# Chapter 43
# Taxonomic Parsing of Bacteriophages Using Core Genes and In Silico Proteome-Based CGUG and Applications to Small Bacterial Genomes

**Padmanabhan Mahadevan and Donald Seto**

**Abstract** A combined genomics and in situ proteomics approach can be used to determine and classify the relatedness of organisms. The common set of proteins shared within a group of genomes is encoded by the "core" set of genes, which is increasingly recognized as a metric for parsing viral and bacterial species. These can be described by the concept of a "pan-genome", which consists of this "core" set and a "dispensable" set, i.e., genes found in one or more but not all organisms in the grouping. "CoreGenesUniqueGenes" (CGUG) is a web-based tool that determines this core set of proteins in a set of genomes as well as parses the dispensable set of unique proteins in a pair of viral or small bacterial genomes. This proteome-based methodology is validated using bacteriophages, aiding the reevaluation of current classifications of bacteriophages. The utility of CGUG in the analysis of small bacterial genomes and the annotation of hypothetical proteins is also presented.

**Keywords** Bacterial genomes · CGUG · Core genes · Genomics · Pan-genome · Proteomics approach · Web-based tool

## 43.1 Introduction

The continuing and predicted explosion of whole genome data is staggering, with de novo determinations of genomes from previously unsequenced genomes as well as multiple determinations of related genomes, e.g., multiple *Escherichia coli* genomes. If this can be pictured as a "tsunami" to place visually the enormity of the data flow and to understand the immense amount of data that need to be mined (and given a relatively sparse array of software tools to do so), then the parallel and earlier capture of bacteriophage and small bacterial whole genome data may be described as a

P. Mahadevan (✉)
Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA
e-mail: padmahadevan@gmail.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978-1-4419-5913-3_43, © Springer Science+Business Media, LLC 2010

“seiche,” both in terms of the smaller sizes of the genomes they represent and the apparent smaller “stage” upon which they play. Nonetheless, bacteriophages are long-studied and remain valuable for their contributions to basic biology knowledge.

“CoreGenesUniqueGenes” (CGUG) is a user-friendly web-based tool that performs “on-the-fly” protein–protein analyses in order to determine the “core” set of proteins of a set of small genomes. It is a redevelopment of an existing tool, CoreGenes [1], and is an enhancement based on suggestions from the wet-bench bacteriophage research community. CGUG has been validated in two different functions, as will be presented in this report. First is in the annotating and comparing of small bacterial genomes, ca. 2 Mb and less. The second is in reexamining the complex and long-standing relationships of bacteriophages. The latter is very important, given the extensive genetic and biochemical studies in the past. Integrating genomics and in silico proteomics with that data gives another dimension to the value of genome determinations and databases.

The International Committee on the Taxonomy of Viruses (ICTV) is an organization of researchers in a particular field that considers the classification of viruses in their field [2], for example, the bacteriophages. A current classification system accepted by the ICTV is based on a hierarchical system that groups viruses by the characteristics that they share. These characteristics reflect the technologies and methodologies that were available at the time, and may be limited by the same. Specifically, in the case of bacteriophages, in the past and currently, these metrics include morphology, genome size, host range, proteins (immunochemistry), and the physical characteristics of the genome, e.g., whether the genome is linear, circular, or supercoiled [3]. In response to the newly available genome and the resulting proteome data, researchers are now considering these data in the scheme of viral relationships and classifications.

As an example of this, a proteome-based approach has been used recently to reexamine and suggest a reclassification of bacteriophages by computationally building a proteome tree based on BLASTP analyses [4]. The disadvantage of this approach is that there was no readily accessible tool that performs this analysis and generates the proteome tree for inspection and analysis. Ideally, a web-based tool that performs the BLASTP analyses and produces easily interpretable output would be very useful to wet-bench biologists. CoreGenes is a tool that was developed earlier and is used currently to determine the “core” or common set of proteins in a set of genomes. CGUG is an upgrade and a modification that incorporates suggestions from several members of the ICTV who were interested in using the software for their research. Core sets of genes have been used to reconstruct ancestral genomes [5], organismal phylogenies [6], and organism classifications [7]. It can be and has been applied to the bacteriophage studies [7].

## 43.2 CoreGenesUniqueGenes Algorithm

CGUG is implemented in the Java programming language, using a combination of servlets and HTML. The algorithm is based on the GeneOrder algorithm to determine gene order and synteny [8]. The algorithm accepts between two and five genome accession numbers. These genomes are then retrieved from GenBank, and the protein sequences are parsed and extracted from the GenBank files. One genome is designated as the reference genome, and the rest are the query genomes. If only two genome accession numbers are entered, all against all protein similarity analyses are performed for each protein of the query genome against the reference genome using WUBLASTP from the WUBLAST package. The results from the protein similarity analyses are parsed according to a previously specified threshold BLASTP score (default = "75"). If the scores from the protein alignments are equal to or greater than the threshold score, the protein pairs are stored and a table of proteins common to the two genomes (that is, "core" proteins) is created as the output.

In the case of more than two genomes, a consensus genome is created from the results of the similarity analysis between the reference genome and the first query genome. This consensus genome becomes the new reference genome. Protein similarity analyses are performed with the second query genome against this new reference genome using WUBLASTP. The algorithm proceeds in an iterative manner, analyzing the subsequent query genomes against the newly created reference genomes. The final output is a set of "core" proteins between a set of up to five small genomes in the form of a table. The unique proteins to a pair of genomes are also presented in tabular format below the "core" protein table. CGUG is available at http://binf.gmu.edu:8080/CoreGenes3.0 and can also be accessed at http://binf.gmu.edu/geneorder.html.

At the request of wet-bench bacteriophage researchers, a homolog count function has also been implemented. This is displayed as the sum of proteins in each column of the table. In addition, in recognition that some genomes may be newly sequenced and desired to be analyzed before submission into public databases, custom data can also be entered for CGUG analysis using the "custom data" interface.

The groups analyzed to demonstrate the function of reconfirming and verifying existing (ICTV) genera are the T7-like bacteriophages (*Escherichia* phage T7, *Yersinia* phage φA1122, *Yersinia* Berlin, *Escherichia* phage T3, *Yersinia* phage φYeO3-12, *Escherichia* phage K1F, *Vibrio* phage VP4, and *Pseudomonas* phage gh-1). Based solely on the GC content and length, it may be difficult to gain meaningful information about the bacteriophages in order to classify them. But an analysis of their proteomes using CGUG yields more informative results. All CGUG analyses are performed at the default threshold setting of "75."

## 43.3 Results and Discussion

### *43.3.1 Bacteriophage Genomes Application: Verification of Existing Genus of the Podoviridae*

Bacteriophages are notoriously difficult to classify because of their genome variations due to horizontal transfers [9]. Recently, several researchers, who are members of the ICTV, have used CoreGenes to reanalyze and reclassify bacteriophages using proteome data [7]. Their previous experiences in applying CoreGenes to earlier work gave insights as to what additional features were needed; these have been incorporated into CGUG and integrated into the later portions of their analyses of these genomes. The bacteriophage families examined to date include the Podoviridae and the Myoviridae [10]. To illustrate the usefulness of CGUG, the T7 genus reanalysis data from a recent collaborator [7] are presented here to emphasize the utility of this approach in the reclassification of the bacteriophages.

The T7-like phages constitute a genus of the Podoviridae family. Using a homologous protein cutoff of 40%, CGUG analysis reveals that the members of the T7-like phages all share greater than 40% homologous proteins with bacteriophage T7. This cutoff is used because it has been used previously to produce clear relationships between bacteriophage genera of the Podoviridae [7]. This shared protein analysis reconfirms and verifies the existing ICTV classification of these phages as belonging to the T7-like phage genus.

### *43.3.2 CGUG Analysis and Reclassification of T7, P22, and Lambda Bacteriophages*

The tailed bacteriophages T7 and P22 are currently and traditionally classified as belonging to the Podoviridae family due to a shared presence of short tails [3]. However, P22 and lambda are more related to each other than to T7, based on the CGUG in silico proteomics analysis. Therefore, P22 should be moved to and classified in the Siphoviridae to which lambda belongs. CGUG analysis reveals that T7 and P22 share only two proteins. T7 and lambda also share only two proteins. The percent identities of these shared proteins are very low, ranging from 14 to 20%. In contrast, P22 and lambda share 19 proteins, several of which show high percent identities (>80%).

Thus, these results show that P22 is more related to lambda than to T7, given the genome and the proteome data. The whole genome percent identity between T7 and P22 is 47.5%, while the identity between T7 and lambda is approximately 46%. This nucleotide level does not provide meaningful information about the relatedness of these genomes. Similarly, the percent GC content of these genomes (between 48 and 50%) is also not informative. The CGUG analysis looks at the in silico proteome, and the relatedness of the T7, P22, and lambda phages can be assessed more meaningfully using this information.

### *43.3.3 Application to Niche Specific Bacterial Strains*

The transcriptomes of two closely related strains of *Burkholderia cenocepacia* were recently mapped by high-throughput sequencing [11]. *B. cenocepacia* strain AU1054 is an opportunistic pathogen found in cystic fibrosis patients, while the

*B. cenocepacia* strain HI2424 is a soil-dwelling organism. Despite the fact that these two organisms live in very different environments, they share 99.8% nucleotide identity in their conserved genes. Even in such highly related bacterial strains, there are cases of hypothetical proteins in one strain that are not annotated with a function that is related to annotated proteins in the other strain. One example is the case of the 3-carboxy muconate cyclase-like protein found in AU1054, while the counterpart protein is annotated as hypothetical in HI2424. These two proteins share only 12.1% identity to each other. However, their lengths are not very dissimilar, and analyses using PFAM (http://pfam.sanger.ac.uk) show that the hypothetical protein appears to contain a "3-carboxy-cis,cis-muconate lactonizing enzyme" domain. Therefore, it is possible that the hypothetical protein shares a function similar to that of the annotated protein in AU1054.

In contrast, there is a case where the hypothetical protein is in AU1054, while the counterpart annotated protein is in HI2424. This annotated protein is an amidohydrolase. The percent identity between these proteins is 19.7%, and their lengths are similar as well at 319 amino acids for the amidohydrolase and 281 amino acids for the hypothetical protein. Functional prediction of this hypothetical protein using the SVMProt server [12] indicates that it belongs to the iron-binding protein family, with a probability of correct classification of 78.4%. Indeed, several enzymes in the amidohydrolase superfamily bind metal ions [13]. The catalytic activity of one amidohydrolase, cytosine deaminase, is highest with iron [14]. Analysis using the Phyre fold recognition server [15] indicates that the hypothetical protein apparently belongs to the "N-terminal nucleophile aminohydrolases". However, it must be noted that the E values from Phyre and the percent precision are not significant. Nevertheless, this taken together with the fact that CGUG puts these two proteins together suggests that the hypothetical protein may indeed be an amidohydrolase. Further wet-bench experiments are needed to confirm this prediction.

## 43.4 Conclusions

Whole genome and in silico proteome analysis tools are necessary to obtain meaningful information about organisms when the nucleotide data are not especially informative. CGUG is a user-friendly tool that is especially suited to wet-bench biologists with little interests in compiling code or deconstructing software in order to analyze proteomes from small genomes such as those from viruses, chloroplasts, and mitochondria, as well as small bacterial genomes. The utility of CGUG is illustrated in the verification of current classifications of bacteriophages and in the proposal of new classifications based on CGUG and other data. The current ICTV classification of the T7-like phages is verified using the whole proteome CGUG analysis.

Currently, the T7 and P22 phages are classified in the Podoviridae, while lambda phage is classified in the Siphoviridae. The in silico proteome analysis by CGUG shows that P22 is more related to lambda than to T7. That is, P22 and lambda share more proteins with each other than they both do with T7. This means that P22 should be classified in the Siphoviridae like lambda. In this case, proteome analysis is more meaningful than the shared morphological similarity between T7 and P22.

The usefulness of CGUG in annotating hypothetical proteins is illustrated in the case of the two closely related niche-specific *Burkholderia* bacteria. The assignment of putative functions to these hypothetical proteins will provide a starting point to help wet-bench scientists confirm these predictions in the lab.

The web-based nature of CGUG makes it accessible and useful to wet-bench-based biologists. The "on-the-fly" nature of the tool avoids the limitations of precomputed data and allows the retrieval of the genomes directly from GenBank. The ability to enter custom data further enhances the tool immensely. These types of software tools allow biologists to take full advantage of the expanding genome sequence databases.

**Acknowledgments** We thank Drs. Andrew Kropinski and Rob Lavigne for their very helpful suggestions of features to add to CoreGenes. We thank Jason Seto for editorial comments and Chris Ryan for maintaining the server on which the software is hosted. This paper is based on Mahadevan and Seto (BIOCOMP '09).

## References

1. Zafar N, Mazumder R, Seto D (2002) CoreGenes: a computational tool for identifying and cataloging "core" genes in a set of small genomes. BMC bioinformatics **3**, 12
2. Fauquet CM, Fargette D (2005) International Committee on Taxonomy of Viruses and the 3,142 unassigned species. Virology journal **2**, 64
3. Nelson D (2004) Phage taxonomy: we agree to disagree. Journal of bacteriology **186**, 7029–7031
4. Rohwer F, Edwards R (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. Journal of bacteriology **184**, 4529–4535
5. Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nature reviews **1**, 127–136
6. Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS biology **1**, E19
7. Lavigne R, Seto D, Mahadevan P, Ackermann H-W, Kropinski AM (2008) Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. Research in microbiology **159**, 406–414
8. Mazumder R, Kolaskar A, Seto D (2001) GeneOrder: comparing the order of genes in small genomes. Bioinformatics (Oxford, England) **17**, 162–166
9. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proceedings of the National Academy of Sciences of the United States of America **96**, 2192–2197
10. Lavigne R, Darius P, Summer EJ, Seto D, Mahadevan P, Nilsson AS, Ackermann HW, Kropinski AM (2009) Classification of Myoviridae bacteriophages using protein sequence similarity. BMC microbiology **9**, 224
11. Yoder-Himes DR, Chain PS, Zhu Y, Wurtzel O, Rubin EM, Tiedje JM, Sorek R (2009) Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing. Proceedings of the National Academy of Sciences of the United States of America **106**, 3976–3981

12. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic acids research **31**, 3692–3697
13. Sadowsky MJ, Tong Z, de Souza M, Wackett LP (1998) AtzC is a new member of the amidohydrolase protein superfamily and is homologous to other atrazine-metabolizing enzymes. Journal of bacteriology **180**, 152–158
14. Porter DJ, Austin EA (1993) Cytosine deaminase. The roles of divalent cations in catalysis. Journal of biological chemistry **268**, 24005–24011
15. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the web: a case study using the Phyre server. Nature protocols **4**, 363–371