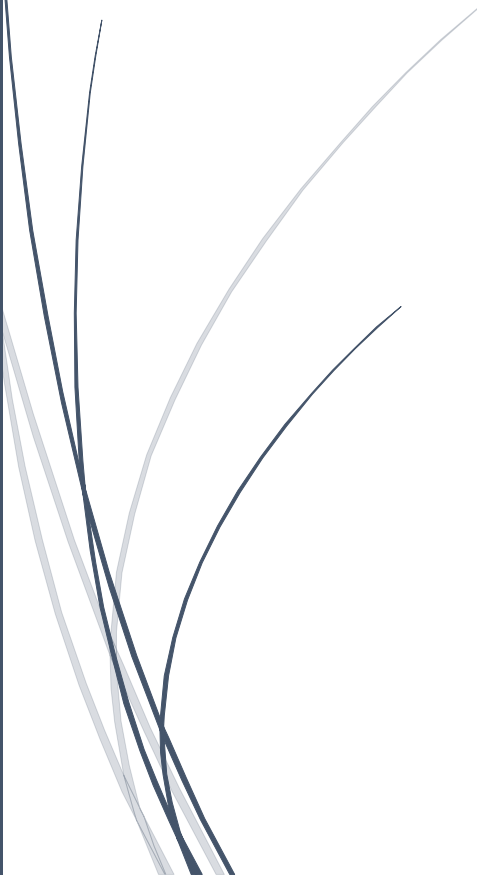




5/10/2014

The occurrences of DNA uptake sequence in *Haemophilus influenzae* and other *Pasteurellaceae* bacteria

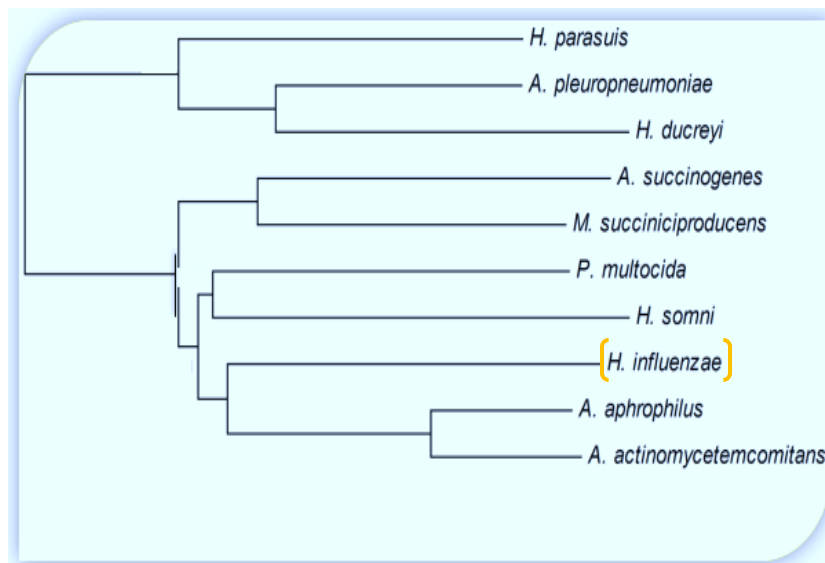


Noha Mudhaffar
JEFF ELHAI - BNFO301

Haemophilus influenzae is a small coccobacillus that includes encapsulated. It is a bacteria that is able to cause a severe infection, occurring mostly in infants and children younger than five years of age. It can cause lifelong disability and be deadly. In spite of its name, Haemophilus influenzae bacteria do not cause influenza (the "flu"). The most common severe form of HIB are: Pneumonia (lung infection), Bacteremia (bloodstream infection), and Meningitis (infection of the covering of the brain and spinal cord). Children can be protected by taken the vaccine. The transmission of this bacteria, including HIB, are through person-to-person by direct contact, or through respiratory droplets like coughing and sneezing [1].

As a bioinformaticians, I used BioBike ⁽¹⁾ to provide the occurrences of DNA uptake sequence in Haemophilus influenzae and other Pasteurellaceae bacteria. Pasteurellaceae (A1) is the family of Haemophilus influenza (A2) that described as Gram-negative bacteria (Taxonomy Browser). Therefore, A mixture of homologous and foreign DNAs in Haemophilus influenzae,

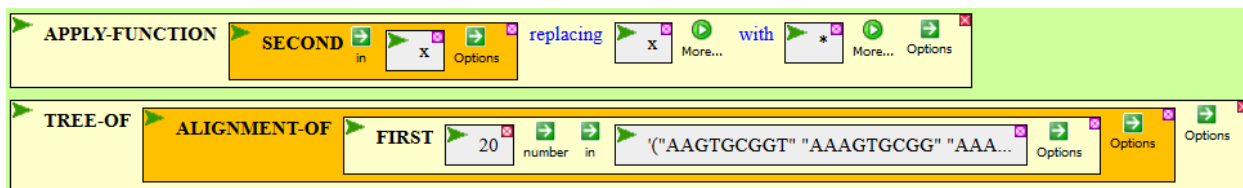
Domain:	Bacteria
Kingdom:	Eubacteria
Phylum:	Proteobacteria
Class:	Gammaproteobacteria
Order:	Pasteurellales
Family:	Pasteurellaceae



show preferential uptake of the homologous DNA [2].

Jeff Elhai¹, Arnaud Taton¹, JP Massar², John K. Myers. BioBIKE: A Web-based, programmable, integrated biological knowledge base. W28–W32 Nucleic Acids Research, 2009, Vol. 37, Web Server issue doi:10.1093/nar/gkp354. Published online 11 May 2009.

Haemophilus-influenzae-86-028NP is one of H-influenzae that has a genome with 1913428 nucleotides long. 62% of the genome is As and Ts, and 38% is Gs and Cs nucleotides. To find the DNA Uptake Signal Sequence (USS) by BioBike, I used COUNTS-OF-K-MERS adding name of bacteria in first box and the number of nucleotides in second box which is 9 (9-nt sites important for certain bacteria to recognize exogenous DNA as self) [4], and choosing BOTH-STRANDS option. I got a long list of common short sequences. The highly repeated sequence is AAGTGCGGT which occurs 1516 times. I wanted to make a graph that shows all common sequences. So, I took all the resulted list from COUNTS-OF-K-MERS and used it in APPLY-FUNCTION to make a loop that gives only the sequences without the coordinates (*



means the previous result). Then I

used

TREE-OF to get as result figure

(B1).

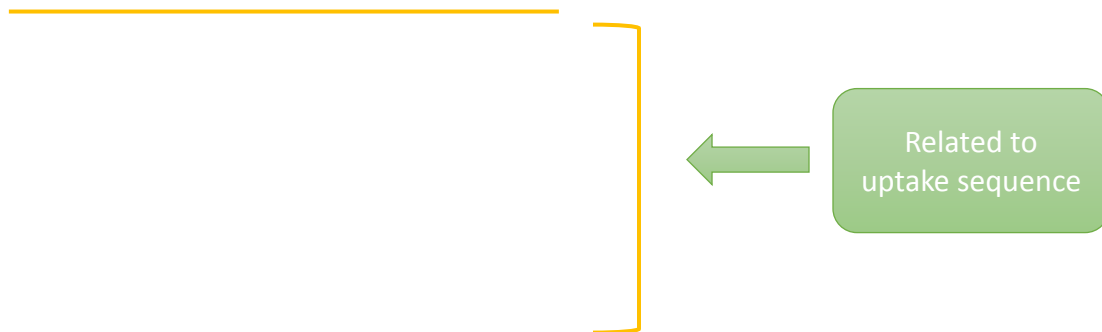
```

Seq 8:AAATA      1 --AAATAAAA A---
Seq 15:AATAA     1 ---AATAAAA AT--
Seq 12:AATAA     1 ---AATAAAA AA--
Seq 20:ATAAA     1 ----ATAAAA AAT-
Seq 17:TAAAA     1 --TAAAAATA A---
Seq 18:AAAAA     1 --AAAAAATA A---
Seq 10:AAAAA     1 ---AAAAATA AA--
Seq 9:AAAAT      1 ----AAAATA AAA-
Seq 14:AAAAG     1 --AAAAGAAA A---
Seq 16:AAAGA     1 ---AAAGAAA AA--
Seq 7:AGTGC      1 ----AGTGCG GTA-
Seq 19:AGTGC     1 ----AGTGCG GTG-
Seq 5:AGTGC      1 ----AGTGCG GTC-
Seq 6:GTGCG      1 -----GTGCG GTCA
Seq 13:GTGCG     1 -----GTGCG GTTA
Seq 1:AAGTG       1 ---AAGTGCG GT--
Seq 2:AAAGT      1 --AAAGTGCG G---
Seq 3:AAAAG      1 -AAAAGTGCG ----
Seq 11:AAAAA     1 AAAAAGTGC- ----
Seq 4:AACCG      1 ---AACCGCA CT--
consensus       1

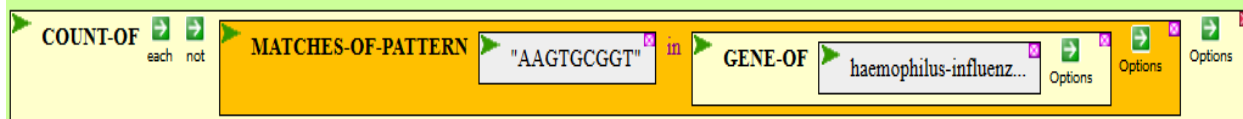
```

AT-rich
sequences

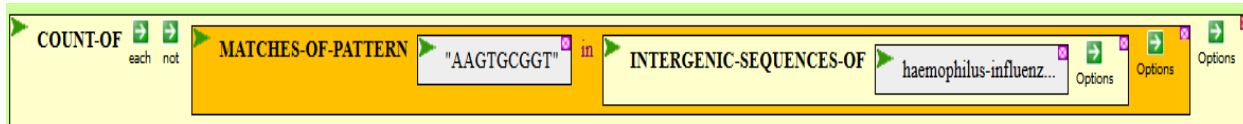




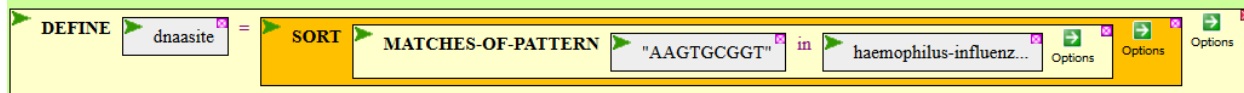
The tree i Figure B1 sequences. The bottom sequences are related to DNA USSs, and the up sequences are occurred because of the rich of As and Ts in the genome. Then, to know there positions, I used two function as shown below to find the number of repeats inside and outside the genes. The first COUNT-OF function showed that there are 1001 DNA USS inside the genes, and the second COUNT-OF function provided the number of



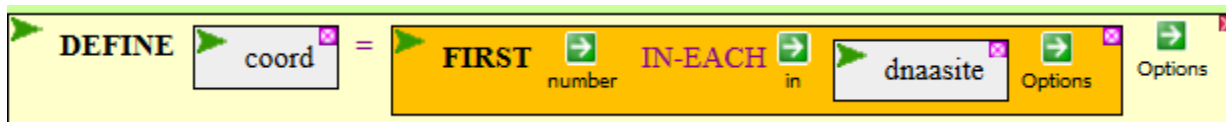
DNA USS between genes which is 479. There are 66 overlap sequences.



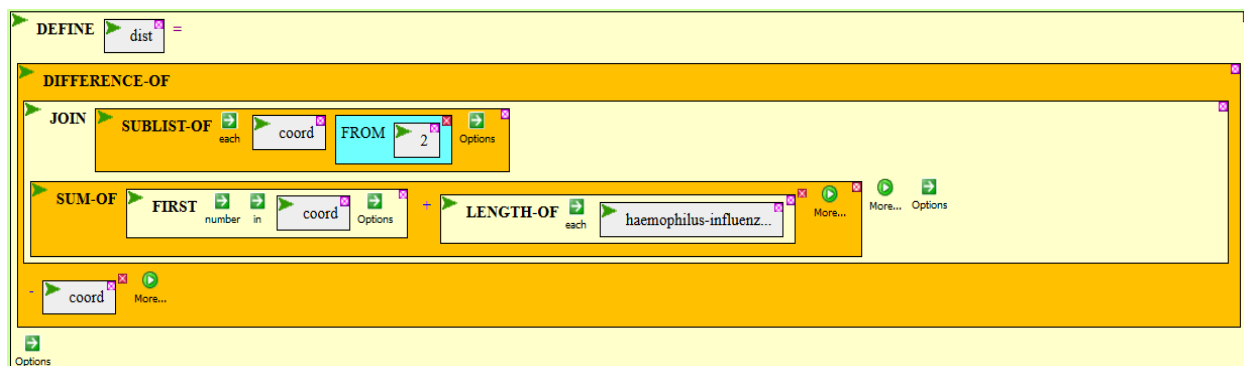
But, what is the distance between the sequences? To know that, I used MATCHES-OF-PATTERN to get the coordinates of the USS sequences. Then sort the result to get more conveniently results. Then defined it as DNA-a-site.



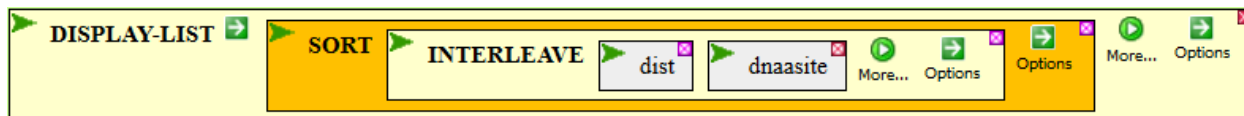
After that, I needed to define a variable called coordinates as a first element from the result I obtained in previous function. Notice the icon IN-EACH. It is important to get the first elements from each line in the list.



Also we need to define a variable says distance that provide the distance between each coordinates. Here is how my function looks like.



Finally I used function INTERLEAVE to mix the distance list with the DNA-a-site list, and sort it, then put the obtained result in DISPLAY-LIST.



I obtained a very long list of distances between sequences. However, it usually is better to use PLOT function to get a nice graph shows the distance. The PIN-INTERVAL icon defined the

scale and the MAX icon defined the maximum coordinates I want to graph it (C1) I chose 1,500,000 because I did not see anything interesting after this coordinates.

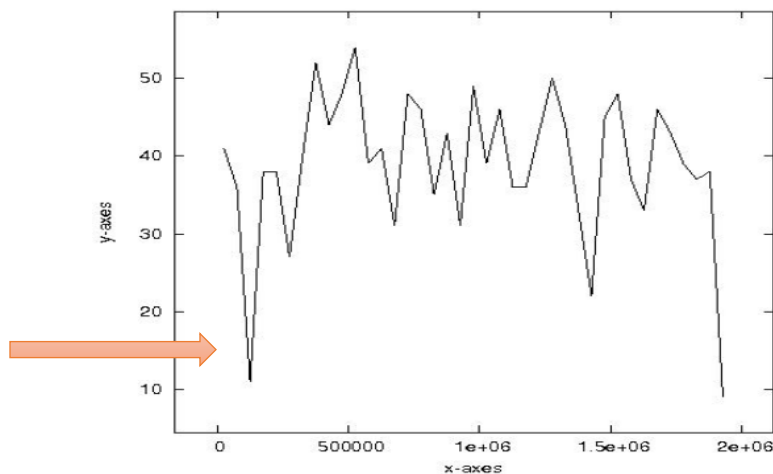
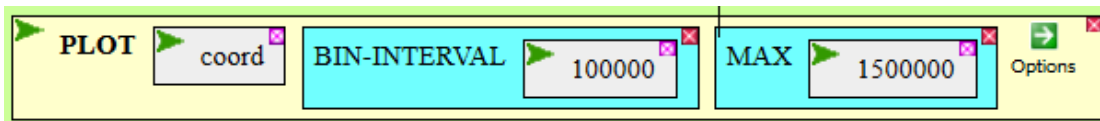
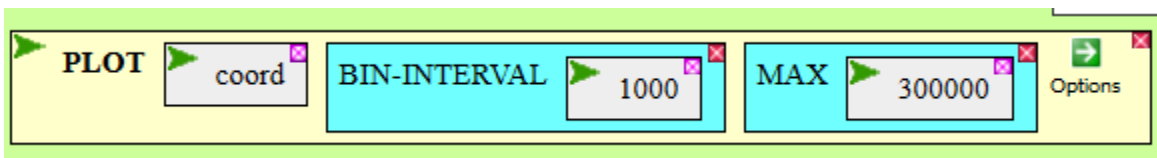


Figure (C1): The row Indicates to the area that has USS very rare

In this graph (C1), the X-axis represents the coordinates, and the Y-axis represents the number of repeats. We can see in coordinates that less than 500,000 nt, there is DNA USS occurs very rare, almost zero, and that something unexpected. So, I zoomed in more to this area by changing the scale to 1000 which I thought is much better, and changing the max to 300,000 nt (C2).



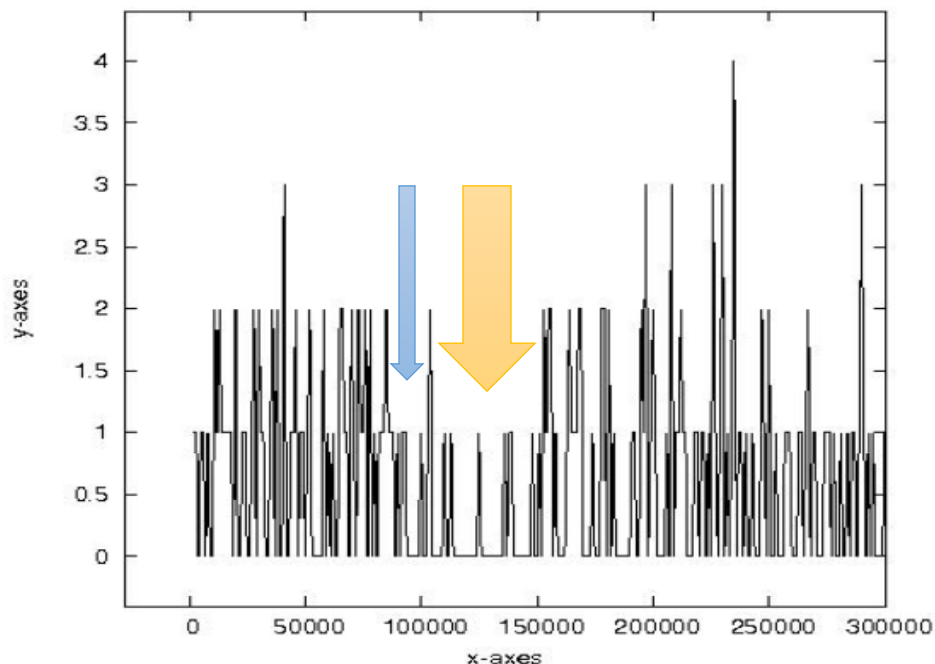


Figure (C2): this graph shows two positions that have rare Occurrence of DNA USSs. First position (blue row) has a plasmid. Second position (yellow row) is bigger than first one and has only 5 DAN USSs

In this graph, we can see two positions that have a rare repeated sequences. In the first small area with a blue row, I found that in a gene called “Hinf-86-028NP.NTHI0101” with coordinates (92334 -> 93170) and description (Chromosome (plasmid) partitioning protein ParA / Sporulation initiation inhibitor protein Soj), there is a plasmid that integrates into the genome. In this area, from 90,000 to 100,000 (10,000 nt), there are only two repeated sequences. The yellow row indicates to the area from 100,000 to 150,000. In this 50,000 nt, there are only 5 repeats of DNA USSs occur, which is less than plasmid area if we estimate that there is only one DNA USS in very 10,000 nt. I think it is called genomic island. “A gene in a genome is defined as putative alien (pA) if its codon usage difference from the average gene exceeds a high threshold and codon usage differences from ribosomal protein genes, chaperone genes and protein-synthesis-processing factors are also high. pA gene clusters in bacterial genomes are relevant for detecting genomic islands (GIs), including pathogenicity islands (PAIs)” [Samuel K, 2001] [5].

Comparing Haemophilus-influenzae-86-028NP with other Pasteurellaceae bacteria and using same functions, I got table (D1) that shows the name of all Pasteurellaceae bacteria, the length, the GC percentage, and the number of DNA USS occurrences.

Organism	Length	GC-FRACTION	Occurrences of DNA USS
Actinobacillus-succinogenes-130Z	2319663	0.44918594	1690
Haemophilus-influenzae-86-028NP	1913428	0.3815231	1516
Actinobacillus-actinomycetemcomitans-HK1651	1995520	0.44412282	1507
Mannheimia-succiniciproducens-MBEL55E	2314078	0.42537978	1485
Haemophilus-influenzae-R2846	1824242	0.37971717	1461
Haemophilus-somnus-2336	2263857	0.37378067	1355
Haemophilus-somnus-129PT	2012878	0.37191722	1245
Haemophilus-influenzae-R2866	1933340	0.38079283	952
Pasteurella-multocida-subsp-multocida-str-Pm70	2257487	0.40404883	927
Haemophilus-influenzae-86028NP	1738864	0.38510486	888
Haemophilus-influenzae-Rd-KW20	1830138	0.38147888	737
Actinobacillus-pleuropneumoniae-L20	2274482	0.41299513	73
Actinobacillus-pleuropneumoniae-serovar-1-str-4074	2292348	0.41376877	63
Mannheimia-haemolytica	2498406	0.40754706	59
Haemophilus-ducreyi-35000HP	1698955	0.38220495	41

Table (D1): comparison between Pasteurellaceae bacteria

The DNA USS occur differently in Pasteurellaceae bacteria. In some bacteria the uptake signal sequence is highly repeats, and in other very rare. Therefore I liked to do some experiments on Haemophilus-influenzae-Rd-KW20. Because of this bacteria known as uncompleted genome sequences, I used MATCH-OFPATTREN to find this nucleotides positions that are not AGCT.

I obtained different letters like "mnsykrw", and I changed it with Gs by using TRANSELITRATE which will give me a new long sequence.

I will use this new sequence to find the occurrence of DNA USS by using MATCH-OF-PATTERN again to find the USS AAGTGCGGT.

Then, count the result to see how many times this sequence occurs.

I got a totally different result from what is in the table (D1). The total number of DNA USS is 1417 repeats. To find the DNA-a-site, coordinates, and the distance, I will use same functions as before.

The screenshot shows a 'DEFINE' window with a variable 'dist' set to a 'DIFFERENCE-OF' operation. The 'DIFFERENCE-OF' operation is composed of a 'JOIN' step and a 'SUM-OF' step. The 'JOIN' step uses 'SUBLIST-OF' on 'coord' from index 2. The 'SUM-OF' step calculates the difference between the 'FIRST' element of 'coord' and the 'LENGTH-OF' of 'haemophilus-influenz...'. There are 'More...' and 'Options' buttons for each sub-step.

The screenshot shows a 'DISPLAY-LIST' window with a 'SORT' operation. The 'SORT' operation is an 'INTERLEAVE' of 'dist' and 'dnaasite'. There are 'More...' and 'Options' buttons for the 'INTERLEAVE' step.

After that, I decided to make a graph to see where it occurs more and where less occurs.

The screenshot shows a 'PLOT' window with 'coord' as the data source. The 'BIN-INTERVAL' is set to 100000 and the 'MAX' value is set to 180000. There is an 'Options' button.

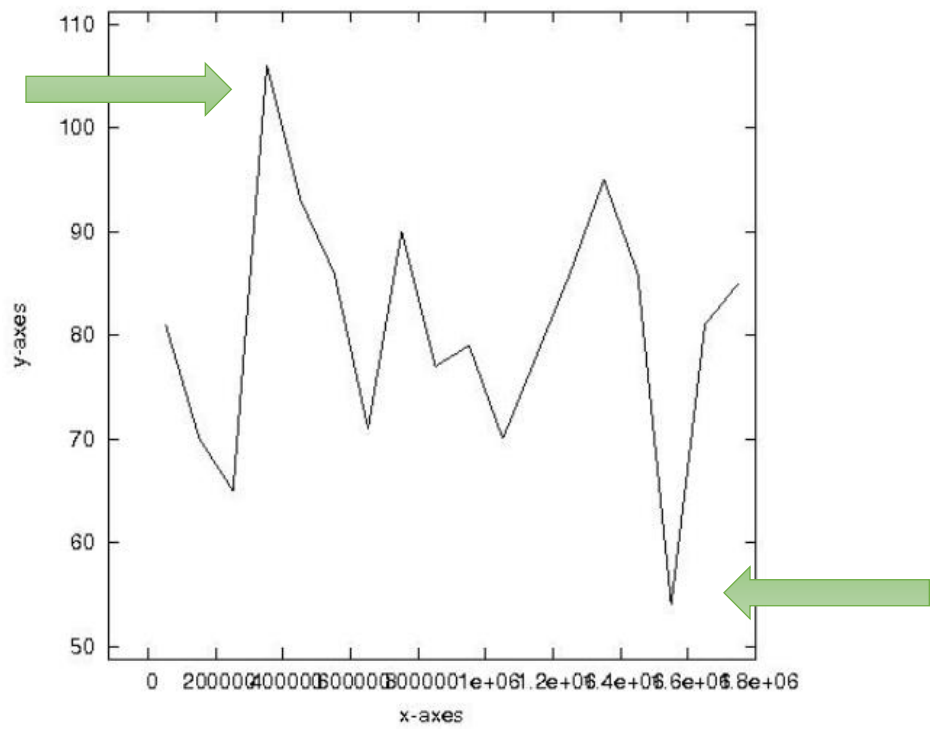


Figure (E1): one high region and one low region

Because of these two regions in fig (E1), I made zoom in for these two areas by changing the scale and the max number.

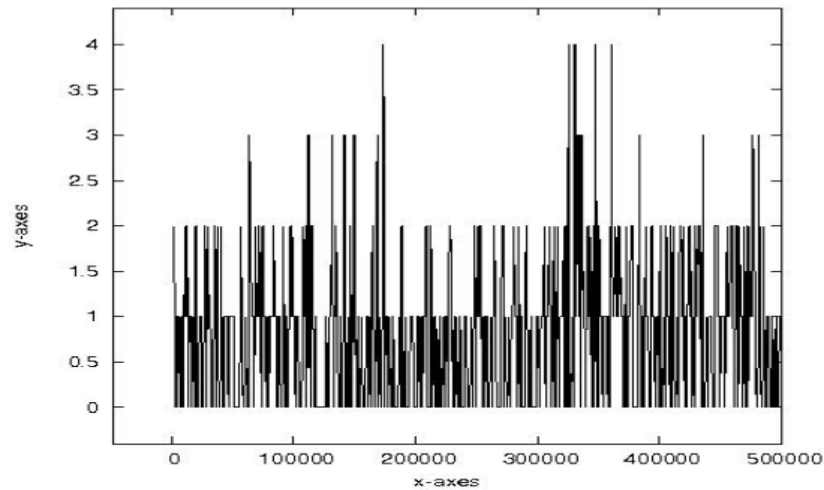
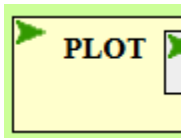


Figure (E2): the repeats of UUS from 1 to 500,000 nt

Fig (E2) shows that there is at least one DNA USS. So I will look to another region that is from 1,500,000 to the end of the genome which is approximately 1,800,000.

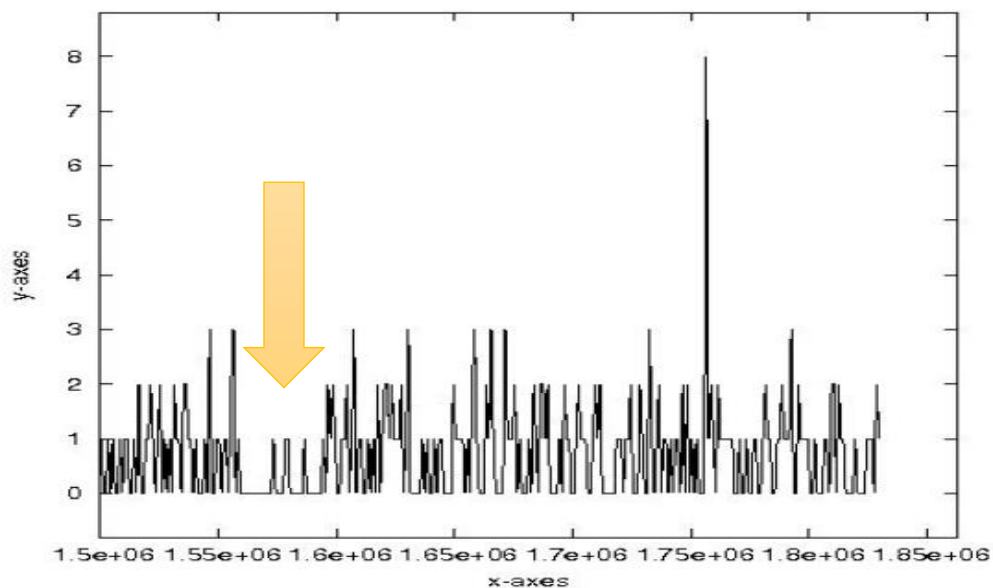
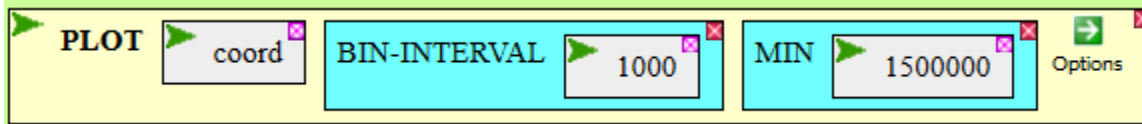


Figure (E3): the repeats of UUS from 1,500,000 to the end of the genome

This yellow row in fig (E3) indicates the area between 1,560,000 until 1,590,000. So I will zoom in more to this region and see what's especial it has.

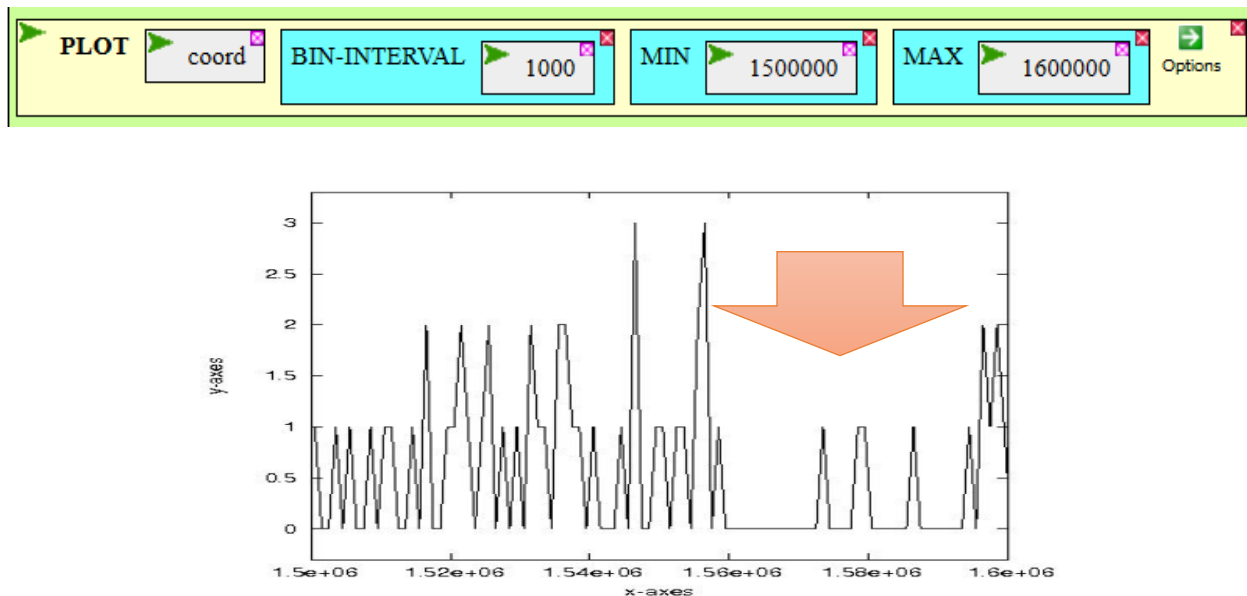


Figure (E4): the orange big row indicates the rare repeats of UUS

This big row in fig (E4) indicates the area that has a very rare DNA USS. By using COUNT-OF I knew that there are only 3 repeated sequences and that's make me look on the genes in this area, but most of the genes have unknown function.

From all the working in BioBike, plying with function, experiments ideas, and getting unexpected result, I learned that bioinformatics world is very big and will never end. As much as

you work will find interesting things that waiting for someone to find them. I would like to work more and more, and have more experiments and results.

References

1. Murphy TF1, Faden H, Bakaletz LO, Kyd JM, Forsgren A, Campos J, Virji M, Pelton SI. Nontypeable *Haemophilus influenzae* as a pathogen in children. *Pediatr Infect Dis J*. 2009 Jan;28(1):43-8. doi: 10.1097/INF.0b013e318184dba2.
<http://www.ncbi.nlm.nih.gov/pubmed/19057458>
2. Hamilton O. Smith*, Michelle L. Gwinn, Steven L. Salzberg. DNA uptake signal sequences in naturally transformable bacteria. *The Institute for Genomic Research*, 9712

Medical Center Drive, Rockville, MD 20850, USA. Res. Microbiol. 150 (1999) 603–616
© 1999 Éditions scientifiques et médicales Elsevier SAS. All rights reserved.

3. Zhuofei Xu¹., Min Yue¹., Rui Zhou¹, Qi Jin², Yang Fan², Weicheng Bei^{1*}, Huanchun Chen^{1*}. Genomic Characterization of *Haemophilus parasuis* SH0165, a Highly Virulent Strain of Serovar 5 Prevalent in China. May 2011 | Volume 6 | Issue 5 | e19631.
www.plosone.org
4. Jeffery A, Introduction to Bioinformatics. Research Group - Very Short Dispersed Repeats. 2014.
5. Samuel K. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. TRENDS in Microbiology Vol.9 No.7 July 2001. 0966-842X/01/\$ – see front matter © 2001 Elsevier Science Ltd. All rights reserved. PII: S0966-842X(01)02079-0. <http://tim.trends.com>