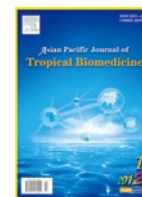




Contents lists available at ScienceDirect

Asian Pacific Journal of Tropical Biomedicine

journal homepage: www.elsevier.com/locate/apjtb

Document heading

In silico sequence and structure analysis for mycobacteriophages

Vasanthi Rajendran, Sameer Hassan, Jerrine Joseph, Nagamiah Selvakumar, Vanaja Kumar*

Department of Bacteriology, Tuberculosis Research Centre, Chennai – 600 031, India

ARTICLE INFO

Article history:

Received 18 January 2012

Received in revised form 21 January 2012

Accepted 2 March 2012

Available online 28 April 2012

Keywords:

Mycobacteriophages
Comparative genomics
BLAST
PFAM

ABSTRACT

Objective: To do a comparative homology search for the 32 mycobacteriophages. **Methods:** This can be estimated as 3 400 proteins from 32 mycobacteriophages assuming each phage has 80–100 genes. The algorithm most widely used for homology detection in comparative genomics is Basic Local Alignment Search Tool (BLAST). Usually a stringent score cutoff is applied to distinguish putative homolog's from possible false positive hits. As a consequence, some BLAST hits are discarded or put into insignificant hits that are in fact homologous. Assigning function to protein sequences is important. Here, we review the status of sequence (BLAST) based and domain based approaches to proteins that can provide functional insights such as database on Protein Families (PFAM). **Results:** The findings showed that 31% of proteins showed similarities with homologous protein with known function. **Conclusions:** In the present study only 31% of mycobacteriophage proteins functions were able to be predicted. Only for 6 proteins, the template structures were available but then they were not directly involved in phage lifecycle. Hence this emphasizes the importance of exploring the structures for mycobacteriophage proteins in order to understand their function and evolutionary significance. Ab initio methods for protein structure prediction can be only alternative for the rest of the proteins but accuracy of prediction is not highly dependable.

1. Introduction

Currently, most approaches to protein function prediction rely on searching sequence databases to identify homologous sequences with prior annotation. The most widely used search tools are Position-Specific Iterative BLAST (PSI-BLAST)[1]; at the National Center for Biotechnology Information alone, 70 000 BLAST searches are performed each day for the general public. It is certainly no coincidence that the BLAST algorithm was the most highly cited paper of the last decade, surpassing all biology publications[2].

PSI-BLAST is an iterative method that uses results from a BLAST search to create a profile (position-specific scoring matrix). The profile is used to search the database for additional homologues, and these results can be used to further improve the profile. A profile captures family-specific information, including functionally and structurally

important residue positions, and can therefore identify distant homologues not recognized by alignment to a single sequence.

However, recent studies have shown that, simply on the basis of overall similarity, it is generally impossible to infer the function of one protein from another below 40% sequence identity. More pessimistically a study by Tian and Skolnick[3] found that precise function diverges below identities of 60%, which decreases the value of iterative database search methods because confident functional assignment cannot be achieved. In this way, the utility of popular curated databases such as Pfam[4], CDD, PRINTS, and PROSITE[5] is restricted by the ability to correlate protein relationships with a similarity in function. With the experimental determination of many new protein structures in recent years and the development of more sensitive remote homologue detection methods that exploit rapidly growing sequence databases, it has become increasingly likely that a protein of biological interest but unknown three-dimensional structure will have a homologue of known structure.

From a homologue of known structure, it is possible to build a model of the target sequence of unknown structure using methods developed by many research groups over

*Corresponding author: Dr. Vanaja Kumar, Scientist F, Department of Bacteriology, Tuberculosis Research Centre, Mayor V R Ramanathan Road Chetput, Chennai – 600 031, India.

Tel: +91-44-2836 9659

Fax: +91-44-2836 2528

E-mail: vanaja_kumar51@yahoo.co.in

the last 30 years. Comparative modeling helps to bridge the gap between primary and tertiary structure by allowing the construction of models that may be used to identify critical residues involved in catalysis, binding, or structural stability; *etc.* Comparative modeling, or homology modeling, is usually based on a number of steps where identifying the template and alignment between the target and template is the most critical.

For proteins with known structure, we can evaluate the sequence alignment by testing how well it describes the similarity between structures of both proteins. Structure can provide clues to function in many cases, even if powerful sequence methods have failed to provide a conclusive functional assignment.

2. Materials and methods

All the protein sequences were searched against Non-Redundant and Protein Data Bank database using Position-

Specific Iterative BLAST (PSI-BLAST) with default setting. Each of the sequences was also searched against PFAM database to identify their functional domains. Proteins having no detectable similarity against PDB[6] database and having functional domain in PFAM, the PDB ids of the proteins listed in PFAM database for the corresponding domain were taken. Each of the structure was further analyzed using PFAM to identify different domains present in the selected structures.








3. Results

In the current study we have selected 32 mycobacteriophages (Table 1) consisting of 3 400 proteins. These proteins were subjected to PSI-BLAST and PFAM for predicting their functions. Of the 3 400 proteins, for 611 proteins, their functions were predicted based on homology using BLAST analysis. Further for 447 out of 611 proteins, their corresponding domains were mapped using PFAM

Table 1.
Analysis of 32 mycobacteriophages.

Mycobacteriophage genome name	Total number of genes	BLAST (Unknown function)	Number of genes with predicted function Generally using BLAST	Number of genes with predicted function BLAST without domain	Number of genes with predicted function BLAST with Domain	Unknown function using BLAST with Domain predicted by Pfam
DD5	87	63	24	15	9	0
Fruitloop	102	86	16	2	14	2
Gumball	88	72	16	5	11	1
Predator	92	83	9	3	6	0
Jasper	94	73	21	10	11	1
KBC	89	68	21	10	11	0
Pukovnik	88	68	20	5	15	1
Kostya	143	121	22	4	18	0
Lockley	90	66	24	12	12	0
Konstantine	95	85	10	1	9	0
Nigel	94	81	13	8	5	0
Ramsey	108	82	26	10	16	0
Solon	86	65	21	10	11	1
adjutor	86	74	12	2	10	1
angel	61	49	12	6	6	0
Boomer	105	82	23	7	16	0
Bps	63	46	17	10	7	0
Brujita	74	57	17	5	12	0
Butterscotch	86	72	14	3	11	1
Chah	104	93	11	4	7	0
Cjw1	141	117	24	3	21	0
Wildcat	148	125	23	3	20	0
Che8	112	92	20	6	14	0
Che12	98	70	28	13	15	1
Plot	89	75	14	3	11	1
Porky	147	122	25	5	20	0
Troll4	84	72	12	2	10	1
Pacc40	101	80	21	4	17	0
Phaedrus	98	86	12	4	8	0
Phlyer	103	90	13	4	9	0
Spud	222	191	31	7	24	2
Cali	222	194	28	5	23	3
Total	3400	2800	600	191	409	16

Table 2.
Identification of homologous proteins having experimental structures.

Protein name	Genome name	Homologues against PDB	Domains	Structures with similar domain (remote homologues)
Glycosyl transferase family 2	Cali-gp232, Spud-gp234	NA		1H71, 1H7q, 1QG8, 1QGQ, 1QGS (15 more)
GerE Bacterial regulatory proteins, luxR family	Fruitloop-gp57	NA		1A04, 1FSE, 1H0M, 1JE8, 1L3L (15 more)
Metallophos Calceineurin-like phosphoesterase domain	Jasper-gp53, Pukovnik-gp55, Solon-gp50, Che12-gp57	NA		1aui, 1fjm, 1g5b, 1ho5, 1hp1(69 more)
Ribbon-helix-helix protein, copG family Domain	Fruitloop-gp72	NA		1B01, 1EA4, 1Q5V, 1X93, 2BA3 (12 more)
ParB-like nuclease domain	Cali-gp78, Spud-gp82	NA		1R71, 1VK1, 1VZ0, 1XW3, 1XW4 (4 more)
LysM domain	Cali-gp92	NA		1E0G, 1Y7M, 2DJP
Phage_prot_Gp6 Phage portal protein, SPP1 Gp6-like	GUMBALL_8, Konstantine-gp7, Adjutor-gp9, Butterscotch-gp9, Plot-gp9, Troll4-gp9	NA		2JES

database (Table 2). The remaining major chunk of 2 342 proteins (69%) is still left in the lurch which could be either not having any homolog's with known function or incapable of picking up the remote homologues. Here it may be mentioned that many in the unknown may be predicted by this approach if the domain information and solved structure were available for related proteins, then this number would increase. We have still a long way to go in order to be able to justify and predict functions for the remaining huge percentage only by other parallel methods of approaches like modeling and threading methods.

It is unfortunate that the function for the proteins are being missed or undetected by BLAST across the mycobacteriophage genomes. This loss is depicted in Table 2.

4. Discussion

For a very meager number of sequence domains, their three dimensional structures are available in the Protein Data Bank which have their sources from other organisms. The structures of the domain available will pave way to predict the 3 dimensional structures for the target and elucidate their function and detect their catalytic triad which was not otherwise possible by sequence based search.

The 31% of proteins with predicted function, the templates were not identified by searching against PDB using BLAST. Hence in these situation proteins with known PFAM domains the PDB ID's for the corresponding domains were selected as templates can be selected for modeling studies.

Conflict of interest statement

We declare that we have no conflict of interest.

Acknowledgement

The authors wish to acknowledge the financial support from Department of Biotechnology, New Delhi, India.

References

- [1] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997; **25**(17): 3389–3402.
- [2] George RA, Spriggs RV, Bartlett GJ, Gutteridge A, MacArthur MW, Porter CT, et al. Effective function annotation through catalytic residue conservation. *Proc Natl Acad Sci USA*. 2005; **102**(35): 12299–12304.
- [3] Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies from a structural perspective. *J Mol Biol* 2001; **307**(4): 1113–1143.
- [4] Bateman A, Birney E, Cerruti L, Durbin R, Ewlinger L, Eddy SR, et al. The Pfam protein families database. *Nucleic Acids Res* 2002; **30**(1): 276–280.
- [5] Edouard de Castro, Christian JA Sigrist, Alexandre Gattiker, Virginie Bulliard, Petra S Langendijk-Genevaux, Elisabeth Gasteiger, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* 2006; **34**(2): W362–W365.
- [6] Berman HM. The protein data bank. *Nucleic Acids Res* 2000; **28**(1): 235–242.