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 are 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 constituent 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
**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**

<table>
<thead>
<tr>
<th>Domain name</th>
<th>Molecular function terms</th>
<th>Biological process terms</th>
<th>Cellular component terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUF 2726</td>
<td>77% – Melanocortin receptor activity</td>
<td>65% – Cellular metabolism</td>
<td>75% – Voltage-gated sodium channel complex</td>
</tr>
<tr>
<td></td>
<td>67% – Melanocyte stimulating hormone receptor activity</td>
<td>60% – Cellular physiological process</td>
<td>60% – Cell</td>
</tr>
<tr>
<td></td>
<td>58% – Ion binding</td>
<td>56% – Physiological process</td>
<td>59% – Intracellular</td>
</tr>
<tr>
<td></td>
<td>57% – Binding</td>
<td>56% – Cellular macromolecule metabolism</td>
<td>59% – Membrane</td>
</tr>
<tr>
<td></td>
<td>57% – Cation binding</td>
<td>55% – Pyruvate family amino acid metabolism</td>
<td>56% – Cytoplasm</td>
</tr>
<tr>
<td></td>
<td>56% – Transition metal ion binding</td>
<td>54% – Cellular process</td>
<td>56% – Intracellular organelle</td>
</tr>
<tr>
<td></td>
<td>52% – Metal ion binding</td>
<td>54% – Primary metabolism</td>
<td>55% – Plasma membrane</td>
</tr>
<tr>
<td></td>
<td>50% – Protein binding</td>
<td>53% – Macromolecule metabolism</td>
<td>54% – Nucleus</td>
</tr>
<tr>
<td></td>
<td>50% – Nucleotide binding</td>
<td>53% – Metabolism</td>
<td>53% – Intracellular membrane-bound organelle</td>
</tr>
<tr>
<td></td>
<td>50% – Receptor binding</td>
<td>52% – Cellular biosynthesis</td>
<td>53% – Exosome (RNase complex)</td>
</tr>
<tr>
<td></td>
<td>49% – Nucleic acid binding</td>
<td>52% – Biopolymer metabolism</td>
<td>52% – Intracellular non-membrane-bound organelle</td>
</tr>
<tr>
<td></td>
<td>51% – Nucleotide binding</td>
<td>51% – Cellular protein metabolism</td>
<td>50% – Nucleus</td>
</tr>
<tr>
<td></td>
<td>46% – Peptide receptor activity</td>
<td>49% – Amino acid and derivative metabolism</td>
<td>52% – Cytoskeleton</td>
</tr>
<tr>
<td></td>
<td>46% – Cytoskeletal protein binding</td>
<td>49% – Nucleotide, nucleoside, nucleotide and nucleic acid metabolism</td>
<td>43% – Integral to plasma membrane</td>
</tr>
<tr>
<td></td>
<td>45% – Peptide binding</td>
<td>47% – Amino acid metabolism</td>
<td>39% – Integral to membrane</td>
</tr>
<tr>
<td></td>
<td>43% – RNA binding</td>
<td>45% – Aminobiochemical</td>
<td>39% – Non-membrane-bound organelle</td>
</tr>
<tr>
<td></td>
<td>41% – Purine nucleotide binding</td>
<td>45% – Cellular catabolism</td>
<td>33% – Microtubule cytoskeleton</td>
</tr>
<tr>
<td></td>
<td>40% – Calcium ion binding</td>
<td>45% – Aminobiochemical</td>
<td>28% – Membrane-bound organelle</td>
</tr>
<tr>
<td></td>
<td>37% – Catalytic activity</td>
<td>43% – Transport</td>
<td>28% – Aminobase-1</td>
</tr>
<tr>
<td></td>
<td>35% – Calcium ion binding</td>
<td>43% – Nitrogen compound biosynthesis</td>
<td>27% – Organelle</td>
</tr>
<tr>
<td></td>
<td>35% – Actin binding</td>
<td>43% – Cell motility</td>
<td>20% – Microtubule</td>
</tr>
<tr>
<td></td>
<td>34% – Peptide receptor activity</td>
<td>42% – DNA metabolism</td>
<td>20% – Integral to plasma membrane</td>
</tr>
<tr>
<td></td>
<td>33% – X-Pro dipeptidyl-peptidase activity</td>
<td>41% – Locomotion</td>
<td>19% – Microtubule</td>
</tr>
<tr>
<td></td>
<td>29% – Oxidoreductase activity</td>
<td>32% – Peptide receptor activity, G-protein coupled</td>
<td>28% – Aminobase-1</td>
</tr>
</tbody>
</table>

*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 (ERRRRRD) 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

* Hypothetical functional partners of YihA revealed by STRING. Functionally important protein- RepA and RepB are matched by gene neighborhood and co-occurrence pattern.
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\(^2^6\). It was also found that some nucleoid associated proteins can play role in gene regulation and horizontal gene transfer in enterobacteria\(^2^7\). It was also found that some protein can have independent nuclear and nucleoid localization functions\(^2^8\). 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


