Identification of Tandem Repeats in Geobacillus and Heamophilious Phages

Introduction:

 Repeated sequences are DNA sequences witch occur with some periodicity and have a distinct pattern. These sequences are present in a wide variety of organisms and have a number of functions. One type of repeated sequence is a tandem repeat, one of the simpler kinds of repeats. Tandem repeats present themselves as a set of one or more nucleotides repeated directly adjacent to one another. When the total length of the repeat is between 10 and 60 nucleotides it is referred to as a mini-satellite, those with fewer are known as micro-satellites, or short tandem repeats3-5. Tandem repeats are especially important as tools in forensic identification and are associated with many genetic disorders, one class of which are the trinucleotide repeat disorders1. This set of disorders includes Huntington’s disease and Fragile X syndrome1.

 While the study of tandem repeats is still ongoing, much less work has been conducted in bacteria and bacteriophages as has been conducted in humans. The similar principles of forensic identification using tandem repeats can be conducted in bacteria as is humans, however characterization of tandem repeats on an organism per organism basis is still not widespread6-7. Here we attempt to identify possible tandem repeats in two heamophilious and two geobacillus bacteriophages phages and characterize where they occur.

Methods:

 In order to identify tandem repeats in the 4 organisms of interest, geobacillus phages E2 and GBSV1 as well as heamophilious phages HP1 and HP2, their full genomes were run through an autocorrelate function in BioBike/PhAnToMe Bike8. The autocorrelate function, shown in figure 1, finds tandem repeats by iterating over the sequence of interest, in this case the phage genomes, in 100 nucleotide increments and analyzes each position and the 7th position after it for a match. In other words, a frame of 7 is shifted over the 100 nucleotide sequence, assessing the first nucleotide of each frame to the next. The autocorrelate function scores each 100 nucleotides by adding a value of 1 for each match. The score for the 100 nucleotides is determined and the next 100 nucleotides are analyzed.



Figure 1: The autocorrelate function from Biobike.

Results:

 The full genomes of the aforementioned phages were run through this function, and a plot generated for each genome (Figure 2). The results were then analyzed for their location, as well as the associated genes and information (Table 1). All four of the phages had autocorrelation scores over 40, though none had scores over 50. The likelihood of a score of 40 occurring naturally is 0.039%.



Figure 2: The plotted results of the autocorrelation that was run on the 4 phages. Starting from the top left clockwise they are Heamophilious HP1, Heamophilious HP2, Geobacillus GBSV1, Geobacillus E2. The y-axis is a measure of the score, and the x axis represents the position in the genome.

|  |  |  |  |
| --- | --- | --- | --- |
| Organism | Start Coordinate (s) | Score(s) | Context |
| Heamophilious HP1 | 116 | 41 | Intergenic, HP1p42-HP1p01 |
| Heamophilious HP1 | 17864-17865, 17868-17906 | 41-44 | HP1p27 |
| Heamophilious HP2 | 118 | 41 | Intergenic, HP2p37-HP2p01 |
| Heamophilious HP2 | 2165-2168, 2170-2171 | 41-42 | HP2p02 |
| Heamophilious HP2 | 13309 | 41 | HP2p17 |
| Heamophilious HP2 | 29241 | 41 | HP2p36 |
| Geobacillus E2 | 31132-31133 | 41 | GBVE2\_gp040 |
| Geobacillus GBSV1 | 6317-6318 | 41-42 | GPGV1\_gp11 |
| Geobacillus GBSV1 | 6319-6325 | 41-42 | Intergenic, GPGV1\_gp11- GPGV1\_gp12 |

Table 1: A charactarization of the tandem repeats found in the 4 phages.

|  |  |
| --- | --- |
| Protein | Function |
| HP1p42 | Phage protein #ACLAME 180 |
| HP1p01 | Phage Integrase, site-specific tyrosine recombinase |
| HP1p27 | Phage tail completion protein |
| HP2p37 | Phage protein #ACLAME 180 |
| HP2p01 | Phage Integrase, site-specific tyrosine recombinase |
| HP2p02 | Phage protein |
| HP2p17 | Phage Capsid scaffolding protein |
| HP2p36 | Phage Protein |
| GBVE2\_gp040 | DNA helicase, phage associated |
| GPGV1\_gp11 | Transcriptional regulator ArpU or LmaC associated with virulence in Listeria |
| GPGV1\_gp12 | Phage Protein |

Table 2: The proteins associated with the identified tandem repeats.

Discussion:

 The tandem repeat analysis yielded some surprising results. Firstly, there were no scores above 45, which is a low score. The lack of reasonably high scores (>50), leads to the suspicion that there exist no significant tandem repeats within the phage genome. While this analysis is not all conclusive, and very like some potential candidates were not caught by this analysis, it is still something that draws attention. Another interesting finding is the overwhelming proportion of possible tandem repeats found within genes. In order to make a more confident conclusion about the distribution of tandem repeats in phage genomes not only do more phages need to be analyzed, but also the proportion of the genome that is non-coding, as phage genome are overwhelmingly coding.

 Insofar as the data as a whole is concerned, no patterns about the genes associated with tandem repeats can be made, and nothing emerges in this analysis. Certain proteins functions could not be ascertained with preliminary investigation in BioBike, thus further analysis is needed. However cursory, this type of analysis is needed for more organisms, and more advanced algorithms are needed to better detect possible tandem repeats. Given that they are an import class of sequences, this kind of analysis may lead to better understanding of their role in phage and bacterial genomes.

References:

1. Siwach, Pratibha. "Tandem Repeats in Human Disorders: Mechanisms and Evolution." *Frontiers in Bioscience*: 4467. Print.
2. Benson, G. "Tandem Repeats Finder: A Program to Analyze DNA Sequences." *Nucleic Acids Research*: 573-80. Print.
3. Van Belkum A, Scherer S, van Alphen L, Verbrugh H. Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Review. 1998, 6(2):275-293.
4. . Lupski J, Weinstock G. Short, interspersed repetitive DNA sequences in prokaryotic genomes. Journal of Bacteriology. July, 174(14):4525-4529
5. Mazel D, Houmard J, Castets AM, and Tandeu N. Highly repetitive DNA sequences in cyanobacterial genomes. Journal of Bacteriology. 1990, 172(5):2755-2761.
6. Lindstedt B, Heir E, Gjernes E, Vardund T, Kapperud G. DNA Fingerprinting of Shiga-toxin producing Escherichia coli O157 based on multiple-locus variable-number tandem-repeats analysis (MLVA). Annals of Clinical Microbiology and Antimicrobials. 2003, 2:12.
7. Lindstedt, Bjørn-Arne. "Multiple-locus Variable Number Tandem Repeats Analysis for Genetic Fingerprinting of Pathogenic Bacteria." Electrophoresis: 2567-582. Print.
8. Elhai, J., A. Taton, J. Massar, J. K. Myers, M. Travers, J. Casey, M. Slupesky, and J. Shrager. "BioBIKE: A Web-based, Programmable, Integrated Biological Knowledge Base." Nucleic Acids Research (2009): W28-32. Print.
9. Korcovelos E (2014). Determination of tandem repeats in Bacillus bacteriophages.