Viral Metagenome Analysis

Rick Pascual
Introduction of Bioinformatics
Final Report

Virginia Commonwealth University
Department of Life Sciences

Submitted May 8, 2009

INTRODUCTION

Viruses are the most abundant organisms on the planet (Edwards and Rohwer 2005). Although, when it comes to viral metagenomes and their biological properties, there is still a good amount we do not know about (Schoenfeld *et al.* 2008). The Schoenfeld *et al* study isolated metagenome profiles from two hot springs from Yellowstone National Park, Bear and Octopus. The data from the study were collected by the Department of Life Sciences at Virginia Commonwealth University in the form of sheared DNA sequence reads. Taking pointers from the Schoenfeld *et al*., the students of the Introduction to Bioinformatics (BNFO 301) course for the Spring 2009 semester attempted to do their own analysis of the reads to see if they could find anything interesting for themselves about these viral metagenome reads.

A program called BioBIKE, sponsored by the National Science Foundation, was used for most of the metagenome analysis of these reads along with tools from the National Center of Biotechnology Information (NCBI). The main goal of this particular viral metagenome analysis is to look into the reads obtained from the Schoenfeld *et al*., construct longer sequence contigs out of these reads and find anything about them that would be deemed of significance in the field of molecular biology.

CONTRUCTION AND ANALYSIS OF CONTIGS

For this particular study, two DNA reads from the Bear Paw Hot Springs were used: BPHS.AOIX1739-b2 and BPHSAOIX1739-g2. For the initial steps in contig (set of over-lapping DNA sequences) construction, the two reads were compared to all the reads within the Bear Paw read database. This is where BioBIKE functions were first utilized; the program allows for the simplification of DNA analysis by the computerization of redundant operations usually needed for such analyses.

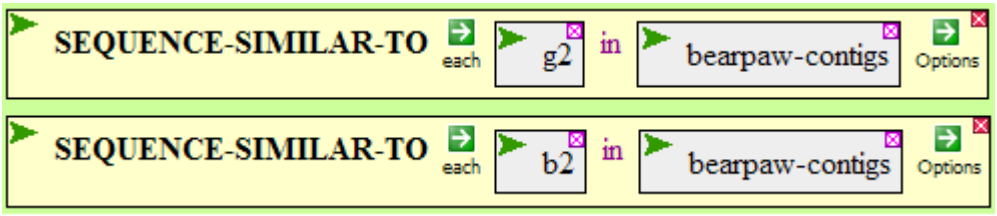

Figure 1 – SEQUENCE-SIMILAR-TO function within BioBIKE that compares a given DNA read to a list of DNA reads. This tries to find certain amounts of nucleotide matches within the read and returns what it finds in the form of the name of the matched sequence, their respective nucleotide coordinate location on the read, as well as values pertaining to its sequence identity (how much the respective reads match up with each other) in the form of a percentage.

For the purposes of convenience, the BPHS.AOIX1739-b2 and BPHSAOIX1739-g2 reads were defined as “b2” and “g2” within BioBIKE. Similarly the list of Bear Paw metagenome reads were defined as “bear-paw.” Using the sequence illustrated in Figure 1, the two reads were compared to the entire list of Bear Paw metagenome reads for any type of similarity. The results of this showed an unusually high level of similarity between the beginning parts (first ~100 nucleotides starting from the 5’-3’ direction) of the two reads along with almost every other read within the Bear Paw metagenome

database. Table 1 and 2 in the Tables and Figures section shows the results. To look into this further, another BioBIKE function was used that was capable to looking at a list of reads it was given and display nucleotide-per-nucleotide the matches within each DNA sequence it found. Figure 2 in Tables and Figures shows a small portion of the alignment found. The highlighted portion shows one of the areas that showed a high amount of repetition and matching-up between the different reads.

It was later found out that Schoenfeld *et al.* (2008) had used linker sequences when acquiring the reads and that one particular linker (mentioned in the study to have been ligated to the metagenome profiles when amplified) containing the sequence 5'-GGAGCAGTATCAGATACAAGCGGCCGCATC-3' when analyzed alongside the reads matched up almost completely in certain segments of it. Further analysis showed that segment 5'-GCCGCATC-3' (the last seven nucleotides of the mentioned linker) matched up almost completely It was later found out by doing a count of this segment among all the reads (used a function in BioBIKE that sped up the process) that it matched up with 6170 of the 8352 reads.

Using this information, a function within BioBIKE was drawn up that, simply put, first looked for this segment in each of the reads. If it found it then it would take the read's respective coordinates and use that to determine where to cut out the "garbage" sequence (the linker segment along with everything before it). Otherwise, it wouldn't consider it. Such a process was deemed relatively unreliable, as it did not discriminate between sequences being read from the 3'-5' and that of 5'-3' among the reads as well as throwing out a good amount of reads (~30%). It was around this time that Jeff Elhai from the Department of Life Sciences at VCU gave out the edited sequences of the viral

metagenome. From that point on, the construction proceeded using the edited reads as substitute. It was assumed and proven by doing an alignment of the edited reads to their non-edited counterparts that the garbage linkers at the beginning were taken out.

The edited sequences were re-defined as “b2-e” and “g2-e” while the Bear Paw contigs as “contigs-e.” Doing a SEQUENCE-SIMILAR-OF highlighted an important aspect of the reads. There was good amount of matching up between the first parts of my particular read and the last parts of other sequences within the Bear Paw database. The boxed out data from Figure 3 show this.

	QUERY	Q-START	Q-END	TARGET	T-START	T-END	E-VALUE	%ID
1.	BPHSe.AOIX1739-b2	1	894	BPHSe.AOIX1739-b2	1	894	0.0	100.0
2.	BPHSe.AOIX1739-b2	38	649	BPHSe.AOIX4283-g2	181	792	9.0d-94	82.03
3.	BPHSe.AOIX1739-b2	38	282	BPHSe.AOIX2071-g2	627	872	8.0d-45	84.55
4.	BPHSe.AOIX1739-b2	17	273	BPHSe.AOIX3916-b2	723	467	8.0d-42	83.27
5.	BPHSe.AOIX1739-b2	495	652	BPHSe.AOIX3916-b2	245	88	3.0d-35	86.71
6.	BPHSe.AOIX1739-b2	496	652	BPHSe.AOIX650-b4	476	632	6.0d-18	82.17
7.	BPHSe.AOIX1739-b2	80	276	BPHSe.AOIX650-b4	60	256	1.0d-15	80.2
8.	BPHSe.AOIX1739-b2	441	649	BPHSe.AOIX2334-b2	282	490	1.0d-15	79.9
9.	BPHSe.AOIX1739-b2	38	155	BPHSe.AOIX4302-b2	652	769	2.0d-14	83.05
10.	BPHSe.AOIX1739-b2	240	276	BPHSe.AOIX650-b2	277	313	2.0d-8	94.59
11.	BPHSe.AOIX1739-b2	606	705	BPHSe.AOIX2510-g2	413	512	8.0d-8	82.0
12.	BPHSe.AOIX1739-b2	804	886	BPHSe.AOIX2510-g2	611	692	3.0d-7	84.34
13.	BPHSe.AOIX1739-b2	801	884	BPHSe.AOIX2544-b2	107	189	2.0d-5	83.33
14.	BPHSe.AOIX1739-b2	801	884	BPHSe.AOIX2544-g2	707	625	2.0d-5	83.33

Figure 3 – Sequence similarities around similar coordinates in the BPHSe.AOIX1739-b2 read compared to the ending parts of other Bear Paw reads.

The data was tabulated and drawn out using the relative coordinates between the reads. As shown in Figure 4, a noticeable congregation of overlapping sequences, by this I mean reads wherein there is a high amount of nucleotides within the area appeared to match up with each other among the different reads. Figure 5 shows a closer version of this overlap along with the respective coordinates of each read where an over-all alignment was obtained. It should be noted that this alignment of DNA sequences was made even between areas wherein linkage was not reported between reads by BioBIKE.

This resulted in a sequence alignment that showed a high amount of nucleotide similarity even in those areas. Most of the time, the reason why there wasn't an over-all consensus (all reads having the same nucleotide at a specific sequence coordinate) was because only one out of the four had a different nucleotide in it. Using this information, an alternative sequence for that area was manually constructed using the dominant nucleotides between each of the reads. For instance, if three of the four sequences had an Guanine (G) in a specific non-consensus coordinate with only one other sequence having Adenine (A), the alternative sequence would use G to count for all four sequences. If two had one base while the other two had another base, one of the two would be used. Figure 6 better illustrates this.

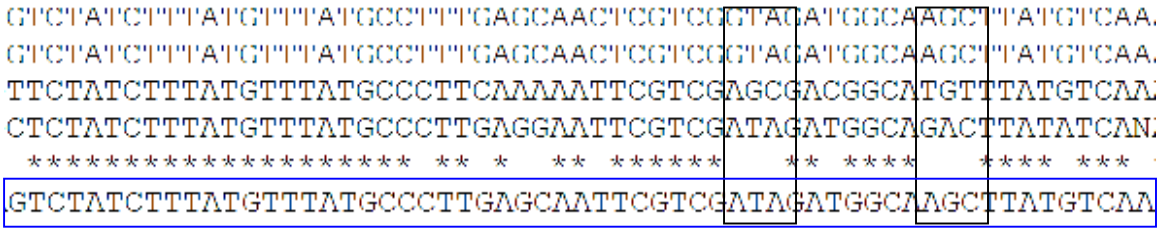

Figure 6 – Short segment of the four aligned sequences with portions boxed out showing how the alternative sequence (lowest sequence) was made.

The resulting alternative sequence of 222 nucleotides was then used to join my BPHSe.AOIX1739-b2 sequence along with the sequence that had the closest identity to it, BPHSe.AOIX4283-g2. The resulting sequence was then transferred on to NCBI to check if there were any matching genes within the database based on its open reading frames (ORF, portions of DNA sequence with a start codon and a stop codon). The results found that one of the ORFs closely resembled the protein structure of gene AprM in *Streptomyces tenebrarius*.

FURTHER DISCUSSION AND FUTURE DIRECTIONS

Such data can be very helpful in further analyzing the viral metagenome of the Hot Springs. Due to time limitations in the project, many of its initial goals were not completed as had hoped for. However, the analysis talked about in this paper show many possible paths that could be taken using the information and insights gained from it. One of these is the further expansion of the contigs using sequence similarities between the reads. Another possible option for further analysis can be the incorporation of both Bear Paw and Octopus hot springs when considering sequence similarities. Lastly, taking from the information about a high amount of nucleotide alignment seen in the matched up reads, a phylogenetic tree can ideally be constructed that uses these variations found in these contigs to find common ancestries between them. The vast amount of future analysis that can be done for these reads as well as the types of information that can be eventually obtained from such analysis can prove to be a very exciting prospect for an interested scientist.

LITERATURE CITED

Schoenfeld, T., M. Patterson, P. Richardson, K.E. Wommack, M. Young, D. Mead. 2008. Assembly of Viral Metagenomes from Yellowstone Hot Springs. Applied and Environmental Microbiology 74: 4164-4174.

Edwards, R and F. Rohwer, 2005. Viral Metagenomics. Nature Reviews Microbiology 3: 504-510.

TABLES AND FIGURES

	QUERY	Q-START	Q-END	TARGET	T-START	T-END	E-VALUE	%ID
1.	Seq1	1	952	BPHS.AOIX1739-b2	13	964	0.0	100.0
2.	Seq1	96	707	BPHS.AOIX4283-g2	240	851	1.0d-93	82.03
3.	Seq1	96	340	BPHS.AOIX2071-g2	696	941	1.0d-44	84.55
4.	Seq1	75	331	BPHS.AOIX3916-b2	798	542	1.0d-41	83.27
5.	Seq1	553	710	BPHS.AOIX3916-b2	320	163	4.0d-35	86.71
6.	Seq1	3	59	BPHS.AOIX2659-b2	10	66	5.0d-25	100.0
7.	Seq1	6	60	BPHS.AOIX3216-b2	16	70	8.0d-24	100.0
8.	Seq1	7	61	BPHS.AOIX3138-b2	21	75	8.0d-24	100.0
9.	Seq1	6	59	BPHS.AOIX940-b4	7	60	3.0d-23	100.0
10.	Seq1	6	59	BPHS.AOIX1741-b2	18	71	3.0d-23	100.0
11.	Seq1	6	59	BPHS.AOIX1137-b4	10	63	3.0d-23	100.0
12.	Seq1	6	58	BPHS.AOIX949-b4	10	62	1.0d-22	100.0
13.	Seq1	6	58	BPHS.AOIX617-b4	9	61	1.0d-22	100.0
14.	Seq1	6	58	BPHS.AOIX537-b4	10	62	1.0d-22	100.0
15.	Seq1	6	58	BPHS.AOIX4351-b2	20	72	1.0d-22	100.0
16.	Seq1	7	59	BPHS.AOIX4335-b2	21	73	1.0d-22	100.0
17.	Seq1	1	61	BPHS.AOIX3726-b2	11	70	1.0d-22	98.36
18.	Seq1	6	58	BPHS.AOIX3676-b2	11	63	1.0d-22	100.0
19.	Seq1	6	58	BPHS.AOIX2270-b2	15	67	1.0d-22	100.0
20.	Seq1	6	58	BPHS.AOIX1856-b2	20	72	1.0d-22	100.0

Table 1 – First 20 SEQUENCE-SIMILAR-OF results between BPHS.AOIX1739-b2 (Seq1 under QUERY) and the Bear Paw sequences (given with their respective names under TARGET).

	QUERY	Q-START	Q-END	TARGET	T-START	T-END	E-VALUE	%ID
1.	Seq1	1	49	BPHS.AOIX3543-g2	12	62	7.0d-15	96.08
2.	Seq1	1	40	BPHS.AOIX3506-g2	5	44	7.0d-15	100.0
3.	Seq1	1	49	BPHS.AOIX3033-g2	11	61	7.0d-15	96.08
4.	Seq1	1	48	BPHS.AOIX2855-g2	2	48	7.0d-15	97.92
5.	Seq1	1	40	BPHS.AOIX2374-g2	7	46	7.0d-15	100.0
6.	Seq1	1	40	BPHS.AOIX2227-g2	8	47	7.0d-15	100.0
7.	Seq1	1	49	BPHS.AOIX2101-g2	7	57	7.0d-15	96.08
8.	Seq1	2	50	BPHS.AOIX2049-g2	9	59	7.0d-15	96.08
9.	Seq1	1	49	BPHS.AOIX1647-g2	6	56	7.0d-15	96.08
10.	Seq1	1	49	BPHS.AOIX1032-g2	5	55	7.0d-15	96.08
11.	Seq1	1	49	BPHS.AOIX1003-g2	6	56	7.0d-15	96.08
12.	Seq1	1	39	BPHS.AOIX1962-g2	6	44	3.0d-14	100.0
13.	Seq1	1	39	BPHS.AOIX950-g2	5	43	3.0d-14	100.0
14.	Seq1	1	39	BPHS.AOIX922-g2	5	43	3.0d-14	100.0
15.	Seq1	1	39	BPHS.AOIX740-g2	6	44	3.0d-14	100.0
16.	Seq1	1	39	BPHS.AOIX639-g2	6	44	3.0d-14	100.0
17.	Seq1	1	50	BPHS.AOIX4122-g2	7	57	3.0d-14	96.08
18.	Seq1	1	39	BPHS.AOIX4118-g2	6	44	3.0d-14	100.0
19.	Seq1	2	40	BPHS.AOIX4093-g2	2	40	3.0d-14	100.0
20.	Seq1	1	39	BPHS.AOIX4059-g2	7	45	3.0d-14	100.0

Table 2 – First 20 SEQUENCE-SIMILAR-OF results between BPHS.AOIX1739-g2 (Seq1 under QUERY) and the Bear Paw sequences (given with their respective names under TARGET).

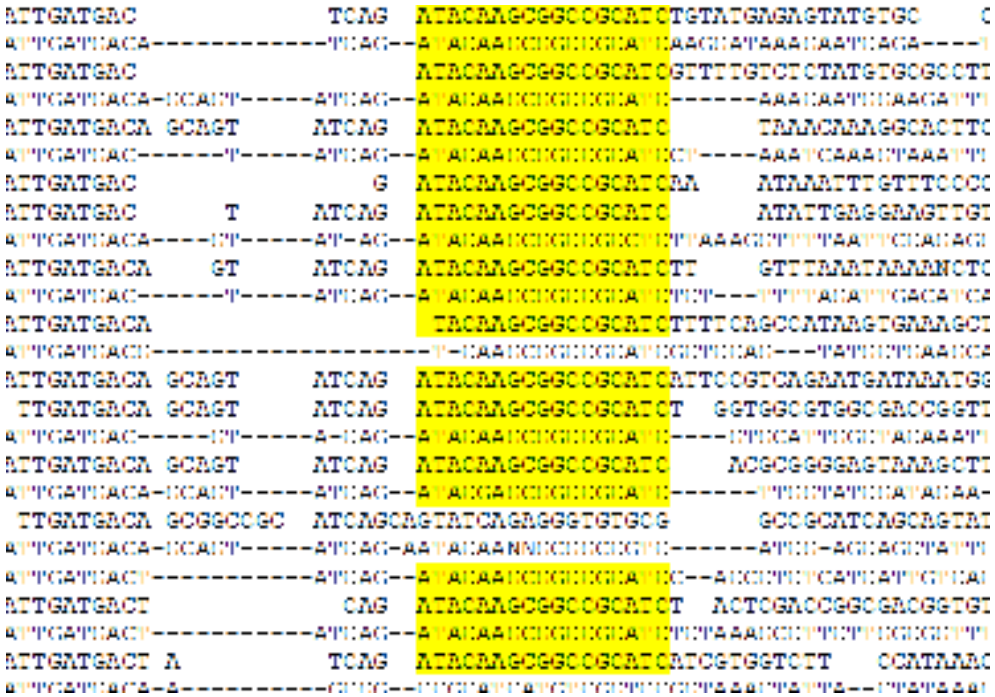

Figure 2 – Alignment of different reads within the Bear Paw database.

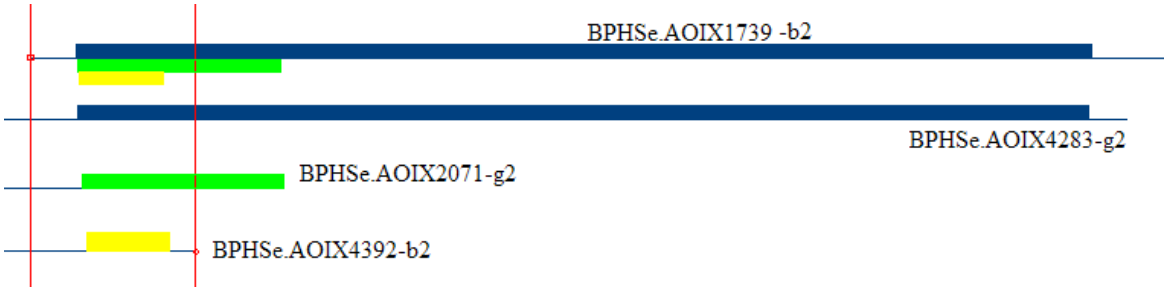

Figure 3 – First identified sequence linkages between four chosen reads including BPHSe.AOIX1739-b2.

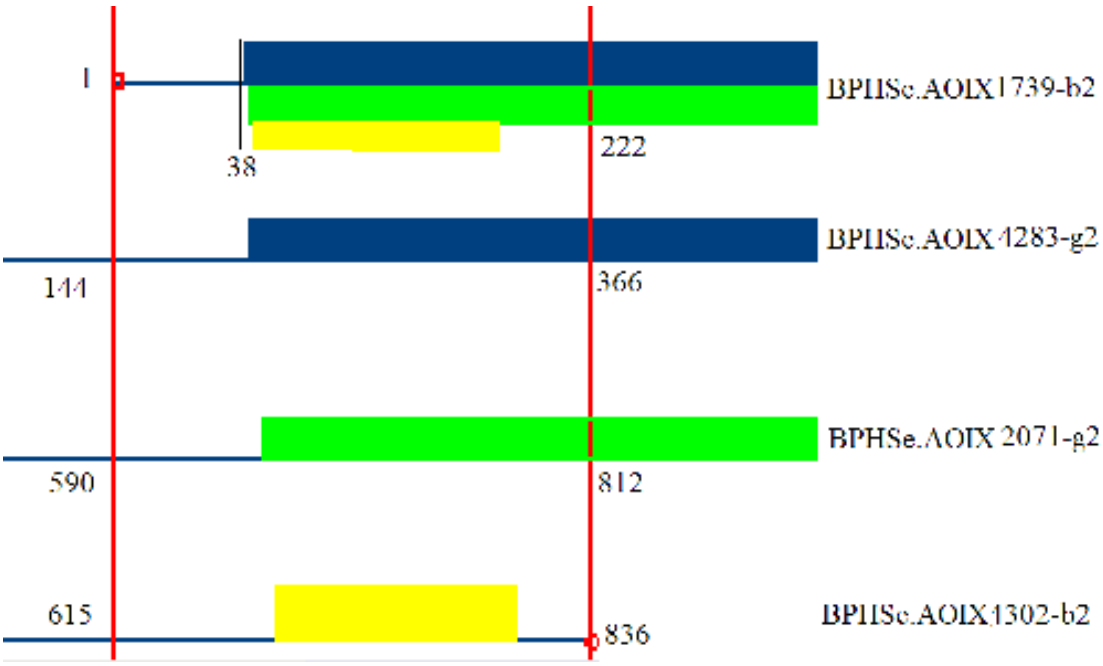

Figure 5 – Respective coordinates from each of the four reads where an over-all general alignment was taken and observed.