Analysis DNA uptake sequences in pathogenic species of the *Haemophilus* genus

**Introduction**

 A large portion of eukaryotic genomes are non-coding, meaning that they do not encode protein sequences. Indeed, more than 98% of the human genome does not encode protein sequences (Elgar, Vavouri 2008). Bacterial genomes, on the other hand, do not possess nearly as many noncoding regions, and the fraction of non-coding DNA tends to account for almost 6%-24% of their genomes (Elhai et al. 2008). While these non-coding segments were previously discounted as “junk” DNA, their functions and roles in genome regulation and evolution are becoming increasingly recognized. Much of the non-coding DNA comes in the form of repeated sequences, which are classified based on the structure and content of the repeated sequence, as well as the frequency of repeats, their lengths, and characteristics. Some important classes of repeats are tandem repeats, dispersed repeats, and CRISPRs (Avershina, Knut 2015). Dispersed repeats are highly variable in length, so this serves as a useful property to allow for classification of the sequences; generally, four different kinds of repeats have been identified based on length; long sequences repeats, short sequence repeats, very short sequence repeats, and short clustered repeats (Buijn et al, 38-43). All dispersed repeats share the property of being transposable elements that are interspersed throughout the genome (Buijn et al, 38-43).

 Very short tandem repeats includes a number of categories of repeated sequences, including Chi sites, Highly Iterated Palindromes, and DNA uptake signals. DNA uptake signals are 9-12 nt long sequences that occur over 100 times its expected value (or constitute approximately 1% of the genome) (Smith et al. 1999). DNA uptake sequences (DUS) regulate DNA uptake and mediate a process known as transformation, a form of communication between bacterial species (Frye et al, 2013). DNA uptake sequences have primary been characterized in members of the *Neisseriaceae* and *Pasteurellaceae* families (the *Neisseria* and *Haemophilus* genera), and specific patterns for DNA uptake have been characterized in these bacteria (Maughan et al, 2010). Gram-negative bacteria (such as *Haemophilus influenzae* Rd and *Neisseria gonorrhoeae*) tend to uptake only homologous DNA from related species, while Gram-positive bacteria appear to efficiently uptake DNA from any source (Smith et al, 1999).

 Frequency and distribution of DUS have been characterized in the *Haemophilus Influenzae* Rd genome. (Smith et al, 1995). They specify DNA uptake by recognizing the 9-base pair sequence 5’-AAGTGCGGT. 1465 copies of the uptake site have been identified in the genome. An alignment of these sites reveals an extended consensus region of 29 base pairs containing the 9-base pair region and two A/T rich regions (Smith et al, 1995). The DUS are typically located in the base paired stem of the transcription terminators (downstream to ends of genes) (Smith et al, 1999). While the uptake signal sequences have been extensively studied in *H. Influenzae*, their presence, location, frequency have not been investigated in other members of the *Haemophilus* genus. Here, five pathogenic species of *Haemophilus* were chosen to observe for the presence of DNA uptake sequences and to characterize their frequency and sequence. These species are associated with a number of diseases in humans, including meningitis, epiglottitis, bacteremia, and cellulitis (*H. influenzae)*, pneumonia (*H. parainfluenzae)*, genital chancre (*H. ducreyi)*, conjunctivitis (*H. aegyptius*), and endocarditis (*Haemophilus aphrophilus*, now named *Aggregatibacter aphrophilus*) (Musher, Chapter 30).

**Methods**

Bacterial strains were selected based on their ability to cause disease in humans. This was determined by examining a Medical Microbiology text by Daniel Musher, but it is important to note that the remainder of the methodology works (potentially with slight modification) on any bacterial genome. However, the *Haemophilus Influenzae* genome was include due to being highly characterized and serving as a useful validation for the methods being used. The primary tool used to carry out the methods was the BioBIKE application (specifically the PhAnToMe/BioBIKE server), a cloud based bioinformatics program that facilitates computational biology. Code for the commands described will be included as well as an image for the implementation of each command. Additionally, the methodology performed will be described for the *H. Influenzae* genome, but the same functions were used to analyze the genomes of the remaining bacteria for DNA uptake sequences.

 Identification of the DNA uptake sequences was done via the COUNTS-OF-K-MERS function. Because the length of the length of the DUS is 9, a window-size of 9 is chosen, and the THRESHOLD (5) option was selected as well (Figure 1). To count the number of times this DUS occurred in the sequence, the COUNT-OF and MATCHES-OF-PATTERN (“AAGTGCGGT”) functions were used in the genome of *H.* *Influenzae*. The MATCHES-OF-PATTERN option was selected in order to detect similar (but not necessarily exact) copies of the target sequence (and the ONE-STRAND option was also selected).



Figure 1: Quantifying the number K-mers (and hence the most overrepresented sequences and the DUS)



Figure 2: Number of occurences of the DUS (and sequences similar to it) in the H. Influenzae genome.

Both intergenic and gene sequences of *H. Influenzae* were separately analyzed for counts of the DUS, using the commands below:



Figure 3: Number of occurrences of the DUS (and sequences similar to it) in intergenic-sequences-of the H. Influenzae genome.



Figure 4: Number of occurrences of the DUS (and sequences similar to it) in genes of the H. Influenzae genome.

To obtain a visual comparison of the most overrepresented sequences, an alignment was performed. The commands for the alignment are as shown below:



Figure 5: Command for alignment.

**Results and Discussion**

The most common 9 nt sequence in the *H. influenzae* genome was “AAGTGCGGT”, which is the same DUS reported by Smith et al (1995). This serves, to an extent, as a validation of the methods used to determine the DUS. Table 1 contains a list of the DNA uptake sequences for the 5 pathogenic *Haemophilus* species that were observed. Additionally, presented in Table 1 is the count of the occurrences of DNA uptake sequences in the genomes of the organisms, and the count in specific regions of the organism.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Organism** | **DNA uptake sequence (DUS)** | **Count of DUS in genome** | **Count of DUS intergenically** | **Count of DUS within genes** |
| *H. influenzae* | AAGTGCGGT | 737 | 241 | 393 |
| *H. parainfluenzae* | ACCGCACTT **AAGTGCGGT** | 742, 718 |  |  |
| *H. ducreyi* | TTTTTTATT | 130 | 55 | 66 |
| *H. aegyptius*  | ACCGCACTT**AAGTGCGGT** | 516, 512 |  |  |
| *H. aphrophilus* | ACCGCACTT **AAGTGCGGT** | 945, 912 |  |  |

The sequence identified in *H. influenza*e as a DUS was also found in *H. parainfluenzae, H. aegyptius,* and *H. aphrophilus*. However, it was not found in the genome of *H. Ducreyi*. Further analysis showed that the most overrepresented 9 nt sequence (“TTTTTTATT”, see above) only occurred 130 times. This does not meet the criteria for being considered a DUS, as it is not overrepresented enough (for comparison, see the DUS in the other species analyzed). In the other species analyzed, while the DUS identified in *H. Influenzae* was present, it was not the most frequently occurring 9mer. In fact, another sequence (“ACCGCACTT”) had more occurrences in several of the analyzed bacteria genomes (see Table 1). While it may not be apparent, the other frequent sequence is the reverse complement of the DUS, which may account for its high frequency. In order to obtain visualization for the DUS, an alignment was done (see Fig 6).



Figure 6: DUS alignment. Note that the sequence from H. Ducreyi does not align well with the other sequences, indicating that H. Ducreyi does not possess DUS.

* Intend to create a plot for the distance between the DUS sequences for each bacterium and potentially analyze areas with irregular frequencies.

**Works Cited**

<http://www.ncbi.nlm.nih.gov/pubmed/18514361>

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2527708/#B33>

http://www.sciencedirect.com/science/article/pii/S0888754314002821

Bruijn, F. J. De, James R. Lupski, and George M. Weinstock. Bacterial Genomes: Physical Structure and Analysis. New York: Chapman & Hall, 1998. Print.

<http://www.ncbi.nlm.nih.gov/pubmed/10673000>

<http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.10034583>

<http://www.ncbi.nlm.nih.gov/pubmed/20628039>

<http://www.ncbi.nlm.nih.gov/pubmed/7542802>

Musher Daniel. Medical Microbiology, 4th edition. University of Texas Medical Branch at Galveston, 1996. Online. (<http://www.ncbi.nlm.nih.gov/books/NBK8458/>)

http://biobike.csbc.vcu.edu/