Estudio de endolisinas

Ricardo Elías Ávalos Rojas.
Estudiantes Ingeniería Civíl Matemática.
Especialista externo: O'bryan Cárdenas Andaur
Profesor Guía Pedro Gajardo.
ricardo.avalos.14@sansano.usm.cl
Universidad Técnica Federico Santa María
Departamento de Matemáticas

1. Contexto y objetivo.

El estudio de virus es un área en constante investigación, de una de estas investigaciones surge la necesidad de identificar, las endolisinas (proteínas), que están presentes en los fagos (virus que infectan a las bacterias) menos exitosos y los más exitosos, este exito se mide por medio de un termino de acuñación propia del Centro de Biotecnología Doctor Daniel Alkalay Lowitt UTFSM, este termino es la promiscuidad, se denomina que un fago es altamente promiscuo cuando, el fago esta presente en una gran cantidad de bacterias, de manera análoga cuando este se encuentra en una pequeña parte de las bacterias.

Teniendo las endolisinas asociadas a los fagos bajamente y altamente promiscuos, se debe realizar un estudio estadístico que permita relacionar las endolisinas con el éxito de los virus. Para lo cual se cuenta con un listado de más de 800 fagos que se encuentran en el formato bioestadistico FASTA. El formato FASTA es utilizado para representar secuencias bien de ácidos nucleicos, bien de péptido, y en el que los pares de bases o los aminoácidos se representan usando códigos de una única letra.

De los 800 fagos listados, se tiene la promiscuidad de aproximadamente 120 de ellos, debido a que son los fagos de la lista de 800 que se encontraron en un muestreo de bacterias, estos 120 fagos son los que se separan en altamente promiscuos y bajamente promiscuos.

2. Definiciones y caracterizaciones del problema.

X_i Es el identificador de un virus i. i $\in$ { 1, 2, . . ., 800 }

Estos objetos son strings de largo variable, pero teniendo un máximo de 500 caracteres estos caracteres pertenecen a un conjunto de letras que se encuentran en formato FASTA. Procurando que las palabras tengan el mismo largo estas se extenderán en caso de ser necesarios con caracteres J letra no definida en el formato FASTA esto con el fin de poder obtener el mismo largo para el trabajo de las funciones, haciendo que las funciones identifiquen a J como un carácter nulo.

$\overline{X} = \cup_{i=1}^{800}$ es el listado de los virus con los que se trabajara.

3. Primeros acercamientos.

El estudio de homología que se realiza requiere generar de manera computacional cadenas de variaciones de endolisinas, en otras palabras a partir de una base de datos se van construyendo variaciones de las endolisinas conocidas, para así compararlas con secuencias de la cadena X_i esto pues puede haber una pequeña mutación de esta endolisina, por esto surge el porcentaje de homología que representa una comparación entre la endolisina original y la detectada en la cadena, en primera instancia se trabajo en la automatización de este proceso por medio de una herramienta llamada hh-suite [1]. Luego de tener problemas con algunas bases de datos se opta por tomar la herramienta HHpred [2] que tiene incorporada a HHblits un programa que cumple la función de hh-suite y cuenta con herramientas adicionales como lo son la comparación de estructura secundaria, para mejorar la homología. Lamentablemente se vuelve a tener problemas con cierta base de datos que ya no existe. Ante esto se empieza a realizar una búsqueda de bases de datos que puedan realizar el trabajo necesario, por lo cual se comienzan a contrastar los resultados emanados de las nuevas bases de datos, con los del estudio "Molecular Aspects and Comparative Genomics of Bacteriophage Endolysins"[3]

4. Enfoque matemático.

Lo anterior desde el enfoque matemático viene siendo.

Se define la función *F* función que entrega pares ordenados que contienen un identificador de endolisinas y un porcentaje de homología.

a_j:= identificador de la endolisina j.

r_j:= porcentaje de homología entre una secuencia de X_i y la endolisina j.

$$F : \overline{X} \rightarrow M[nx2]$$

$$X \rightarrow \begin{pmatrix} a_1 & r_1 \\ a_2 & r_2 \\ \vdots & \vdots \\ a_n & r_n \end{pmatrix}$$

$$\begin{pmatrix} a_1 & r_1 \\ a_2 & r_2 \\ \vdots & \vdots \\ a_n & r_n \end{pmatrix} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_n \end{pmatrix}$$

Observación: Notar que el numero de pares ordenados que se encuentran en la matriz, depende del identificador de fago al que se le aplique la función, por lo que estas dimensiones no son constantes.

5. Ideas.

Como ya se ha mencionado el estudio requiere poder identificar qué endolisinas se encuentran en un virus especifico, por lo que es necesario introducir la siguiente función

G Función que determina si el virus X_i contiene la endolisina a_j

$G(\overline{X}, \overline{a}) \rightarrow \{0,1\}$; 0 no se encuentra relación y 1 si existe relación.

Adicionalmente se buscan endolisinas que estén presentes en la muestra de 120 fagos los cuales tienen asignados sus índices de promiscuidad, por lo que se define.

$R(a_j) = \frac{\sum_{i=1}^{120} G(X_i', a_j)}{m}$; donde X' es un identificador de los 120 fagos "promiscuos".

R representa el porcentaje de presencia de una endolisina en una muestra de virus, por lo que es claro que el recorrido de R es [0. 1]

Luego de esto se debe utilizar un criterio para determinar qué endolisinas influyen positiva o negativamente en la promiscuidad de un fago. Por lo cual se podría utilizar la misma idea anterior, pero restringiendo los fagos a los más promiscuos y a los menos promiscuos.

6. Resolución.

El problema lamentablemente se encuentra estancado de momento, pues las bases de datos equivalentes no han sido encontradas, tras simulación de datos se presentan diferencias importantes a los resultados de control del estudio mencionado [2] por lo cual no se ha podido realizar el estudio estadístico que se plantea como objetivo. Por lo que la generación de un criterio para determinar qué endolisinas influyen positiva y/o negativamente en la promiscuidad de los fagos es un criterio bastante básico que no se pudo refinar.

7. Conclusiones.

En términos generales se podría decir que el trabajo depende principalmente de encontrar una base de datos para generar alineamientos de cadenas de aminoácidos satisfactoria, con lo cual se podría realizar un trabajo un poco más fino, respecto al objetivo del estudio, teniendo una vez los datos, tal ves estas presenten ciertas cualidades de distribuciones estadísticas, lo cual permitiría mejorar criterios expuestos o cambiarlos.

Otra consideración a tener en cuenta es cuanto porcentaje de homología es el mínimo aceptable para considerar que la homología es significativa y tenerla como una "garantía", si bien se cito que el 20 % es una cota mínima, este numero podría ser variado para garantizar la confiabilidad del estudio.

8. Bibliografía

1 hh-suite programa de codigo libre, para distribución en Linux

`https://github.com/soedinglab/hh-suite`

.

2 HHpred

`https://toolkit.tuebingen.mpg.de/#/tools/hhpred`

.

3 "Molecular Aspects and Comparative Genomics of Bacteriophage Endolysins"

`http://jvi.asm.org/content/87/8/4558.full`

.