction) is a web-server for prediction of protein function developed by Kihara Lab at Purdue University. It results most probable protein function (GO terms) to given protein sequence based on its novel algorithm[16,18]. SVM prot classifies proteins into functional family from its primary sequence. SVM prot classification system is trained from representative proteins of a number of functional families and seed proteins of Pfam curated protein families[17].

Search for nuclear localization signal

PredictNLS server[19] was used to determine the presence of nuclear localization signal. NLSdb is a comprehensive source of information regarding NLSs and proteins translocated into the nucleus by signal sequences. Targeting signal recognition is a key control point in the regulation of nuclear transport[19]. It is therefore a used for identifying targeting signals in their sequence.

Conservation analysis

Jalview[20] was used to see conserveness of this localization signal among domain of other sequences. It is a tool that is used to view and edit multiple sequence alignments.

Transmembrane helices

TMHMM[16] was used to determine the presence of transmembrane helices. TMHMM is a web based server developed to predict TM helices and their topologies. It is a method based on support vector machines (SVM) in a hierarchical framework to predict TM helices first, followed by their topology prediction[21].

Secondary structure analysis

Secondary structure was analyzed using dompred[22]. DomPred is a server designed to predict putative protein domains and their boundaries for a given protein sequence[23].

Analysis for functional partners

STRING[24] was used to determine functional partners of this protein. The database STRING is a precomputed global resource for the exploration and analysis of protein-protein associations. It is used to retrieve and display the repeatedly occurring neighbourhood of a gene. The repeated occurrence of genes in each other's neighbourhood on genomes has been shown to indicate a functional association between the proteins they encode[25].

Results

Initial functional features were predicted by using PFP and SVM prot. PFP result shows a different number of possible functional classifications of protein (Table 1). PFP result showed that this protein group may belong to many functional protein families including, nucleotide binding properties. Further analysis was done to verify and predict the correct functional properties of this protein group. SVM prot shows similarity of this protein with a number of protein families include: transmembrane metal-binding, nucleotide binding, hydrolases – acting on halide bonds, DNA repair, copper-binding, calcium-binding, magnesium-binding etc.

Presence of nuclear localization signal was analyzed using predictNLS server. It was found that nuclear localization signal was present in input sequence. Using Jalview it was found that this nuclear localization sequence is conserved in Domain organization of *Shigella* spp. and *Escherichia coli.* Conserved nuclear localization sequence is ERRRRD. It is present in 176-181 a.a sequence. The conserved sequence was found from Jalview shown in Figure 1.

TMHMM result showed that input sequence may contain one transmembrane helics. It has predicted that amino acid sequence from 1-2 are present inside, 3-22 are present in transmembrane region and 23-196 are present in outside region (Figure 2). DomPred showed that DUF2724 contains more helix-residues than strand and coil-residues. By using DomPred a probable secondary structure of this domain group was predicted (Figure 3 and Figure 4).

STRING was used to predict functional partners of this protein (Table 2). Hypothetical functional partners are shown in Figure 5. Among the functional partners of this protein, repA and repA2 are functionally associated with replication. RepA (285 aa) is a replication protein and RepB (86 aa) is a replication regulatory protein. RepA is of two types, repA2 and repA. repA2 is involved in the determination of copy number in gene replication and RepA play role as initiator protein for mRNA synthesis.

Table 1. *Functional classification of protein from its primary sequences using PFP*

Domain name	PFP result		
	Molecular function terms	Biological process terms	Cellular component terms
	Probability – Description	*Probability – Description*	*Probability – Description*
DUF 2726	77% – Melanocortin receptor activity	65% – Cellular metabolism	75% – Voltage-gated sodium channel complex
	67% – Melanocyte stimulating hormone receptor activity	60% – Cellular physiological process	60% – Cell
	58% – Ion binding	56% – Physiological process	59% – Intracellular
	57% – Binding	56% – Cellular macromolecule metabolism	59% – Membrane
	57% – Cation binding	55% – Pyruvate family amino acid metabolism	56% – Cytoplasm
	56% – Transition metal ion binding	54% – Cellular process	56% – Iintracellular organelle
	52% – Metal ion binding	54% – Primary metabolism	55% – Plasma membrane
	50% – Protein binding	53% – Macromolecule metabolism	54% – Nucleus
	50% – Nucleotide binding	53% – Metabolism	53% – Intracellular membrane-bound organelle
	49% – Receptor binding	52% – Cellular biosynthesis	53% – Exosome (RNase complex)
	49% – Nucleic acid binding	52% – Biopolymer metabolism	52% – Intracellular non-membrane-bound organelle
	49% – Adenyl nucleotide binding	51% – Cellular protein metabolism	
	46% – Peptide receptor activity	49% – Amino acid and derivative metabolism	52% – Cytoskeleton
	46% – Cytoskeletal protein binding	49% – Nucleobase, nucleoside, nucleotide and nucleic	43% – Integral to plasma membrane
	45% – Peptide binding	acid metabolism	39% – Integral to membrane
	43% – DNA binding	47% – Amine metabolism	35% – Non-membrane-bound organelle
	43% – Unfolded protein binding	46% – Organic acid metabolism	33% – Microtubule cytoskeleton
	41% – Purine nucleotide binding	45% – Cellular catabolism	28% – Membrane-bound organelle
	40% – RNA binding	45% – Amine biosynthesis	28% – Aminin-1
	40% – Hormone binding	44% – Transport	28% – Aminin complex
	37% – Calcium ion binding	43% – Nitrogen compound biosynthesis	27% – Organelle
	35% – Catalytic activity	43% – Cell motility	20% – Microtubule
	34% – Actin binding	42% DNA metabolism	
	33% X-Pro dipeptidyl-peptidase activity	41% – Locomotion	
	33% – Peptide receptor activity, G-protein coupled		
	29% – Oxidoreductase activity		

*Initial functional features of query protein were predicted by using PFP and SVM prot. PFP result showed a different number of combinatorial possible functional classifications of protein. It showed that this protein group may belong to many functional protein families including, nucleotide binding properties.

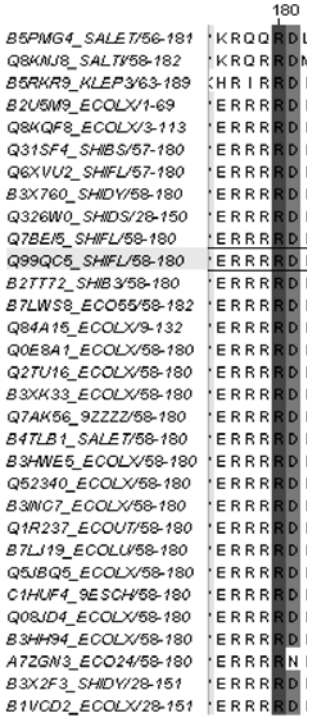

Figure 1. *Conserved NLS signal on domain of DUF2726. The conserved sequence of NLS signal (ERRRRD) was found to be present in 176-181 aa in domain organization of Shigella spp. and Escherichia coli. Conserved NLS signal is shown in multiple sequence alignment using Jalview.*

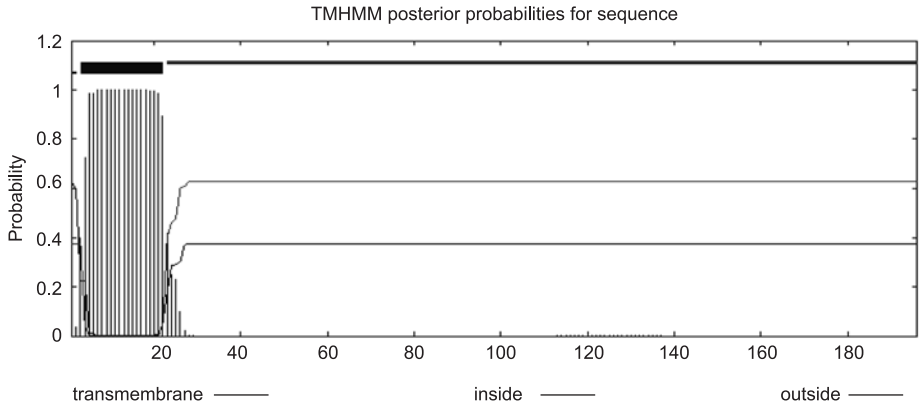

Figure 2. *Results from TMHMM for DUF 2726. TMHMM was used to determine the amino acid positions that are inside/outside of the transmembrane helices. It was found that amino acid positions from 23-196, that covers the NLS signal is located outside of the transmembrane region.*

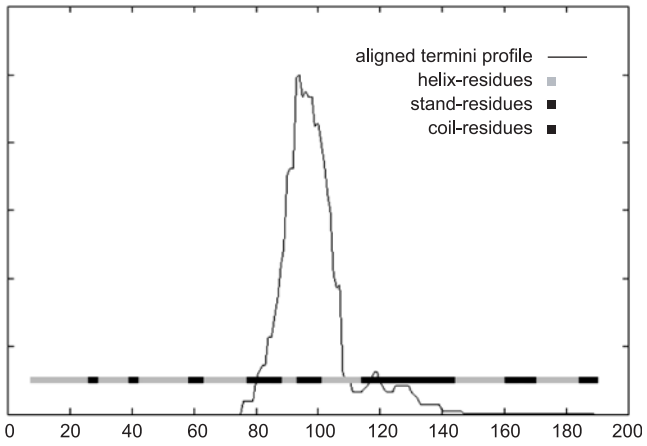

Figure 3. *Predicted secondary structure of protein Q99QC5. Secondary structure was predicted using dompred. It showed that DUF2724 contains more helix-residues than strand and coil-residues. The graph is derived by dompred from the N- and C-termini positions from PSI-BLAST local alignments. Large values indicate positions where sequence discontinuities occur, thus locates putative domain boundaries. A putative domain is predicted with the domain boundaries at residue positions of 100.*

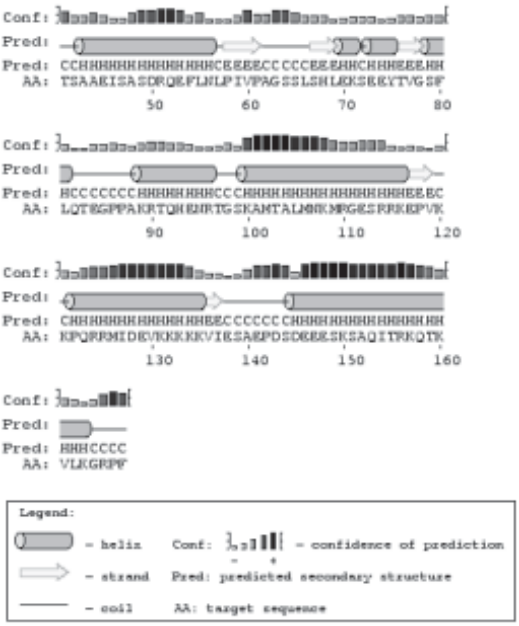

Figure 4. *PSI-BLAST alignment profile. Predicted secondary structure and PSI-BLAST alignment profile for Q99QC5 showing helix, strand and coil portions of different amino acids and predicted confidence values.*

Table 2. *List of predicted functional partners found on STRING*

Predicted Functional Partners:		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
ccdB	plasmid maintenance protein CcdB (101 aa)			•				•		0.847
ccdA	plasmid maintenance protein CcdA (72 aa)			•				•		0.814
traX	conjugal transfer pilus acetylation protein TraX (248 aa)	•		•						0.786
repB	replication protein; This protein is involved in the determination of copy number in gene repli [...] (86 aa)	•						•		0.763
stbA	plasmid stable inheritance protein (319 aa)			•						0.712
yccB	hypothetical protein (323 aa)			•						0.701
SF2042	hypothetical protein (92 aa)			•						0.681
repA	replication protein (285 aa)	•		•						0.634
ORF47	hypothetical protein (160 aa)							•		0.609
S0240	putative coat protein (431 aa)							•		0.609

* Hypothetical functional partners of YihA revealed by STRING. Functionally important protein- RepA and RepB are matched by gene neighborhood and co-occurrence pattern.

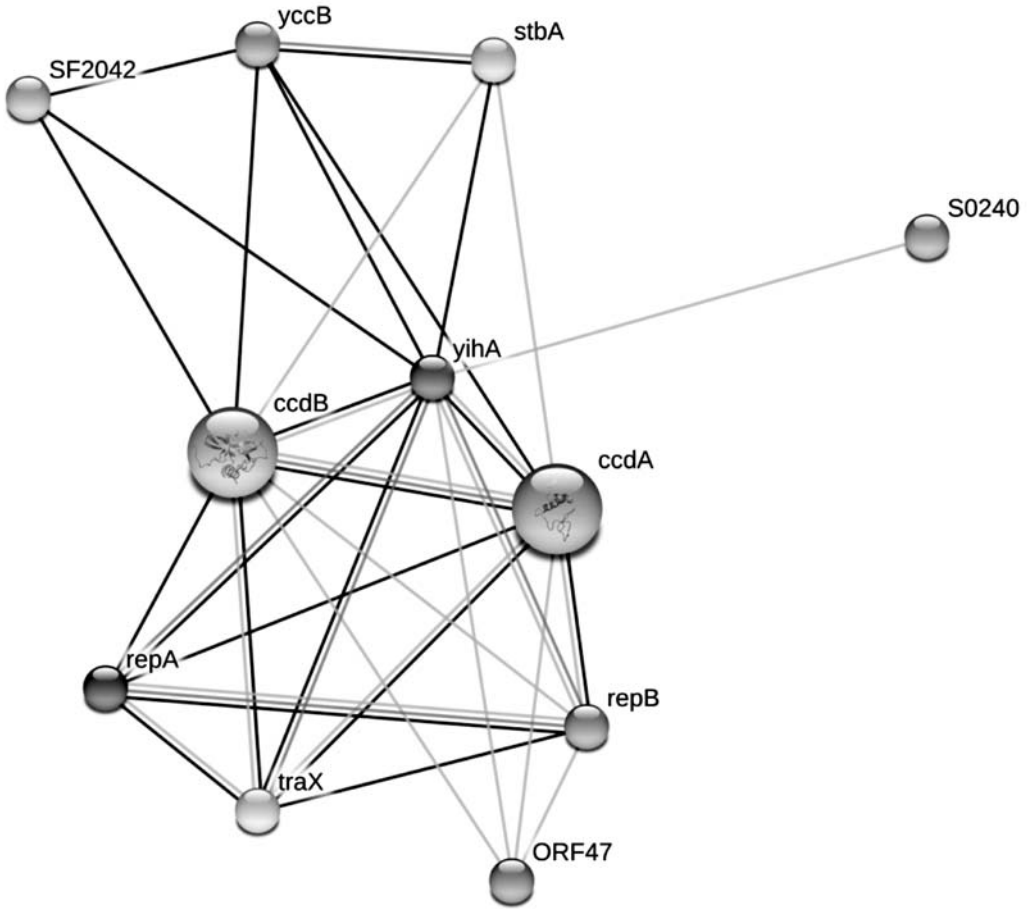

Figure 5. *Hypothetical functional partners of protein yihA explored by STRING. Protein network represents predicted functional partners of protein yihA explored by STRING. Hypothetical functional partners are shown in 3D balls with associated interaction pattern. Among the functional partners repA and repB are found functionally important and functionally associated with replication.*

Discussion

This study was targeted to analyze DUF2726 member proteins *in silico*. DUF2726 family member proteins are mainly present on different bacterial species of Gamma-proteobacteria and have a unique domain organization. We had analyzed the protein sequences of domain DUF2726 by utilizing different bioinformatics tools and databases. We have found that this domain contains a nuclear localization signal peptide, which is conserved in *Escherichia* spp. and *Shigella* spp. In some previous studies it was found that bacteria contain nuclear localization signal NLS in their chromosome and/or plasmid and that NLS can work efficiently upon infection of mammalian cells and has the capability to transduce protein into mammalian cells[26]. It was also found that some nucleoid associated proteins can play role in gene regulation and horizontal gene transfer in enterobacteria[27]. It was also found that some protein can have independent nuclear and nucleoid localization functions[28]. In this study it was predicted that it has nucleic acid binding properties. Analyzing protein-protein interactions using STRING functional partners associated with DUF2726 were explored. We have found that it has gene neighbourhood and co-occurrences with protein repA and repB. RepA and repB are functionally associated with replication. RepA is a replication protein and RepB is a replication regulatory protein. RepA2 is involved in the determination of copy number in gene replication and RepA can play role as initiator protein for mRNA synthesis. Presence of a nucleic acid binding properties, a NLS signalling peptide, and possible interaction pattern with other replication proteins, infers its possible role as a NLS like signalling peptide associated with gene regulation.

Conclusion

This study predicts that DUF2726 can act as a NLS like signalling peptide. Data found from different tools or databases coincides with the predicted function and showed significant evidence on behalf of it. Further analysis in wet lab will verify this result.

References

1. Coggill P, Finn RD and Bateman A. 2008. Identifying protein domains with the Pfam database. *Curr Protoc Bioinformatics*. Chapter 2, Unit 2.5.

2. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL,

Eddy SR and Bateman A. 2010. The Pfam protein families database. *Nucleic Acids Res.* **38**(Database issue): D211-222.

3. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A and Finn RD. 2012. The Pfam protein families database. *Nucleic Acids Res.* **40**(Database issue): D290-301.
4. Bateman A, Coggill P and Finn RD. 2010. DUFs: Families in search of function. *Acta Crystallogr Sect F Struct Biol Cryst Commun.* **66**(Pt 10): 1148-1152.
5. Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon AM, Wilson IA and Godzik A. 2009. Exploration of uncharted regions of the protein universe. *PLoS Biol.* **7**(9): e1000205.
6. Sammut SJ, Finn RD and Bateman A. 2008. Pfam 10 years on: 10,000 families and still growing. *Brief Bioinform.* **9**(3): 210-219.
7. Heller K. 2009. Charting an unknown protein universe. *PLoS Biol*, **7**(9): e1000206.
8. Pena-Castillo L, and Hughes TR. 2007. Why are there still over 1000 uncharacterized yeast genes? *Genetics.* **176**(1): 7-14.
9. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P and Von Mering C. 2009. STRING 8 – A global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37**(Database issue): D412-416.
10. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO and Eisenberg D. 2004. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* **5**(5): R35.
11. Soding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics*. **21**(7): 951-960.
12. Madera M. 2008. Profile comparer: A program for scoring and aligning profile hidden Markov models. *Bioinformatics.* **24**(22): 2630-2631.
13. Jung I and Kim D. 2009. SIMPRO: Simple protein homology detection method by using indirect signals. *Bioinformatics.* **25**(6): 729-735.
14. Bateman A and Finn RD. 2007. SCOOP: A simple method for identification of novel protein superfamily relationships. *Bioinformatics.* **23**(7): 809-814.
15. Boutet E, Lieberherr D, Tognolli M, Schneider M and Bairoch A. 2007. Uniprotkb/Swiss-Prot. *Methods Mol Biol*. **406**: 89-112.
16. Hawkins T, Chitale M, Luban L and Kihara D. 2009. PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins.* **74**(3): 566-582.
17. Cai CZ, Han LY, Ji ZL, Chen X and Chen YZ. 2003. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acid Res.* **31**(13): 3692-3697.
18. Hawkins T, Luban S and Kihara D. 2006. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Sci*. **15**(6): 1550-1556.
19. Nair R, Carter P and Rost B. 2003. NLSdb: Database of nuclear localization signals. *Nucleic Acids Res.* **31**(1): 397-399.
20. Clamp M, Cuff J, Searle SM and Barton GF. 2004. The jalview java alignment editor. *Bioinformatics*, **20**(3): 426-427.
21. Zhou H and Zhou Y. 2003. Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden Markov model based method. *Protein Sci.* **12**(7): 1547-1555.
22. Bryson K, Cozzetto D and Jones DT. 2007. Computer-assisted protein domain boundary prediction using the Dom-Pred server. *Current Protein Peptide Sci*. **8**(2): 181-188.
23. Bujnicki JM, Elofsson A, Fischer D and Rychlewski L. 2001. Structure prediction meta server. *Bioinformatics.* **17**(8): 750-751.
24. Snel B, Lehmann G, Bork P and Huynen MA. 2000. STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **28**(18): 3442-3444.
25. Mering CV, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA and Bork P. 2005. STRING: Known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acid Res.* **33**(Suppl 1): D433-D37.
26. Kaczmarczyk SJ, Sitaraman K, Hill T, Hartley JL and Chatterjee DK. 2010. Tus, an *E. coli* protein, contains mammalian nuclear targeting and exporting signals. *PLoS One.* **5**(1): e8889.
27. Banos RC, Vivero A, Aznar S, Garcia J, Pons M, Madrid C and Juarez A. 2009. Differential regulation of horizontally acquired and core genome genes by the bacterial modulator H-NS. *PLoS Genet.* **5**(6): e1000513.
28. Redrejo-Rodriguez M, Munoz-Espin D, Holguera I, Mencia M and Salas M. 2013. Nuclear and nucleoid localization are independently conserved functions in bacteriophage terminal proteins. *Mol Microbiol.* **90**(4): 858-868.