

## **Ensamble del genoma de un bacteriófago de *Vibrio cholerae***

Mauricio Torres, Alonso Peón, Diego Gimeno y Diego Morett

**Colegio Marymount**

Estrella del Norte #6 Col. Rancho Tetela CP. 62160, (777)3124277

[colegio@marymount.edu.mx](mailto:colegio@marymount.edu.mx)

**Asesor:** Ricardo Grande (IBt UNAM)    **Co-asesor:** Leonardo Collado (IBt UNAM)

### **Resumen**

Este proyecto consistió en el ensamble de un bacteriófago encontrado en el Estado de Hidalgo, en un lago de aguas negras llamado Endhó. Investigadores de la UNAM estipularon que este virus ataca a la bacteria *Vibrio cholerae*, lo cual los llevó a creer que se podía tratar de una especie nueva de virus. Los investigadores de la UNAM establecieron esta hipótesis porque solo existen dos bacteriófagos de *Vibrio cholerae* conocidos, y ninguno de los dos es nativo de México. Nuestro trabajo consistió en verificar si este bacteriófago era una especie nueva. Para saber esto, lo que hicieron investigadores de la UNAM fue tomar una muestra del ADN del bacteriófago para después secuenciarlo en el *Genome Analyzer-Ilx*. Se nos proporcionó la secuencia del virus y comenzamos a realizar el ensamblado usando diferentes programas como *Velvet*. Una vez obtenido el genoma ensamblado, éste se comparó con otros genomas en una base de datos en línea llamada *Blast* para poder apreciar las diferencias en la estructura entre el genoma del virus nuevo y los genomas de los virus en las bases de datos. Se obtuvo como resultado que el genoma del virus que fue secuenciado ( $\phi$ iVC8) es 98 % idéntico al genoma de otros dos virus ( $\phi$ iVC2 y  $\phi$ iVC5) que atacan a la misma bacteria (*Vibrio cholerae*). Esta diferencia permitió concluir que el virus  $\phi$ ivc8 sí es suficientemente diferente para ser considerado una especie de virus nuevo.

### **Introducción y Antecedentes**

Los bacteriófagos son virus que infectan a bacterias, su nombre significa “devorador de bacterias”; Felix d’Hérelle los descubrió en 1917 y los nombró así (Encyclopedia Britannica, 2010). Estos organismos no están vivos ya que no se pueden reproducir por sí mismos y necesitan forzosamente un hospedero para poder replicarse.

Frecuentemente, los bacteriófagos son muy selectivos y sólo atacan a una especie de bacteria. Los *fagos* que atacan bacterias patógenas se podrían utilizar para prevenir enfermedades.

El ADN es el material genético donde está almacenada toda la información que necesita un organismo para poder llevar a cabo sus funciones. El ADN está formado por grupos fosfato ( $\text{PO}_4^{-3}$ ), desoxiribosa, y cuatro bases nitrogenadas: adenina (A), timina (T), guanina (G) y citosina (C). Las bases se unen por puentes de hidrógeno y son complementarias; la A siempre se junta con la T y la C con la G. La secuencia de bases se divide en tripletes llamados *codones*; cada codón codifica para un aminoácido, los cuales son las bases de las proteínas. Existen veinte aminoácidos con los cuales se forman las proteínas de todos los seres vivos (Claros, sin fecha).

Por medio de la secuenciación masiva se puede conocer toda la cadena de bases nitrogenadas del ADN de un organismo. Las técnicas de secuenciación han avanzado mucho durante los últimos años. Actualmente se puede secuenciar el genoma humano en diez días. Los nuevos métodos de secuenciación necesitan romper el genoma en cadenas pequeñas, y amplificar éstas miles de veces para poder reconocer las diferentes bases. Después, las partes secuenciadas se deben de volver a armar para reconstruir el genoma del organismo.

Investigadores de la Facultad de Medicina de la UNAM han hecho estudios sobre bacteriófagos que atacan a la bacteria *Vibrio cholerae*, que es la que causa el cólera: una enfermedad que causa diarrea, vómito y malestar general (Wikipedia, 2010). Se ha observado que esta enfermedad presenta ciclos; hay épocas en las que se propaga más, seguidas por épocas en las que casi no se presentan casos. Los investigadores de la Facultad de Medicina de la UNAM encontraron un virus bacteriófago en el lago Endhó, un lago de aguas negras

ubicado en el estado de Pachuca. El *fago* que se encontró aniquila a la bacteria del cólera.

Este virus se secuenció unas semanas antes del inicio del presente proyecto en el Instituto de Biotecnología de la UNAM, en la Unidad Universitaria de Secuenciación Masiva de DNA (<http://www.uusmd.unam.mx/>). Nuestro trabajo consistió en armar la secuencia del ADN del virus y compararlo para determinar si tiene las características para ser considerado un organismo nuevo.

### **Hipótesis**

El bacteriófago secuenciado será semejante a algunos previamente conocidos, pero lo suficientemente diferente para ser considerado un tipo de organismo nuevo.

### **Objetivo**

Ensamblar la secuencia del bacteriófago para comparar la estructura genómica del virus armado con los virus de las bases de datos.

### **Materiales**

Los investigadores de la UNAM utilizaron el *Genome Analyzer Iix*, una máquina de secuenciación masiva que se encuentra en el Instituto de Biotecnología de la UNAM, para secuenciar el ADN del bacteriófago. Para secuenciar primero se tiene que extraer una muestra de ADN del organismo que se quiere estudiar. Esta muestra se replica muchas veces y se fragmenta usando un proceso mediante el cual el ADN disuelto en fenol (tipo de alcohol) se pasa por tubos delgados a presión, dispersándose al salir, y ocasionando la fragmentación de las cadenas de ADN. Los fragmentos de ADN se corren por un gel especial para poder escoger los fragmentos de un tamaño específico. Los fragmentos se introducen a la máquina con indicadores específicos para cada una de las bases y mediante un microscopio y un láser, se toman fotografías. En estas fotografías cada base nitrogenada aparece de un color diferente y así se conoce la secuencia de ese fragmento (Illumina, 2010).

Ya con la secuencia se utilizó el programa *Velvet* para ensamblar el genoma del virus, el cual nos fue proporcionado por nuestro co-asesor Leonardo Collado, y fue utilizado en el Instituto de Biotecnología, ya que el programa sólo puede ser operado en el sistema operativo Linux.

Una vez ensamblado el genoma, se comparó usando una base de datos llamada *Blast*, que es un servidor gratuito donde se almacena información genómica de muchas instituciones, principalmente universidades.

## **Métodos**

El proceso del desarrollo del proyecto se inició cuando nos proporcionaron la secuencia del virus. Lo primero que se hizo fue tomar los *contigs* (ver glosario al final) dados por el secuenciador, que tenían una longitud de 36 bases nitrogenadas, y se armó nuestra “biblioteca”. La biblioteca consiste en armar los llamados *K-mers*, que son palabras de longitud K, en este caso utilizamos palabras de 23 bases. Lo que esto significa es que a partir de los *contigs*, se van creando estos *K-mers* tomando primero las primeras 23 letras. Después se crea el segundo *K-mer* tomando de la letra 2 a la 24 y así sucesivamente. Esto resulta en 13 secuencias diferentes desplazadas una base con respecto a la anterior. Si se topa con dos o más *K-mers* iguales, en vez de crear dos entradas de biblioteca diferentes, sólo se marca la frecuencia del determinado *K-mer*.

*Velvet* es un programa usado para ensamblar genomas usando secuencias de corta lectura. Esto significa que cuando se trata de ensamblar, se usan cadenas pequeñas de nucleótidos entre 25 y 100. Nosotros usamos los *K-mers* que son de 23 bases nitrogenadas. En realidad la función de *Velvet* es quitar los errores del ensamblaje y también poder saber cuál es el orden del genoma completo. En algunos puntos puede llegar a hacerse nudos (también llamados “burbujas”) y el programa ayuda a saber cómo eliminar estos nudos y también saber con un alto grado de certeza cuál es el verdadero orden de la secuencia.

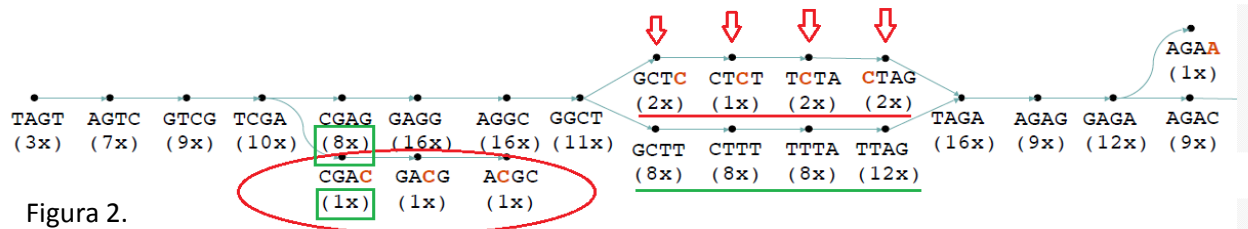
La figura 1, acompañada por el siguiente párrafo, describe el proceso mencionado previamente.



Figura 1.

El primer renglón donde está señalando la flecha es una parte del genoma ya ensamblado. Lo que hace *Velvet* es usar todos los *K-mers* (ver glosario al final) e ir armando el genoma. En este caso los *K-mers* serían todas las cadenas de siete bases nitrogenadas que están debajo del genoma. Se van alineando como se puede ver en donde están los tres rectángulos grises de la figura 1. Cuando hay Guanina, Timina y Citosina consecutivamente se alinean y los siguientes *K-mers* se van ajustando a este proceso. En el caso del rectángulo rojo hay puras Timinas, pero en un *K-mer* se encuentra una Citosina (en color rojo). La frecuencia de la Timina es de 11x ya que hay once Timinas en ese punto en específico y la de la Citosina es de 1x. La probabilidad de que haya una Citosina ahí es más baja que la de la Timina por lo que se deja la Timina en el genoma (representado por el primer renglón, señalado con una flecha roja).

Este mismo proceso se puede ilustrar con más claridad en la figura 2, acompañada por una explicación en el siguiente renglón.



Los cuadrados verdes señalan dónde está la frecuencia de cada *K-mer*. En este caso la CGAG tiene una frecuencia de 8x y el GAGC tiene una de 1x. Los dos siguientes *K-mers*, que se encuentran en el círculo rojo, son también de 1x. Debido a que la cadena ahí termina y la frecuencia comparada con los *K-mers* de arriba es mucho menor, por lo que todo lo que se encuentra en el círculo se elimina. Pasa lo mismo en la “burbuja” siguiente (donde se encuentran las flechas rojas). Donde se encuentra la línea roja tiene menor frecuencia que donde se encuentra la línea verde, por lo que se eliminan todos los *K-mers* que están sobre la línea roja.

El genoma del bacteriófago es de aproximadamente 39,500 bases nitrogenadas. Una vez ensamblado el genoma, se guardó el ordenamiento de las bases en un procesador de textos que utiliza el formato .FASTA, el cual es un formato muy parecido al bloc de notas, pero sólo cuenta con un tipo de letra y está especializado para guardar las letras representantes de las bases nitrogenadas: A, T, C, G. Este archivo se subió a un servidor llamado BLAST para comparar a nivel de estructura genómica el genoma de nuestro virus con todos los virus existentes encontrados en las bases de datos. BLAST (<http://blast.ncbi.nlm.nih.gov/>) es un servidor al cual cualquier persona puede acceder, ya que se encuentra en línea. En este servidor están archivados todos los genomas secuenciados que se han hecho públicos. Por medio de este servidor se puede comparar una secuencia con las bases de datos, y obtener las secuencias de los otros organismos a los que

ésta se parece, y en qué porcentaje. La comparación se hace a nivel de bases nitrogenadas; es importante hacer notar esto ya que a nivel de aminoácidos se obtienen algunas diferencias, ya que hay codones que codifican para el mismo aminoácido y aunque las bases sean diferentes pueden ser el mismo aminoácido. La figura 3, junto con el procedimiento a seguir explicado en la parte inferior, ilustra la página en donde se proporciona el genoma del organismo para realizar la comparación.

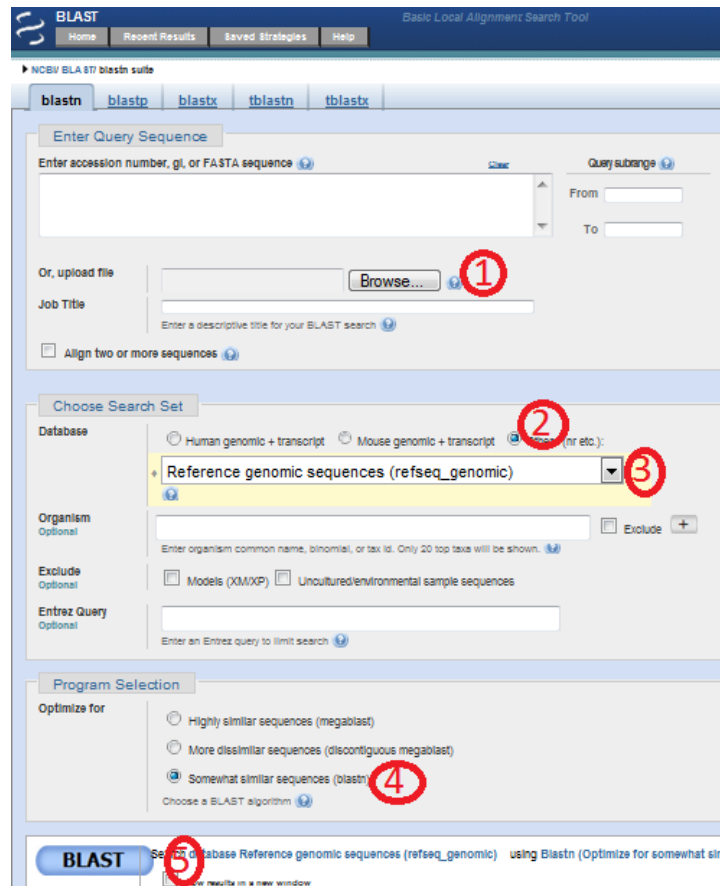


Figura 3.

El procedimiento es el siguiente:

1. En esta sección se puede subir el archivo .FASTA, haciendo click en *Browse* y seleccionando el archivo en la carpeta donde el archivo .FASTA está guardado.
2. Se especifica el tipo de organismo que se quiere comparar. Hay tres opciones: Humano, ratón u otros. En este caso seleccionamos “otros”.

3. Se especifica en qué nivel se quiere comparar el genoma. Se puede comparar a nivel genómico, a nivel de proteínas, a nivel de RNA, etc. En este caso se comparó a nivel genómico, es decir se compara cada una de las letras ATCG con los genomas de otros virus.
4. Se establece qué nivel de semejanza se busca. Muy parecido, algo parecido, o muy poco parecido. En este caso se utilizó algo parecido, ya que esperábamos que los virus fueran similares, pero con genomas diferentes.
5. Se hace click a este botón para iniciar la comparación.

## Resultados

El resultado de la comparación de *Blast* se muestra en la figura 4.

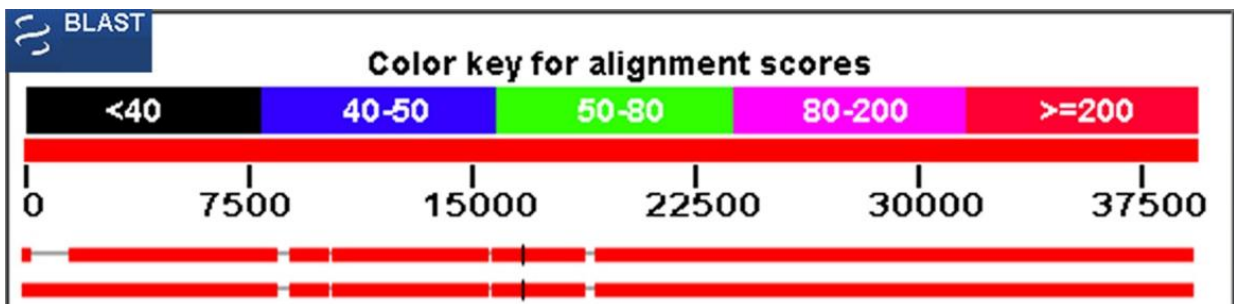


Figura 4.

En esta figura se representa la similitud que existe entre el virus que estudiamos con los otros dos virus que atacan al *Vibrio cholerae*. La barra roja en la parte superior de la figura 3 representa el genoma del virus estudiado. Las dos barras inferiores representan a los otros dos virus VP2 (arriba) y VP5 (abajo). Las secciones de las barras inferiores que se encuentran coloreadas de rojo son las secciones en los genomas que son idénticos al virus que estudiamos. Las secciones blancas con una línea gris horizontal representan las secciones de los genomas que difieren con el virus estudiado. La barra vertical negra también representa una diferencia pero es tan pequeña que se tiene que marcar con una línea para poderla apreciar. No se utilizaron las coloraciones que aparecen en la parte superior de la figura 4 ya que no hubo diferencias con esos niveles de similitud, sólo se presentaron niveles menores que 40 puntos (coloración negra) y

mayores o iguales a 200 puntos (coloración roja). Podríamos imaginar las barras rojas como una palabra de 39,500 letras de ATCG. Lo que hace *Blast* es alinear los genomas de los virus y comparar letra por letra para ver si son iguales o si difieren.

La comparación en *BLAST* mostró que el virus estudiado comparte el 98 % del material genético con otros dos virus que también atacan a *Vibrio cholerae*, llamados VP2 y VP5.

### **Conclusiones**

Aunque este 98 % de similitud parezca un parecido muy grande, sí es una diferencia representativa a nivel genético, sobre todo considerando que el genoma del ser humano es 99 % idéntico al del chimpancé (The Chimpanzee Sequencing and Analysis Consortium, 2005), y por lo tanto se puede concluir que el bacteriófago estudiado es un tipo de organismo nuevo, el cual fue nombrado phiVC8.

### **Reconocimientos**

Agradecemos al Dr. Ricardo Grande por proporcionarnos la secuencia del virus, y asesorarnos con la metodología del proyecto. Agradecemos también al Lic. Leonardo Collado por proporcionarnos las herramientas necesarias para realizar el ensamblado del genoma, y asesorarnos en el proceso mismo. Agradecemos al Dr. Enrique Galindo por supervisar nuestro proyecto.

### **Bibliografía**

- Claros, Gonzalo (-) *Estructura: Bases Nitrogenadas*, consultado (10/02/10)  
[http://sebbm.bq.ub.es/BioROM/contenido/av\\_bma/apuntes/T2/t2\\_bn.htm](http://sebbm.bq.ub.es/BioROM/contenido/av_bma/apuntes/T2/t2_bn.htm)
- Encyclopedia Britannica (2010), *Félix d'Hérelle*, consultado (10/02/10)  
<http://www.britannica.com/EBchecked/topic/262988/Felix-d-Herelle>

Illumina (2010) *Genome Analyzer Iix*, consultado (10/02/10)

[http://www.illumina.com/systems/genome\\_analyzer\\_iix.ilmn](http://www.illumina.com/systems/genome_analyzer_iix.ilmn)

The Chimpanzee Sequencing and Analysis Consortium (2005). *Initial sequence of the chimpanzee genome and comparison with the human genome*. *Nature* **437** (7055): 69–87

Wikipedia (2010) *Cholera*, Wikipedia, consultado (10/02/10)

<http://en.wikipedia.org/wiki/Cholera>

### **Glosario de definiciones básicas**

**Bacteriófago:** un virus que infecta exclusivamente bacterias.

**Contig:** un conjunto de segmentos sobrepuestos de ADN proveniente de una muestra de ADN. Se puede utilizar para deducir la secuencia de ADN.

**Base Nitrogenada:** son compuestos orgánicos cíclicos, que incluyen dos o más átomos de nitrógeno. Son parte fundamental de las cadenas de ADN. Existen cuatro: Adenina, timina, guanina y citosina.

**K-mer:** “palabras” de longitud K, que se usan para armar la secuencia del genoma, superponiéndolos uno sobre el otro.