ISSN 0975-6299

Vol 2/Issue 2/Apr-Jun 2011

International Journal of Pharma and Bio Sciences

RESEARCH ARTICLE

BIOINFORMATICS

# STRUCTURAL & FUNCTIONAL PREDICTION OF HYPOTHETICAL PROTEINS IN BACTERIOPHAGES AGAINST HALOPHILIC BACTERIA- AN *IN SILICO* APPROACH

**SWAPNIL G SANMUKH, WAMAN N PAUNIKAR*, TARUN K GHOSH, TAPAN CHAKRABARTI**

National Environmental Engineering Research Institute, Nehru Marg, Nagpur- 440020, India

*Corresponding author

WAMAN N PAUNIKAR

National Environmental Engineering Research Institute, Nehru Marg, Nagpur- 440020, India

## ABSTRACT

The hypothetical proteins of two Halophilic phages, Halomonas phage phiHAP-1 (length 39245) and Halorubrum phage HF2 (length 77670) were studied. The structure and function predictions of hypothetical proteins were done by the use of bioinformatic web tools. In Halomonas phage phiHAP-1, function prediction for 9 and structure prediction for 3 hypothetical proteins whereas in Halorubrum phage HF2, function prediction for 32 and structure prediction for 6 hypothetical proteins were resolved. The probable function prediction was done by using Bioinformatics web tools like CDD-BLAST, INTERPROSCAN, PFAM and COGs by searching sequence databases for the presence of orthologous conserved domains in the hypothetical sequences. The tertiary structure prediction was done by using $PS^2$ server. This study revealed some of the uncharacterized hypothetical proteins in Halophilic phages whose complete genome sequence is known and can be used for the further understanding of phage genome and its life cycle.

## KEY WORDS

Hypothetical proteins, Halomonas phage phiHAP-1, Halorubrum phage HF2, bioinformatics web tools, orthologous conserved domains.

## INTRODUCTION

The Halophilic bacteriophages are lytic and infect a wide range of haloarchaea including *Halobacterium*, *Halomonas*, *Haloferax*, *Haloarcula*, *Natrialba*, *Haloterrigena* and *Halorubrum*. The complete sequences of the HF1 and HF2 genomes which is linear dsDNA have recently been completed, they are the largest archaeal virus genomes sequenced having 75.9 kb and 77.7 kb respectively (Tang *et al.*, 2004).The two sequences are identical except for single base pair in first 48 kb, but are extremely varying in other regions, suggesting a recombination event between different haloviruses. This suggests that there is a high level of recombination among viruses that live in hypersaline environments (Tang *et al.*, 2004). As only few hypersaline bacteriophages are sequenced there is no available information regarding their life cycle or their survival in such extreme condition.

The complete genome sequence of the two phages under study is known. The Halomonas phage phiHAP-1 has 39245 nucleotides and 46 protein genes whereas; Halorubrum phage HF2 has 77670 nucleotides and 114 protein genes. Out of 22 hypothetical proteins in Halomonas phage phiHAP-1, function predictions for 9 hypothetical proteins and structure prediction of 3 hypothetical proteins were resolved. Similarly, in Halorubrum phage HF2, out of 105 hypothetical proteins, function prediction for 32 proteins and structure prediction of 6 proteins was resolved.

The Bioinformatics provide a platform for prediction of unknown characteristics of protein through genomic and proteomic approach. The development in the genome sequencing and gene marking provided the opportunity to study in detail the coding ability of the organism for proteins and assists in understanding the regulatory and expression pattern of the proteins. Bioinformatic web tool can predict homology in hypothetical proteins by searching the enzymatic domains from defined databases for proper functional characterization. CDD-BLAST, INTERPROSCAN, PFAM and COGs can search the orthologous sequence in biological sequence databases for the target sequence and assist in classification of hypothetical protein in particular family (Edward et al., 2000). Similarly, the 3D-structure for target protein can be constructed using best scored template of orthologous family member (Zafer et al., 2006; Chih-Chieh et al., 2006). The $PS^2$ server was used for the tertiary structure prediction of hypothetical proteins in Halophilic phages. This study will helps us to understand the probable functions of hypothetical proteins in the life cycle of Halophilic phages and derived data could be used in the further research in hypersaline environment.

## MATERIALS AND METHODS

### *Sequence Retrieval*

Complete protein sequences of Halorubrum phage HF2 (Tang, 2002) and Halomonas phage phiHAP-1 (Mobberley, 2008) were downloaded from the Database of KEGG (http://www.genome.jp/kegg/genome).

### Functional Annotations

Hypothetical proteins were screened for the presence of conserved domains using sequence similarity search with close orthologous family members available in various protein databases using the web tools. Four bioinformatics web tools like CDD-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) (Altschul et al., 1997; Schaffer et al., 2001; Aron et al.,2006),INTERPROSCAN (http://www.abi.ac.uk/interpro) ( Zdobnov and Rolf, 2001) , Pfam (http://www.pfam.sanger.ac.uk/) (Alex et al., 2004) and COGs (http://www.ncbi.nih.gov/cog) (Roman et al., 2000) were used, which shows the ability to search the defined conserved domains in the sequences and assist in the classification of proteins in a concern family.

### *Protein Structure Prediction*

$PS^2$ (PS Squared) Protein Structure Prediction Server was used (http://www.ps2.life.nctu.edu.tw/) ( Chih-Chieh et al., 2006; Altschul et al.,1997; Schaffer et al., 2001; Cédric et al., 2000; Wendy et al.,2000), which accepts the protein (query) sequences in FASTA format and uses the strategies of Pair-wise and multiple alignment by combining powers of the programs PSI-BLAST, IMPALA and T-COFFEE in both target – template selection and target– template alignment and resultant target proteins 3D structures were constructed using structural positioning information of atomic coordinates for known template in PDB format using best scored alignment data. Where the selection of template was based on the same conserved domain detected in the functional annotations and which must be available in the structure alignment for modeling purpose.

## RESULTS AND DISCUSSION

### *Functional Annotations*

The web tools based analysis classified each hypothetical protein into particular protein family member based on conserved domain available in the sequence, and are represented in Table 1 for Halorubrum phage HF2 and Table 2 for Halomonas phage phiHAP-1.

Table 1 Probable Conserved Domains in Hypothetical proteins of Halorubrum phage HF2

| Gene Sequence | CDD blast | Interproscan | Pfam | Cogs |
|---|---|---|---|---|
| 929529 | no | no | Tombusvirus P19 core protein | no |
| 929535 | Nucleases | Restriction endonuclease | no | Nucleases |
| 929538 | Phage_integrase | Integrase | Phage integrase | Integrase |
| 929544 | Repressor of nif and glnA expression | no | Ribonuclease | no |
| 929552 | no | no | Prophage tail fibre N-terminal | no |
| 929553 | no | no | Non-histone chromosomal protein MC1 | no |
| 929555 | Band 7 domain of flotillin | Prohibitin, Band 7 protein | Band 7 | Membrane protease |
| 929556 | no | no | MarR family | no |
| 929557 | no | no | TFIIB zinc-binding | Transcription initiation factor IIB |
| 929564 | no | no | PhnA Zinc-Ribbon | no |
| 929566 | Phosphohexomutase | no | no | Phosphomannomutase |
| 929575 | no | no | no | Ribonucleotide reductase |
| 929576 | no | no | Renin receptor-like protein | no |
| 929577 | no | no | RNA polymerase I | no |
| 929583 | no | no | lipoprotein lipid attachment site | no |
| 929587 | no | no | Malarial early transcribed membrane protein | no |
| 929589 | no | no | Biofilm regulator BssS, Bacteriophage lysis protein | no |
| 929595 | no | no | Uncharacterized protein conserved in bacteria | no |
| 929598 | no | no | Ubp3 associated protein Bre5 | no |
| 929599 | no | no | Phage head-tail joining protein | Methionine synthase II (cobalamin-independent) |
| 929600 | no | no | Phage terminase | no |
| 929602 | no | no | Preprotein translocase | no |
| 929605 | no | no | NADH ubiquinone oxidoreductase | no |
| 929612 | no | no | Fimbrial protein | no |
| 929615 | P2 bacteriophage J protein baseplate | Baseplate assembly protein J | Baseplate J-like protein | homolog of phage Mu protein gp47 |
| 929616 | no | no | Nucleoporin protein Ndc1-Nup | no |
| 929617 | no | no | Cytomegalovirus UL20A protein | no |
| 929622 | no | no | T-antigen specific domain | no |
| 929625 | no | no | Nucleotidyltransferase, Geminivirus coat protein/nuclear export factor | no |
| 929628 | Phage_sheath_1 super family | Phage tail sheath protein | Phage tail sheath protein | no |
| 929636 | no | no | Baculovirus polyhedron envelope protein | no |
| 929642 | no | no | PAAR motif | no |

Table 2 Probable Conserved Domains in Hypothetical proteins of Halomonas phage phiHAP-1

| | | | | |
|---|---|---|---|---|
| 5912336 | no | no | Prophage minor tail protein Z (GPZ) | no |
| 5912337 | no | no | Nucleopolyhedrovirus P10 protein, Chordopoxvirus fusion protein,Haemolysin XhlA | no |
| 5912344 | no | unintegrated | Integral membrane protein | no |
| 5912345 | PG_binding_3 super family | Peptidoglycan binding domain | Predicted lysozyme (DUF847), Predicted Peptidoglycan domain | no |
| 5912350 | zf-dskA_traR super family | Zinc finger, DksA/TraR C4-type | Prokaryotic dksA/traR C4-type zinc finger | DnaK suppressor protein |
| 5912352 | no | no | Septum formation initiator, Uncharacterized protein conserved in bacteria | no |
| 5912368 | no | Peptidase C14, ICE, catalytic subunit p20, active site | no | no |
| 5912369 | no | no | 5TMR of 5TMR-LYT | no |
| 5912372 | no | no | no | no |
| 5912379 | no | no | IclR helix-turn-helix domain | no |

## Protein Structure Prediction

The $(PS)^2$ Server built the three dimensional structures for hypothetical proteins and available in PDB format files which can be used in further studies. $(PS)^2$ satisfactorily predicted 3-D structures of 3 hypothetical proteins in Halomonas phage phiHAP-1 and 6 hypothetical proteins in Halorubrum phage HF2 using best scored orthologous template, while, server failed to predict tertiary structures of 19 hypothetical proteins in Halomonas phage phiHAP-1 and 26 hypothetical proteins in Halorubrum phage HF2 due to the lack of defined 3D structures for the aligned templates. These templates with best scoring with hypothetical sequences were represented by their features such as Template ID, Identity, Score and E-value which was represented in Table 3 for Halorubrum phage HF2 and Table 4 for Halomonas phage phiHAP-1. The predicted structures of hypothetical proteins are given in Figure 1-3 for Halomonas phage phiHAP-1 and Figure 4-9 for Halorubrum Phage HF2.

**Predicted structures of hypothetical proteins in Halomonas phage phiHAP-1**

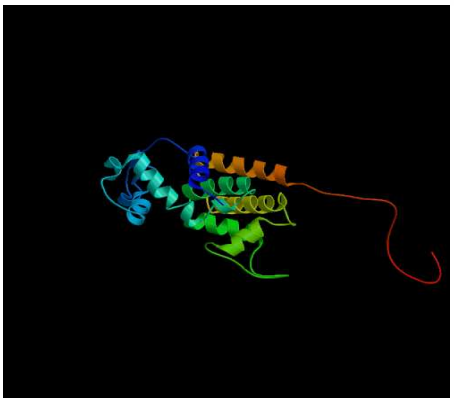

Figure 1 Structure for hypothetical protein with NCBI Gene ID 5912345

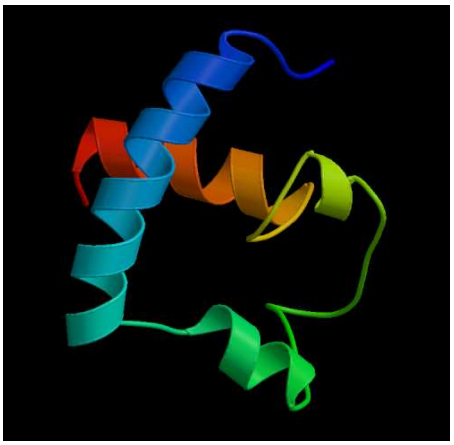

Figure 2 Structure for hypothetical protein with NCBI Gene ID 5912350

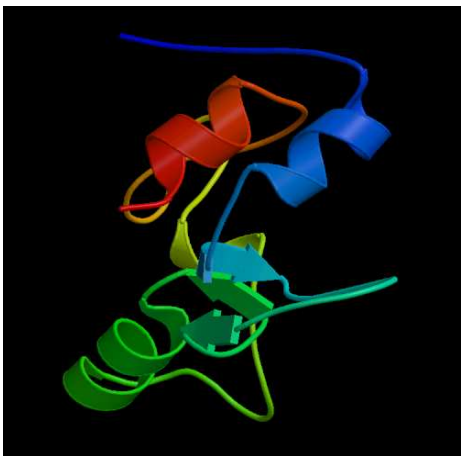

Figure 3 Structure for hypothetical protein with NCBI Gene ID 5912352

**Predicted structures of hypothetical proteins in Halorubrum Phage HF2**

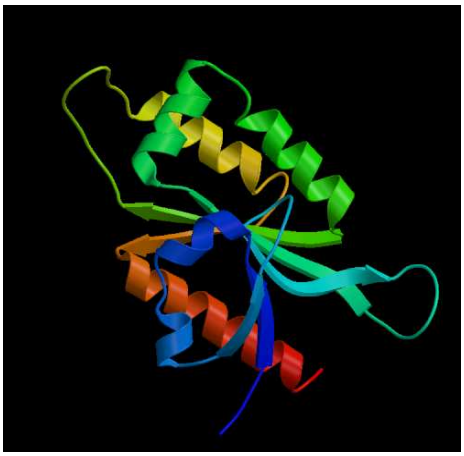

Figure 4 Structure for hypothetical protein with NCBI Gene ID 929535

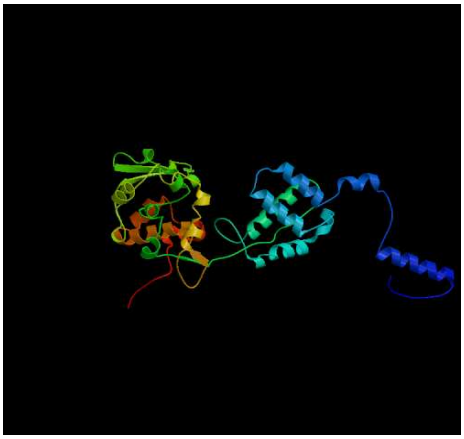

Figure 5 Structure for hypothetical protein with NCBI Gene ID 929538

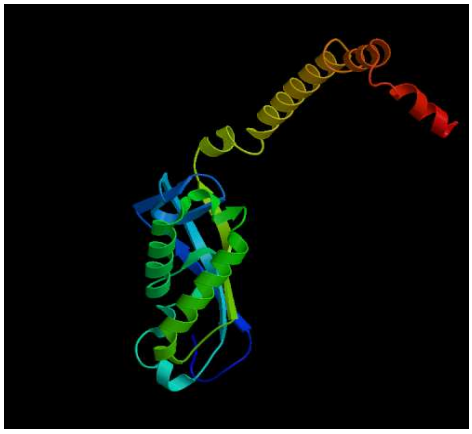

Figure 6 Structure for hypothetical protein with NCBI Gene ID 929555

Figure 7 Structure for hypothetical protein with NCBI Gene ID 929589

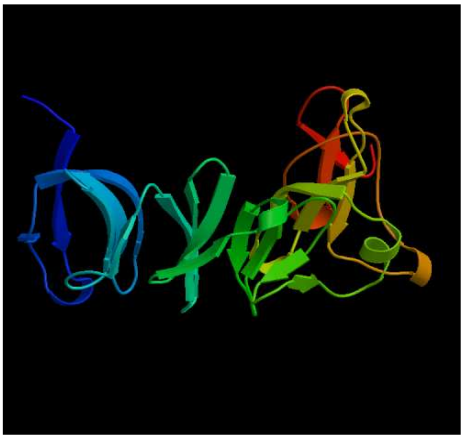

Figure 8 Structure for hypothetical protein with NCBI Gene ID 929612

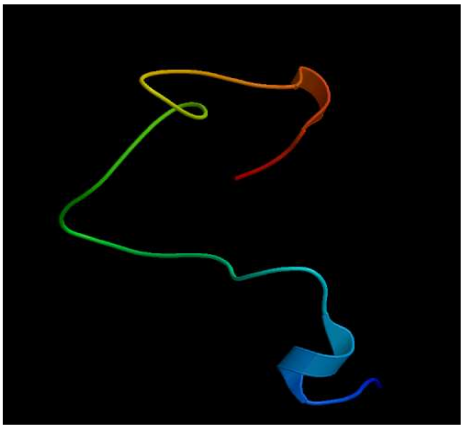

Figure 9 Structure for hypothetical protein with NCBI Gene ID 929636

Table 3 Template ID, Identity, Score and E-value Halorubrum Phage HF2

| NCBI Gene ID | Templates | Identity | Score | E-value |
|---|---|---|---|---|
| 929535 | 1j24A | 22 | 62 | 8.00E-11 |
| 929538 | 1z1bB | 14 | 54 | 4.00E-08 |
| 929555 | 3bk6A | 15 | 120 | 4.00E-28 |
| 929589 | 1e1oA | 28 | 34 | 0.01 |
| 929612 | 2hu5B | 11 | 35 | 6.00E-04 |
| 929636 | 1yy9A | 25 | 30 | 0.01 |

Table 4 Template ID, Identity, Score and E-value of Halomonas phage phiHAP-1

| NCBI Gene ID | Templates | Identity | Score | E-value |
|---|---|---|---|---|
| 5912345 | 2ikbA | 35 | 198 | 5.00E-53 |
| 5912350 | 1tjlA | 23 | 39 | 3.00E-04 |
| 5912352 | 1qssA | 29 | 35 | 0.008 |

## CONCLUSION

This study sorted some functional hypothetical proteins of halophilic bacteriophages by applying the parameters of pairwise and multiple sequence alignment tools along with structure prediction tools, which suggests that many probable functionally uncharacterized proteins are available in the halophilic bacteriophages and their exact role in their lifecycle is still unclear. Development in sequence analysis programming and ever growing genome sequence databases enhanced this methodology to draw conclusive functional relationships in the hypothetical proteins under study. Bioinformatics Web Tools like CDD-BLAST, INTERPROSCAN, PFAM and COGs have shown the ability to predict functions in hypothetical proteins, in that sense assisted in predicting enzymatic activity in some proteins of halophages. $(PS)^2$ serves as fast automated homology modeling web server and helped in predicted three dimensional structures of 9 hypothetical proteins in halophages which may prove beneficial in establishing their role in life cycle of halophilic bacteriophages.

## ACKNOWLEDGEMENT

We are thankful for the help provided by Miss. Kimi patel and Miss. Lekha patel in the work and referencing.

## REFERENCES

1. Alex, B., Lachlan, C., Richard, D., Robert, D. F., Volker, H., Sam, G.J., Ajay, K., Mhairi, M., Simon, M., Erik, L. L. S., David, J. S., Corin Y., Sean, R. E., (2004). The Pfam families' database. Nucleic Acids Research, Vol. 32, D138-D141.
2. Altschul, S., F., Madden, T., L., Schaffer, A., A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., J., (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389-402.
3. Aron, M. Bauer., John, B. A., Myra, K. D., Carol, D. S., Noreen, R. G., Marc, G., Luning, H., Siqian, H., David, I. H., John, D. J., Zhaoxi, K., Dmitri, K., Christopher,

J. L.,Cynthia A. L., Chunlei, L., Fu, L., Shennan, L., Gabriele, H. M., Mikhail, M., James, S. S., Narmada, T., Roxanne, A. Y., Jodie, J. Y., Dachuan, Z., Stephen, H. B., (2006). CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Research, Vol. 35, D237–D240.

4. Cédric, N., Desmond, G. H., Jaap, H., (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205-217.
5. Chih-Chieh, C., Jenn-Kang, H., Jinn-Moon, Y., (2006). (PS)2: protein structure prediction server Nucl. Acids Res. 34, W152-W157.
6. Edward, E., Gary, L. G., Osnat, H., John, M., John, O., Roberto, J. P., Linda, B., Delwood, R., Andrew, J. H., (2000). Biological function made crystal clear-annotation of hypothetical proteins via structural genomics. Current Opinion in Biotechnology 11, 25-30.
7. Roman, L. T., Michael, Y., Galperin, Darren A. Natale, Eugene V. Koonin (2000). The COG database: a tool for genome –scale analysis of protein functions and evolution. Nucleic Acid Research. 28, 33-36.
8. Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S. Spouge, J. L., Wolf, Y. I., Koonin, E. V., Altschul, S. F., (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29(14), 2994-3005.
9. Wendy, B., Alexander, V.D.B., Evelyn, C., Pascal, H., Peter, S., Guenter, S., Mary, A. T., (2000). The EMBL Nucleotide Sequence Database. Nucleic Acid Research. 28, 19-23
10. Zafer, A., Yucel, A., Mark, B., (2006). Protein secondary structure prediction for a single-sequence using hidden semi-Markov models, BMC Bioinformatics, 7, 178.
11. Zdobnov, E. M., Rolf, A., (2001). Interproscan- an integration platform for the signatures recognition methods in InterPro. Bioinformatics 17, 847-848
12. Mobberley, J. M., Authement, R. N., Segall, A. M., Paul, J. H. (2008). The Temperate Marine Phage HAP-1 of *Halomonas aquamarina* Possesses a Linear Plasmid-Like Prophage Genome. J. Virology., Vol. 82, No.13, 6618–6630.
13. Nuttall, S. D., Smith, M. D. (1995). Halophage HF2: Genome Organization and Replication Strategy. J. Virology., Vol. 69, No. 4, 2322–2327.
14. Tang, S. L., Nuttall, S., Smith, M. D. (2004). Haloviruses HF1 and HF2: Evidence for a Recent and Large Recombination Event. J. Bacteriology, Vol. 186, No. 9, 2810–2817