Inquiro, the UAB undergraduate science research journal

Volume 2014 | Number 8 Article 24

2014

Structural Genomics and Its Importance for Gene Function

Michael Falahat

Denise Monti

Structural Genomics and Its Importance for Gene Function

Michael Falahat and Denise Monti

Department of Biology, University of Alabama at Birmingham, Birmingham, AL, USA

Abstract

This study sought to test whether current bioinformatics programs are sufficient to extract valuable data regarding a gene found in the mycobacteriophage *Holli*. Specifically, we examined whether current bioinformatics tools could be used to predict putative gene functions for unknown genes, in this case *Holli* gp 54 (gene product 54), located on the right arm of the genome map. We used protein prediction websites such as I-TASSER to aid in gene function prediction. We found that gene function could not be predicted based on results collected from current bioinformatics tools. We concluded that current bioinformatics tools do not provide concrete gene function predictions, though they do help us in selecting noteworthy genes that can be studied through wet lab experiments for more accurate data.

Introduction

Bacteriophages are viruses that infect bacteria. Phages are found everywhere and are much smaller than their hosts. The basic structure of a phage consists of a capsid, where its DNA or RNA genome is stored, and a tail.[1] A specific type of bacteriophage, called a mycobacteriophage, is a bacteriophage that targets bacteria in the genus *Mycobacterium*. The mycobacteriophage settles onto the surface of the mycobacterium host and then injects its genomic DNA into the host cell. Next, phages reproduce inside the microbe to multiply and eventually explode out of (i.e., "lyse") the host.[2]

Identifying functions for phage genes can be very useful for phage therapy, in which bacteriophages are used to treat infections caused by pathogenic bacteria, by facilitating more targeted and controlled therapy.[3] Many protein-coding genes found in a mycobacteriophage are unique for that phage. The function of gene 54 in the mycobacteriophage *Holli* is unknown and is currently being studied. Gene 54 is located on the right arm of the genome map and consists of 91 base pairs. In recent years, there has been an exponential increase in the number of genomes sequenced in full, but the sequences of genes and their corresponding protein products alone do not provide insight into gene functions in the cell.[4] Rather, examination of the structures of unknown proteins may help us to predict the specific function of the gene product. To understand the function of proteins, wet lab experiments such as recrystallization of the protein can be performed. However, this method is costly and time consuming.

In this study, we used current bioinformatic tools to identify possible structures for putative gene products of gene 54. A program called "Phamerator" helped us to identify the phams, or related sequences, and the clusters to which the gene belongs. [5] In this program, getting the sequence for the same gene from different phages was possible. Programs like Clustal Omega and T-COFFEE were used to compare the sequences of genes from different phages and gave us valuable information that was used to help predict the function of the gene.

Understanding protein functions can help with understanding cell function. When human cells do not work properly, understanding altered protein functions can help with the development of appropriate strategies to treat various diseases. Phage therapy, in which bacteriophages are used to treat bacterial infections, is an active field of research. Understanding the function of proteins within mycobacteriophages can potentially help with the treatment of human diseases.

Materials and Methods

Genes were termed using DNAMaster, a genome annotation and exploration tool. A gene of interest was selected for function prediction. Phamerator was used to identify the gene's pham and clusters represented in the pham. Clustal Omega (Figure 2) was used to assess the conservation of amino acids amongst genes in the same pham. I-TASSER was used to model possible structures for the gene being studied, and C-scores for these structures were determined. The C-score represents the confidence score for estimating the quality of predicted models. It is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations.[7] Higher C-scores represent more reliable structures. Z-scores were also determined. A Z-score greater than 1 represented a good alignment, and in general the Z-score represents the difference between the raw and average scores in units of the standard deviation. Proteins with high structural similarities in PDB (Protein Data Bank) were analyzed for a possible protein function based on structural similarities and protein structures stored in PDB. TM-scores, which compares the query protein structure and the structure in PDB based on their given and known residue equivalency, were taken into consideration as well. TM-score is the scale for measuring the structural similarity between two structures. Higher TM-scores represent better structural alignment, a TM-score

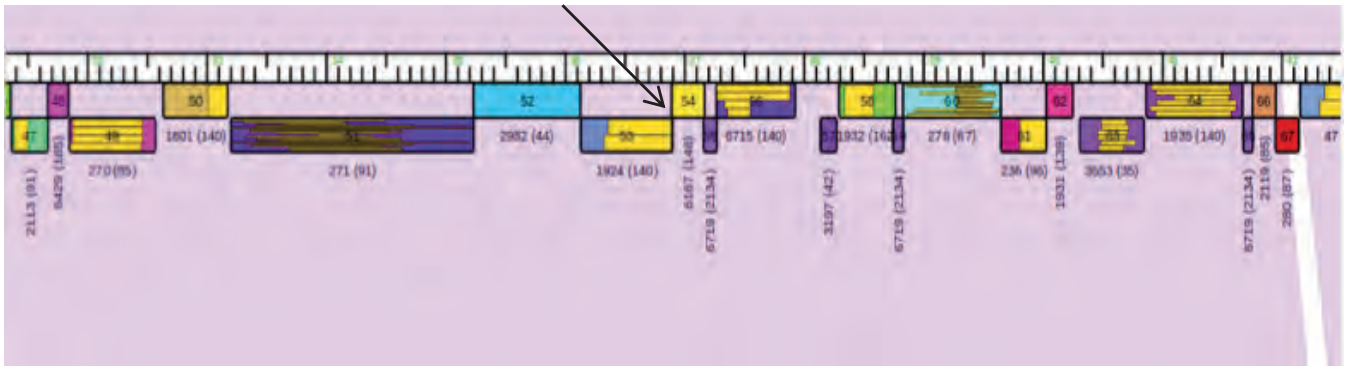

Figure 1. Section of a genome map of mycobacteriophage Holli *showing the gene studied (gene 54) in the right arm of the genome. The number 6167 under gene 54 indicates the pham number, and the number in parentheses indicates the clusters (A4) contained in that pham.*[8]

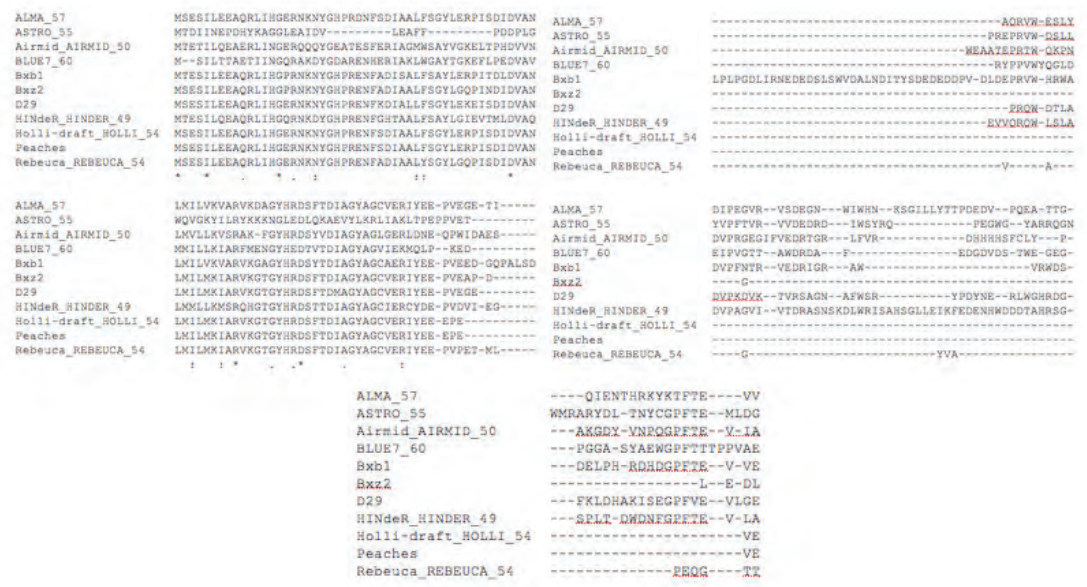

Figure 2. Protein sequence alignment for A1-A10 gene using Clustal Omega

This shows an alignment for gene A1-A10. The star under some of the columns indicate that the nucleotide was a one-to-one match for all the different phages; however, this is a poor alignment since it did not contain many stars. These data were studied and based on this alignment no observation was seen.

greater than 0.5 indicates a model of correct topology, and a TM-score less than 0.17 indicates a random similarity.[7]

Results

Gene 54 is a gene found in mycobacteriophage *Holli*. Part of the genome map for *Holli* is shown in Figure 1. A genome map consists of a left arm and a right arm. This section of the genome map shows the genes that are contained in this section of the right arm of the phage *Holli*. Each box represents a gene, and the width of the box indicates the length of the gene in base pairs. *Holli* gp 54, or gene product 54, is located in the right arm of the genome map. This gene is 91 base pairs long and has 1:1 match with gp 53 of the mycobacteriophage Peaches.[6] *Holli* gp 54 has a pham number of 6167. This pham contains 54 clusters, including A1 through A2, J, and K4.

Based on the data collected from the bioinformatics programs used, gene function could not be assigned to the gene found in *Holli*. The same gene found in clusters A1 through A2 was submitted to Clustal Omega for alignment (Figure 2). The protein sequence for the gene found in *Holli* was submitted to I-TASSER, an online platform for protein structure and function predictions. I-TASSER calculated a C-score representing a confidence score for estimating the quality of predicted models based on the significance

of threading template alignments and the convergence parameters of the structure assembly simulations. The four models with the highest C-scores predicted by I-TASSER are presented in Figure 3. The C-scores, ranging between −5 and −2, were poor. The best C-score was −2.29 (Figure 3).

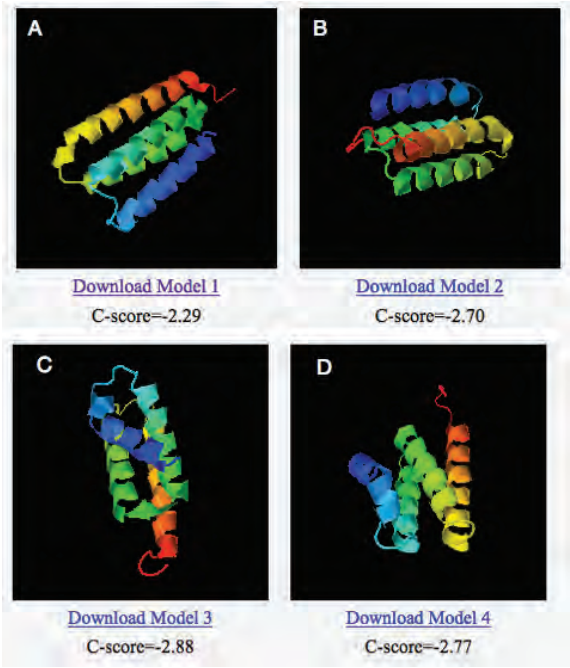

Figure 3. Top four protein models predicted by I-TASSER for gene in phage Holli

I-TASSER generates tens of thousands of conformations, called decoys. Based on the pair-wise structure similarity, I-TASSER reports up to five models that correspond to the five largest structure clusters. The C-score (confidence score) indicates the structures with the largest partition function (or lowest free energy).[4]

I-TASSER also generates many templates; however, I-TASSER uses templates with the highest significance in the threading alignments. This significance is measured by the Z-score, which is the difference between the raw and average scores in units of standard deviation.[7] Each threading template used by I-TASSER contains a normalized Z-score of the threading alignment. In Table 1, the top three threading templates showed poor Z-scores of 0.82, 0.60, and 0.75. These results represent poor normalized Z-scores because they are less than 1. The top threading template, a virulence factor from *Campylobacter jejuni* (1vqrA), has an unknown function; an anthranilate phosphoribosyl-transferase (1o17A) codes for a transferase; and a core-binding domain of bacteriophage lambda integrase (2oxoA) codes for a DNA-binding protein.

Proteins with highly similar structure, as indicated by I-TASSER, were also studied (Table 2). The top three proteins had TM-scores of 0.812, 0.737, and 0.727; however, percentage identities were extremely poor. Percentage identities for the top three proteins were 0.043, 0.043, and 0.00. Prediction could not be made according to this data. Model structures of proteins whose structure is similar to the predicted structure of the gene product of *Holli* gene 54 are shown in Figure 4. Two structures were missing a large portion of the predicted structure of this gene product (Figure 4A and 4B). A third structure (Figure 4C) had no missing portions, but its identity was 0.

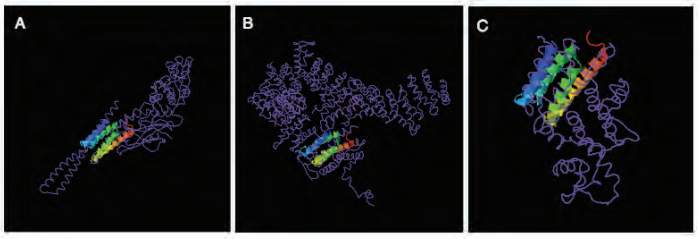

Figure 4. Structures with highly similar structure in PDB to Holli *gene 54*

Structures A, B, and C show the structures for proteins with structures highly similar to that of Holli gene 54 in PDB from Table 2. Structures A and B are missing a large portion of the protein it has been compared to. While structure C contains most of the structure it has been compared to, its identity is 0 according to Table 2.

The same gene from mycobacteriophage Bxb1 was submitted to I-TASSER. Results as shown in Figure 5 had a lower C-score in the top models when compared to the models for the gene in *Holli*. The second model in both Bxb1 and *Holli* codes for a transferase enzyme. This prediction of function could not stand because of a poor Z-score. Figure 6 shows proteins with high structural similarities in PDB for Bxb1. Protein A was predicted to function as a hydrolase inhibitor, and proteins B and C were predicted to function as protein transporters. Compared to the functions observed in *Holli*, the predicted functions were different. In *Holli*, the top three proteins had functions of a transcription protein, a DNA-binding protein, and a signaling protein. Function could not be predicted based on these results.

Table 1. Top 3 Threading Templates for Holli *gene*

Rank	PDB Hit	Iden1	Iden2	Cov.	Norm. Z-score
1	1vqrA	0.12	0.21	0.88	0.82
2	1o17A	0.16	0.18	0.84	0.60
3	2oxoA	0.12	0.24	0.97	0.75

Table 2. Proteins with highly similar structure in PDB (Protein Data Bank) for Holli *gene*

Rank	PDB Hit	TM-score	RMSD[a]	IDEN[a]	Cov.
1	1y1uC	0.812	2.26	0.043	0.989
2	3s4wA	0.737	2.70	0.043	0.967
3	3t5vA	0.727	2.54	0.000	0.967

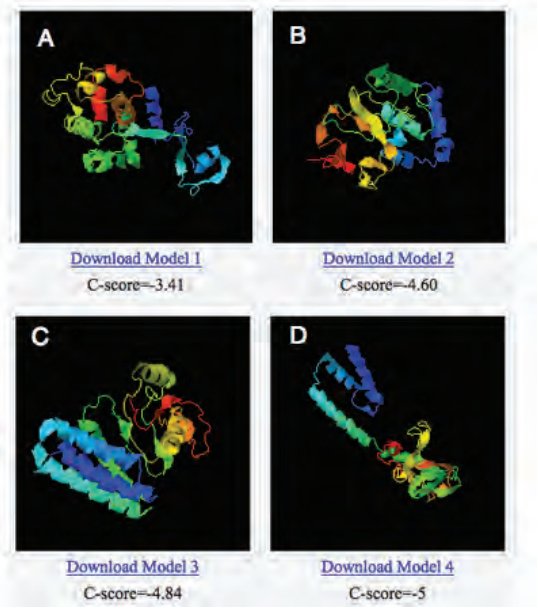

Figure 5. Top four models predicted by I-TASSER for Bxb1 *gene*

For further investigation on predicting the function of the protein encoded by the gene studied in Holli*, the same gene from* Bxb1 *(an A1 subcluster) was also submitted to I-TASSER.* Bxb1 *is one of the members of pham 6167. These top models predicted by I-TASSER had an extremely low C-score. These models look significantly different than the models predicted for the gene in* Holli*. Despite the significant difference in structure, the second model in both* Holli *and* Bxb1 *functions as a transferase; however, this is not enough evidence to support this prediction due to the poor Z-score.*

Discussion

Data gathered from various programs such as I-TASSER indicate that predicting a function for the gene *Holli* using current bioinformatics tools is impossible. Protein structures sent from I-TASSER for the gene in *Holli* did not provide sufficient data to provide predictions of a gene's function. More data were gathered when the gene found in Bxb1 was submitted to I-TASSER. Comparisons among proteins with highly similar structures in PDB could be made according to the data collected from Bxb1; however, inconsistency exists between the data from *Holli* and that from Bxb1. As a result, there were not sufficient results to predict a function.

During the study, we observed that the second best threading template sent from I-TASSER for both *Holli* and Bxb1 coded for a transferase. We came close to predicting that this gene codes for a transferase; however, the Z-score for this threading template in *Holli* is 0.62, which is low. A good Z-score is greater than 1, representing a good alignment. When looking at the second best threading template for Bxb1, the Z-score was 0.58, which is also poor. While it was predicted that the gene coded for a transferase enzyme, there was not strong evidence for this prediction based on the results collected.

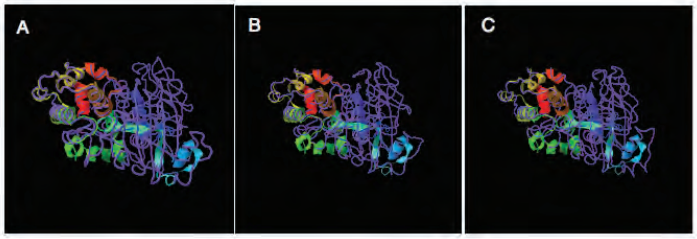

Figure 6. Proteins with highly similar structure in PDB (Bxb1)

In the above figures, protein A represents a hydrolase inhibitor. Proteins B and C represent a protein transporter. These functions are different than the proteins with highly similar structure in PDB for the gene in Holli*. In* Holli*, the top three proteins represents a transcription protein, a DNA-binding protein, and a signaling protein. Due to these dissimilarities, gene function could not be predicted for the gene found in* Holli*.*

Our study indicates that wet lab experiments must be performed to provide accurate function predictions of a gene, despite the fact that this process is costly and time consuming. Our study did show that bioinformatics tools can provide weak predictions to help narrow down or speculatively identify the functions of proteins in a cell. Wet lab experiments can then be guided and performed based on the bioinformatics predictions. Although the gene being studied in *Holli* did not provide adequate information to predict protein function, knowing this type of information helped us determine which genes in *Holli* are worth investigating using wet lab experiments.

References

1. Cresawn, S.G. et al. Comparative genomics of Cluster O mycobacteriophages. *PLOS ONE*. 10, e0118725 (2015).
2. Travis, J. All the world's a phage. *Science News*. 164, 26–28 (2003).
3. Khan Mirzaei, M. & Nilsson, A.S. Isolation of phages for phage therapy: A comparison of spot tests and efficiency of plating analyses for determination of host range and efficacy. *PLOS ONE*. 10, e0118557 (2015).
4. Skolnick, J., Fetrow, J. S., & Kolinski, A. Structural genomics and its importance for gene function analysis. *Nature Biotechnology*. 18, 283–287 (2000).
5. Cresawn, S. G. et al. Phamerator: A bioinformatic tool for comparative bacteriophage genomics. *BMC Bioinformatics*. 12, 395 (2011).
6. Broadway, L. & Engelsen, A. Details for Phage Peaches. *Phagesdb*. (2008). http://phagesdb.org/phages/Peaches/
7. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. The I-TASSER suite: Protein structure and function prediction. *Nature Methods*. 12, 7–8 (2015).
8. Lawrence, J. DNA Master. Department of Biological Sciences. *University of Pittsburgh*. (2009).