It was my friend’s idea to visit Yellowstone National park hot springs. Since I was desparet to explor something new, I was excited and decided to go. Yellowstone hot spring is a place that I have never emagine it would excist its spectacular unique termal futures and landforms are magnificient. How can an organism could live in this kind of termophilic extreme environmental condition? That was the question I just kept asking myself trough out my trip. Now i got a chance to work on a metagenomes from yellowstone hot spring as my bioinformatics final project. What can you do with 896 nuclotoides? I was to determine what is interesting about this 896 nuclotoides from octopus hot spring(my read) Organism-[phage\_octopus\_hot\_springs](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA67160&NAME=-phage_octopus_hot_springs). With the help of biological tool aided with computer(BioBike) I started my jurney to discovery.

Since Genes contain information that determin specificaly about the cell. My first stap was to find a gene in my sequence. The probablity of finding a gene in a read that is only 896 nuclotoides was very unlikely, So I had to make my read biger atlist a caple fo tauthend nuqulotoid long. I find reads that are similar to my read from the metagenome cotopuse inorder to find overlaps; identical highly similar regines between two reads.

([SEQUENCE-SIMILAR-TO](http://biobike.csbc.vcu.edu:9003/help/function?PKG=TIHITINA95546&PACKAGE=BBI&SYMBOL=SEQUENCE-SIMILAR-TO) OCTHS.APNO1593-B2 IN OCTOPUS)

From the list of similar reads result I evaluated each reads on to how they have overlaps with my read.

**QUERY Q-START Q-END TARGET T-START T-END E-VALUE %ID**

**1.** [**OctHS.APNO1593-b2**](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA95546&NAME=OctHS.APNO1593-b2) **1 956** [**OctHS.APNO1593-b2**](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA95546&NAME=OctHS.APNO1593-b2) **1 956 0.0 100.0**

**2.** [**OctHS.APNO1593-b2**](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA95546&NAME=OctHS.APNO1593-b2) **61 926** [**OctHS.ATYB3741-b4**](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA95546&NAME=OctHS.ATYB3741-b4) **931 66 0.0 93.76**

**3.** [**OctHS.APNO1593-b2**](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA95546&NAME=OctHS.APNO1593-b2) **61 716** [**OctHS.APNO3011-g2**](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA95546&NAME=OctHS.APNO3011-g2) **79 738 0.0 98.33**

**4.** [**OctHS.APNO1593-b2**](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA95546&NAME=OctHS.APNO1593-b2) **61 736** [**OctHS.ATYB6889-g2**](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA95546&NAME=OctHS.ATYB6889-g2) **742 67 0.0 96.3**

**5.** [**OctHS.APNO1593-b2**](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA95546&NAME=OctHS.APNO1593-b2) **61 705** [**OctHS.ATYB2608-b2**](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA95546&NAME=OctHS.ATYB2608-b2) **696 52 0.0 97.21**

**6.** [**OctHS.APNO1593-b2**](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA95546&NAME=OctHS.APNO1593-b2) **101 818** [**OctHS.APNO1356-b2**](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA95546&NAME=OctHS.APNO1356-b2) **268 985 0.0**

After a careful evaluation, I determine which read to choose from the list using the Start and End points from my read (Query) and all the other similar reads (Target). I took in to consideration that how much of the query matches and how long the target is to determine how good the match is. I combined the similar reads I have chosen with my read using a function called join on biobike.

([JOIN](http://biobike.csbc.vcu.edu:9003/help/function?PKG=TIHITINA97326&PACKAGE=BBI&SYMBOL=JOIN) TARGET11 ([SEQUENCE-OF](http://biobike.csbc.vcu.edu:9003/help/function?PKG=TIHITINA97326&PACKAGE=BBI&SYMBOL=SEQUENCE-OF) OCTHSE.APNO1593-B2))

By joining all the overlapping reads I managed to get a longer read that has a length of 4131. Ones I make my read long enough to find gene, I used a sequence comparison tool NCBI to identify genes in my sequence read. I found about 8 genes in my sequence. I have also managed to find hypothetical protein in my sequence.

I have also tried to identify an open frame in my sequence I used the biobike tool and find the following: F indicates forward direction and B indicates backward. Similarities to other s were limited to specific open reading.

Open reading Frame:

* ((F 215 529) (F 947 1579) (F 1600 1980) (B 1990 2313) (F 2068 3171) (B 3704 4033))

Using tBlastX in NBCI I blasted my sequence to find comparison similarity with any known viruses. To be accurate on the results I used biobike as well. Comparing the results from the two, the result from biobike gave me a result with only two of virus that was also obtained from the NBCI result. I used the percent significant identity value and the E-value in order to come to a decision whether my sequence might come from one of the organisms’. I conclude that my sequence might come from sulfolobus islandicus rod-shaped virus. This virus has similar filamentous particles containing thermophilic Archaea in the kingdom cerenracheota. This makes sense since my sequence come from hot spring which a thermophilic area.

Sulfolobus islandicus rod-shaped virus 2 was the organism with a significant similarity 88% identity. The Descriptions of this virus is as described below:

## Virion Properties

### Morphology

Virions consist of a capsid. Virus capsid is not enveloped. Capsid/nucleocapsid is elongated and exhibits icosahedral symmetry. The capsid is rod-shaped and rigid.

### Nucleic Acid

The genome is not segmented and contains a single molecule of linear single-stranded DNA.

## Classification

This is a description of an archaeal virus at the species level.

**The following figure describes the description of my read and all the reads that are overlapping with:**

[OctHSe.APNO1593-b2](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA67160&NAME=OctHSe.APNO1593-b2)

[OctHSe.ATYB5969-g2](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA67160&NAME=OctHSe.ATYB5969-g2)

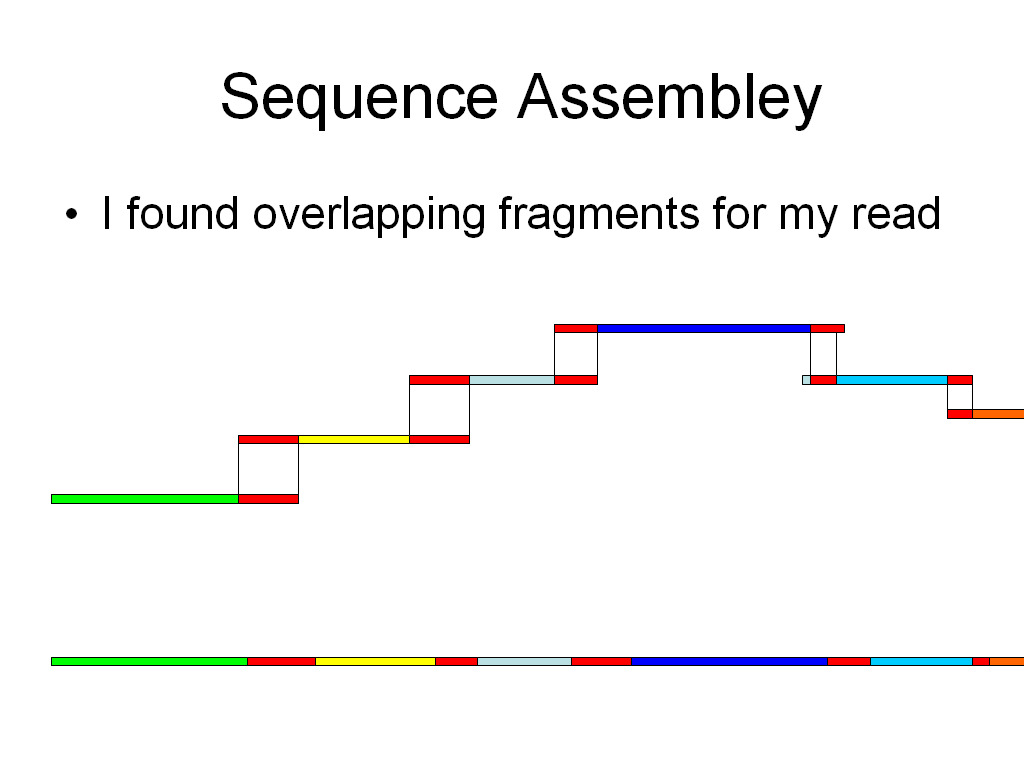
[OctHSe.ATYB3725-g2](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA67160&NAME=OctHSe.ATYB3725-g2)

OctHSe.APNO3549-g2

OctHSe.ATYb2671-g2

OctHSe.ATYB638-b2

Overlaps



**Gene Predictions:**

#### Information on input sequence

Sequence title: Fri May 8 15:22:38 EDT 2009

Length: 4135 bp

G+C Content: 48.73 %

#### Parse predicted by GeneMark.hmm 2.0

GeneMark.hmm PROKARYOTIC (Version 2.6r)

Sequence file name: sequence, RBS: N

Model file name: heuristic\_no\_rbs.mat

Model organism: Heuristic\_model\_for\_genetic\_code\_11\_and\_GC\_30

Fri May 8 15:22:38 2009

Predicted genes

Gene Strand LeftEnd RightEnd Gene Class

# Length

1 + <2 529 528 1

2 + 526 774 249 1

3 + 822 1013 192 1

4 + 1010 1579 570 1

5 + 1604 1771 168 1

6 + 1828 1980 153 1

7 + 2068 3171 1104 1

8 + 3471 >4133 663 1

Predicted proteins:

>gene\_1|GeneMark.hmm|175\_aa|+|2|529 >Fri May 8 15:22:38 EDT 2009

ALKSDRGTCSGPTQYSARHCFGSPLSGVKNRKYEVLGYDVIGRVVKDYSPPLWNLVAQLS

RFFGQKFGIRDVVRVSAPYWSSDNQIAVVAGGVLGVPVGLFAPVPTIEPEDYVDRDVVVD

VDVPEQDTISCYVFDYSTIDIFFDKVPWPNDVFVAQCKKPLRPGNSGGPVRLLRQ

>gene\_2|GeneMark.hmm|82\_aa|+|526|774 >Fri May 8 15:22:38 EDT 2009

MRKVRALLDPILDVFSSLGVYTYSVFGYLTSRLDYISQVMIVLMFTVYTVLKVKEDPVAV

SNYGYFLGGYFIGAQGGVLFTT

>gene\_3|GeneMark.hmm|63\_aa|+|822|1013 >Fri May 8 15:22:38 EDT 2009

MGRRLKVHITATISPESGEYLEWVADQFFNGNRSEALDEIIKWSKENGFHIVVAPFLAKQ

RKS

>gene\_4|GeneMark.hmm|189\_aa|+|1010|1579 >Fri May 8 15:22:38 EDT 2009

MRPKRVHELIVCPQFDNRGYDESRHGGHAVEAILAEILARYGFEAQAEIEYRGIVGHPDF

VRVLDNYVEFIEVKNTGQIMYTHILQVGMYKSLLYKVYNKPVIGYILYVRYRVVFGDSPI

MPRWVQDLGVQYIYLPMDAGENYVNWGLFRVSYRSKLAGPYCAYCRNDECKIKYIIINGP

GNSILDLEE

>gene\_5|GeneMark.hmm|55\_aa|+|1604|1771 >Fri May 8 15:22:38 EDT 2009

VLSHLWLGGTEKTEAMVKGQGFFPEAAGGTGATPAGLGTPRPPQGIKVSLGPTSI

>gene\_6|GeneMark.hmm|50\_aa|+|1828|1980 >Fri May 8 15:22:38 EDT 2009

MPGQYIVLEIVREAENQLGYSFRTVVLGDNKADSRPKVEEAPEERPQDWV

>gene\_7|GeneMark.hmm|367\_aa|+|2068|3171 >Fri May 8 15:22:38 EDT 2009

VSIDAAMLINVPDPMKVILGLNKQLFNAAKPDSFISRPVGYQVRGNSIYVKYPNGVKAML

PLLGKTEEERVRYALKVARSWHVVVANSRGYVVKELMDLVADVEAAKIVVDAVGKRMALA

MAMGLKPEALDGYWHRAAVLITRAHVMELSQPATGKSTFGWYISRVLGGIVVNEPPTVAF

LAGDARDGSFGVAFTSQVLIFDEIDKWTKKGESMSQTMDVLLSGMENCTWARGAGRGLTY

SKSLSTLWFENTPVTLAEVPRDRAKIPEIMTKLLKVNAGPLIDRISIIATEAPRFRTDML

GDMLKPAVIRGLSKILTEEAQQRWDQLRQQGWDARSSWHLAVILTMLEWTKKKEASIDDA

IGIYKMQ

>gene\_8|GeneMark.hmm|221\_aa|+|3471|4133 >Fri May 8 15:22:38 EDT 2009

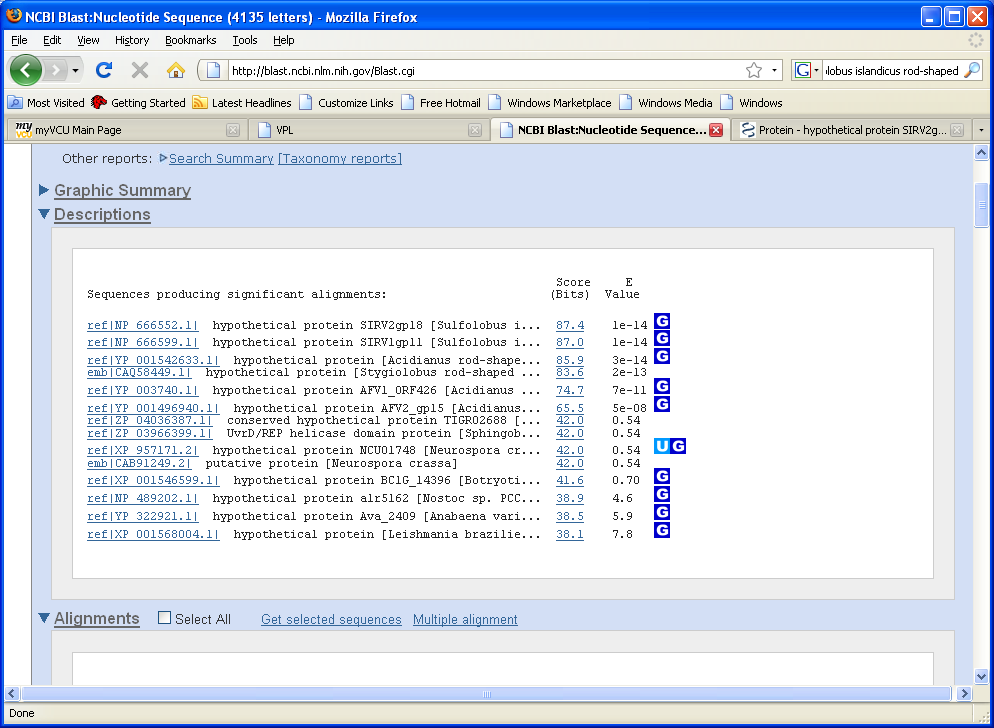
MSAVGSPRVLPHELRLALTYAKMQKRIYAPYIAAIQATLNEAFGPMGILAYGVSRGTITA

DDVSRYTSYVKQTVATRLQTLGVSFDFDAAPVSGGISVAWTAVKGANQHSNSHTITGSVE

DVSDDWIQSLSIAVDRFKRWAQTLRPHKVKDILEAAAVGYEAFMTEGTPSPRQALAVVKH

KLRLSAVNPALYVQQLRNSGDKYQIGYQLAETEVQELIRGA

**Blast result for comparison search:**



**BIOBik comparison search in Know virus:**

QUERY Q-START Q-END TARGET T-START T-END E-VALUE %ID

1. Seq1 2125 3024 [NC\_004086.p-SIRV2gp18](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA97326&NAME=NC_004086.p-SIRV2gp18) 51 375 6.0d-17 24.85

2. Seq1 2125 3024 [NC\_004087.p-SIRV1gp11](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA97326&NAME=NC_004087.p-SIRV1gp11) 55 379 8.0d-17 24.85

3. Seq1 2500 3045 [NC\_005830.p-ORF426](http://biobike.csbc.vcu.edu:9003/frame?PKG=TIHITINA97326&NAME=NC_005830.p-ORF426) 189 380 4.0d-13 28.79

([SEQUENCE-SIMILAR-TO](http://biobike.csbc.vcu.edu:9003/help/function?PKG=TIHITINA97326&PACKAGE=BBI&SYMBOL=SEQUENCE-SIMILAR-TO) Mysequence IN \*KNOWN-VIRUSES\*  
 :TRANSLATED-DNA-VS-PROTEIN)

**The Graphical summary of the results from the comparison such:**

