Predicting the function of a hypothetical protein from Pyrococcus horikoshii OT3 as HD domain containing metal-dependent phosphohydrolase

P Babu, K Harshita, V prasanth, S Chitti

Citation

P Babu, K Harshita, V prasanth, S Chitti. *Predicting the function of a hypothetical protein from Pyrococcus horikoshii OT3 as HD domain containing metal-dependent phosphohydrolase*. The Internet Journal of Genomics and Proteomics. 2009 Volume 5 Number 2.

Abstract

2CQZ, a hypothetical protein from Pyrococcus horikoshii OT3 was selected to study the presence of domain regions based on similarity search using BLASTp (Basic Local Alignment Search Tool for protein) server against non-redundant databases of NCBI (National Centre for Biotechnology Information). Similarity analysis revealed matches with oxetanocin-like protein, metal dependent phosphohydrolases and HD domain containing proteins, respectively. Further multiple alignments were constructed to recognize the regions of conserved residues. From our analysis and Pfam entry PF01966 it was identified that this protein is known to contain a HD domain motif. 2CQZ showed highest similarity with oxetanocin-like protein and moderate similarities with metal-dependent phosphohydrolases and HD domain proteins. Finally, it was observed that the metal-coordinating HD motif (H33, H68, D69, and D137) and other conserved residues (R18, E72, and D77), important for activity are retained in 2CQZ and the protein thus belongs to the superfamily of metal-dependent phosphohydrolases.

INTRODUCTION

Many small bacterial, archaebacterial and larger eukaryotic genomes are currently being sequenced. In all genomes sequenced to date, a large portion of these organisms' protein coding regions encodes polypeptides of unknown biochemical, biophysical, and/or cellular functions. In biochemistry, a hypothetical protein is a protein whose existence has been predicted, but for which there is no experimental evidence that it is expressed in vivo (Zarembinski et al 1998).

The function of a hypothetical protein can be predicted by domain homology searches with various confidence levels and assigning the molecular function of a protein with unknown function starts with determining the three-dimensional structure of the protein by either X-ray crystallography or NMR. The structural sequence is then compared against those of the protein structure database (Protein Data Bank). If there are one or more significant structural homologs, the hypothetical protein will have molecular properties similar to the homologs (Fields et al 1999; Edwards et al 2003). Many protein families have diverged from common ancestors and are represented with one or more domains. Domains are characterized as semi-independent 3D-units, often associated with a particular function, are genetically mobile and frequently moving within and between biological systems by gene or exon shuffling (George et al 2002). Therefore understanding the domain organization of a protein sequence is crucial for structural and functional genomics initiatives.

In this paper we address the prediction of domain regions in a hypothetical protein, 2CQZ, from Pyrococcus horikoshii OT3. In order to understand the functional aspects based on residue conservation among similar proteins, multiple alignment program MAFFT (Multiple Alignment using Fast Fourier Transform) (http://www.ebi.ac.uk/Tools/mafft) was used.

MATERIALS AND METHODS

PDB (Protein Data Bank) (http://www.rcsb.org/pdb) was searched for hypothetical proteins of unknown function and 2CQZ protein (190 residues length) of Pyrococcus horikoshii OT3 was selected. The structure was composed of 68% helical and 1% beta sheet secondary structural elements without any breaks in the protein structure.

PAIR WISE ALIGNMENT

2CQZ sequence in FASTA format was extracted from PDB.

Similarity search was carried out by scanning the sequence against non-redundant database of NCBI (National Centre for Biotechnology Information) (http://www.ncbi.nlm.nih.gov) using protein-protein BLAST pair wise alignment tool, BLASTp (Altschul et al 1990).

MULTIPLE ALIGNMENTS

Multiple sequence alignments of similar proteins resulted from blast analysis was performed by MAFFT (Multiple Alignment using Fast Fourier Transform) using default parameters. Multiple alignments are carried out to identify the regions of residue conservation among homologous proteins (Yamada et al 2006).

RESULTS AND DISCUSSION

BLASTp search analysis of 2CQZ protein sequence against non-redundant database resulted in many hits with varying degrees of percentage identities and similarities. Reported in Table 1 are the results of the analysis, where it can be observed that the oxetanocin-like protein (NP_578124.1), metal-dependent phosphohydrolase (YP_182427.1), HD domain (NP_587821.1), HD domain containing protein 2 (XP_001315286.1), showed similarity ranging from 63-32% and the remaining hits obtained are with other hypothetical proteins. From Table 1, it can be understood that 2CQZ would represent a metal-dependent phosphohydrolase and may contain HD domain, respectively. Therefore multiple alignments are constructed with each protein family versus 2CQZ protein to provide clarity towards probable relational aspects of 2CQZ.

Figure 1

Table 1: Sequence analysis of 2CQZ vs non-redundant database showing %identities, %positives, gaps and residue overlap (query and subject sequences).

From Table 1 it is evident that the number of residues identical with 2CQZ varied considerably, and moreover, both the metal-dependent phosphohydrolase and HD domain containing proteins matched with our protein (based on the number of entries). A search in NCBI nucleotide database for oxetanocin-like protein revealed one entry on complete genome of Pyrococcus furiosus DSM 3638. A similar search in PDB as well as blastp scan using 2CQZ sequence against PDB revealed 63% identity and 82% similarity with 1XX7 (http://www.rcsb.org/pdb/explore.do?structureId=1XX7) , a conserved hypothetical protein from Pyrococcus furiosus Pfu-403030-001. This protein is known to contain a HD domain motif.

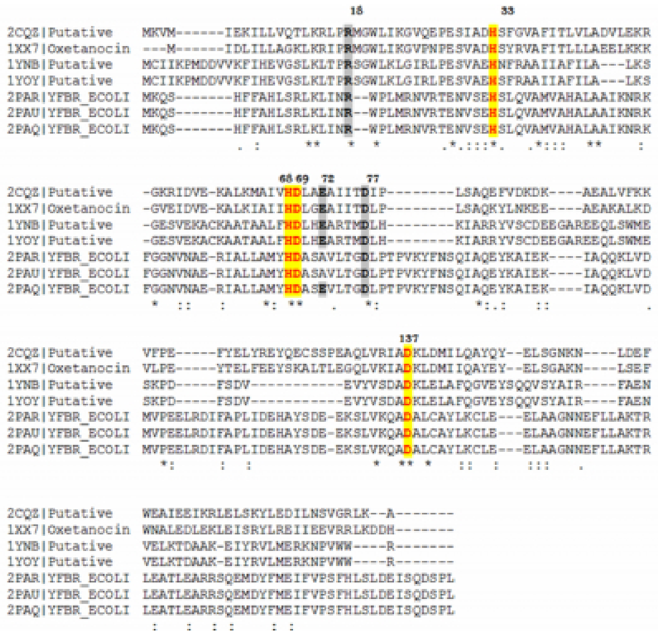
HD domain is found in a superfamily of enzymes either with predicted or known phosphohydrolase activity. These enzymes appear to be involved in the phosphatase or phosphodiesterase activities and accordingly play a role in signal transduction and possibly other functions in bacteria, archaea and eukaryotes. The HD superfamily is reported to possess highly conserved key metal-binding residues, histidines or aspartates, essential for the activity of these proteins (Aravind et al 1998). Also, metal-dependent phosphohydrolases are known to possess HD domain, important for the activity of these proteins.

Therefore, from the above data and from Pfam entry PF01966, (http://pfam.sanger.ac.uk/family? acc=PF01966) it is observed that oxetanocin-like protein has HD domain and are metal-dependent phosphohydrolases. As 2CQZ showed highest similarity with oxetanocin-like protein and moderate similarities with metal-dependent phosphohydrolases and HD domain proteins, it is believed that 2CQZ should also contain a HD domain motif and hence further investigation was directed to find HD domain similarities.

It has been reported by Zimmerman et al (2008), that the nucleotidase YfbR proteins from Escherichia coli are HD domain phosphohydrolases and contains a large cavity accommodating the metal-coordinating HD motif (H33, H68, D69, and D137) and other conserved residues (R18, E72, and D77), important for activity. Hence, PDB proteins having HD domains are selected from Pfam entry PF01966, viz., 2PAR (Escherichia coli), 2PAQ (Escherichia coli), 2PAU (Escherichia coli), 1XX7 (Pyrococcus furiosus Pfu-403030-001), 1YNB (Archaeoglobus fulgidus) and 1YOY (Archaeoglobus fulgidus dsm 4304). From these, 1YNB and 1YOY proteins lack metal ions and multiple sequence alignments between six protein sequences and 2CQZ using MAFFT 4.0 program revealed a clear match with HD motif (Figure 1) which confirms that the hypothetical protein 2CQZ retains HD domain function and belongs to the superfamily of metal-dependent phosphohydrolases.

Figure 2

Figure 1: MAFFT multiple alignments of six HD domain containing proteins showing E72A mutation in 2PAR and 2PAU



```
                                         18               33
2CQZ|Putative     MKVM------IEKILLVQTLKRLPRMGWLIKGVQEPESIADHSFGVAFITLVLADVLEKR
1XX7|Oxetanocin   ---M------IDLILLAGKLKRIPRMGWLIKGVPNPESVADHSYRVAFITLLLAEELKKK
1YNB|Putative     MCIIKPMDDVVKFIHEVGSLKLTPRSGWLKLGIRLPESVAEHNFRAAIIAFILA---LKS
1YOY|Putative     MCIIKPMDDVVKFIHEVGSLKLTPRSGWLKLGIRLPESVAEHSFRAAIIAFILA---LKS
2PAR|YFBR_ECOLI   MKQS-------HFFAHLSRLKLINR--WPLMRNVRTENVSEHSLQVAMVAHALAAIKNRK
2PAU|YFBR_ECOLI   MKQS-------HFFAHLSRLKLINR--WPLMRNVRTENVSEHSLQVAMVAHALAAIKNRK
2PAQ|YFBR_ECOLI   MKQS-------HFFAHLSRLKLINR--WPLMRNVRTENVSEHSLQVAMVAHALAAIKNRK
                             . :     **   *  *       .*.:::*.  .*:::  **     :

                                 68 69 72   77
2CQZ|Putative     -GKRIDVE-KALKMAIVHDLAEAIITDIP---------LSAQEFVDKDK----AEALVFKK
1XX7|Oxetanocin   -GVEIDVE-KALKIAIIHDLGEAIITDLP---------LSAQKYLNKEE----AEAKALKD
1YNB|Putative     -GESVEKACKAATAALFHDLHEARTMDLH---------KIARRYVSCDEEGAREEQLSWME
1YOY|Putative     -GESVEKACKAATAALFHDLHEARTMDLH---------KIARRYVSCDEEGAREEQLSWME
2PAR|YFBR_ECOLI   FGGNVNAE-RIALLAMYHDASAVLTGDLPTPVKYFNSQIAQEYKAIEK----IAQQKLVD
2PAU|YFBR_ECOLI   FGGNVNAE-RIALLAMYHDASAVLTGDLPTPVKYFNSQIAQEYKAIEK----IAQQKLVD
2PAQ|YFBR_ECOLI   FGGNVNAE-RIALLAMYHDASEVLTGDLPTPVKYFNSQIAQEYKAIEK----IAQQKLVD
                   *  ::   :    *: **   .   *:           *:.:   ::           .

                                                137
2CQZ|Putative     VFPE-----FYELYREYQECSSPEAQLVRIADKLDMILQAYQY--ELSGNKN----LDEF
1XX7|Oxetanocin   VLPE-----YTELFEEYSKALTLEGQLVKIADKLDMIIQAYEY--ELSGAKN----LSEF
1YNB|Putative     SKPD-----FSDV-----------EVYVSDADKLELAFQGVEYSQQVSYAIR----FAEN
1YOY|Putative     SKPD-----FSDV-----------EVYVSDADKLELAFQGVEYSQQVSYAIR----FAEN
2PAR|YFBR_ECOLI   MVPEELRDIFAPLIDEHAYSDE-EKSLVKQADALCAYLKCLE---ELAAGNNEFLLAKTR
2PAU|YFBR_ECOLI   MVPEELRDIFAPLIDEHAYSDE-EKSLVKQADALCAYLKCLE---ELAAGNNEFLLAKTR
2PAQ|YFBR_ECOLI   MVPEELRDIFAPLIDEHAYSDE-EKSLVKQADALCAYLKCLE---ELAAGNNEFLLAKTR
                    *:     :  :              *  ** *   ::  :   :::   .

2CQZ|Putative     WEAIEEIKRLELSKYLEDILNSVGRLK--A-------
1XX7|Oxetanocin   WNALEDLEKLEISRYLREIIEEVRRLKDDH-------
1YNB|Putative     VELKTDAAK-EIYRVLMERKNPVWW----R-------
1YOY|Putative     VELKTDAAK-EIYRVLMERKNPVWW----R-------
2PAR|YFBR_ECOLI   LEATLEARRSQEMDYFMEIFVPSFHLSLDEISQDSPL
2PAU|YFBR_ECOLI   LEATLEARRSQEMDYFMEIFVPSFHLSLDEISQDSPL
2PAQ|YFBR_ECOLI   LEATLEARRSQEMDYFMEIFVPSFHLSLDEISQDSPL
                   :   :  : :    : :
```

References

r-0. Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-410.

r-1. Aravind L, and Koonin EV. The HD domain defines a new superfamily of metal-dependent phosphohydrolases. Trends Biochem Sci 1998; 23:469-472.

r-2. Edwards YJ, and Cottage A. Bioinformatics methods to predict protein structure and function. A practical approach. Mol Biotechnol 2003; 23:139-166.

r-3. Fields S, Kohara Y, and Lockhart DJ. Functional genomics. Proc. Natl. Acad. Sci. USA 1999; 96:8825-8826.

r-4. George RA, and Heringa J. Protein Domain Identification and Improved Sequence Similarity Searching Using PSI-BLAST. Proteins 2002; 48:672–81.

r-5. Yamada S, Gotoh O, and Yamana H. Improvement in accuracy of multiple sequence alignment using novel group-to-group sequence alignment algorithm with piecewise linear gap cost. BMC Bioinformatics 2006; 7:524-541.

r-6. Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, and Kim SH. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. Proc. Natl. Acad. Sci. USA 1998; 95:15189-15193.

r-7. Zimmerman MD, Proudfoot M, Yakunin A, and Minor W. Structural insight into the mechanism of substrate specificity and catalytic activity of an HD-domain phosphohydrolase: the 5'-deoxyribonucleotidase YfbR from Escherichia coli. J Mol Biol 2008; 378:215-226.

Author Information

P. Ajay Babu, Ph.D
Bioinformatics Division, Translational Research Institute of Molecular Sciences (TRIMS)

K. Harshita, B.Tech
Department of Biotechnology, Merit-International Institute of Technology

V. Vishnu prasanth, M.Sc
Bioinformatics Division, Translational Research Institute of Molecular Sciences (TRIMS)

Sashikanth Chitti, MS
Bioinformatics Division, Translational Research Institute of Molecular Sciences (TRIMS)