Reporte 2: Predicción de promotores
Bioinformática II: Bioinformática Estructural
Licenciatura en Ciencias Genómicas
Integrantes del equipo:
Jessica Samantha Cruz Ruiz
Lorena Elizabeth Fajardo Brígido

La predicción de promotores ha sido una tarea muy importante a lo largo de los años, debido a que en esta parte del DNA es donde varias proteínas, entre ellas la RNA polimerasa, comienzan el proceso de transcripción. Se han diseñado distintos predictores, basándose en secuencias ya conocidas, como distintas cajas donde ya se sabe que un factor de transcripción puede actuar, pero en este reporte se expondrá el desarrollo de un predictor basado en qué tan inestable está la secuencia promotora tomando en cuenta la energía libre de Gibbs.

Las instrucciones que se siguieron en este trabajo y sus respectivos resultados fueron los siguientes:

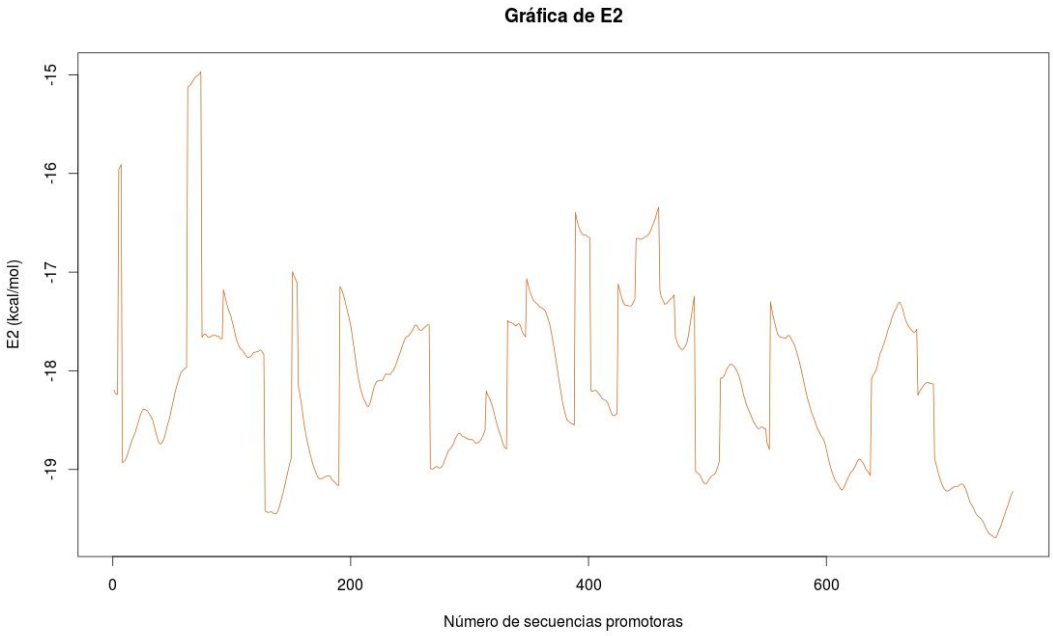
1) Completar el código fuente del programa 1.1 para implementar el predictor de Kanhere y Bansal, justificando los valores de cutoff1 y cutoff2 de acuerdo con las figuras de su artículo.

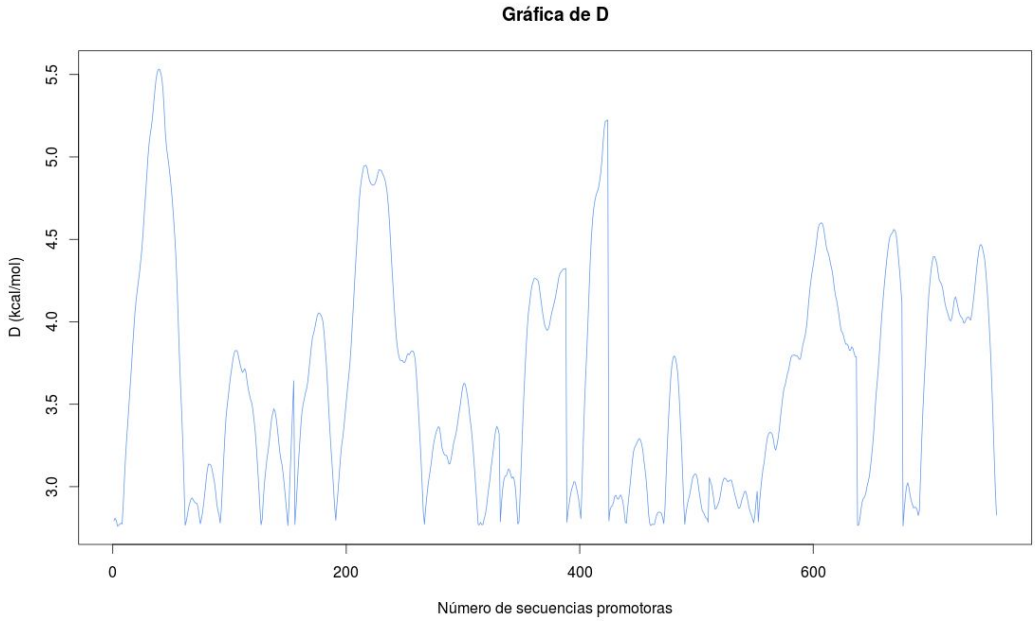
Para la elaboración del predictor de promotores, se utilizó el lenguaje de programación Perl. Al código que ya se nos fue proporcionado en clase se le añadió la visualización de las secuencias por ventanas de 15 nucleótidos, la corrección por simetría y el cálculo de los valores E1, E2 y D. El script se encuentra en el archivo “promotores.pl”.

Los valores de cutoff1 (D) y cutoff2 (E1) fueron elegidos con base en el artículo de Kanhere y Bansal. Ellos grafican y escriben en tablas distintos valores de D y E1, calculando su sensibilidad y su precisión, pero, dado que ambos parámetros se comportan de manera opuesta (con distintos valores de D y E1, cuando los valores de sensibilidad eran muy altos los de precisión muy bajos, o viceversa), elegimos los siguientes valores intermedios:

Sensibilidad	Cut-off D	Cut-off E1	Frecuencia de falsos positivos	
0.50	**2.76**	**-17.53**	1/3914	1/13737

2) Diseñar una figura donde se muestra gráficamente D, E1 y E2 para una posición n.





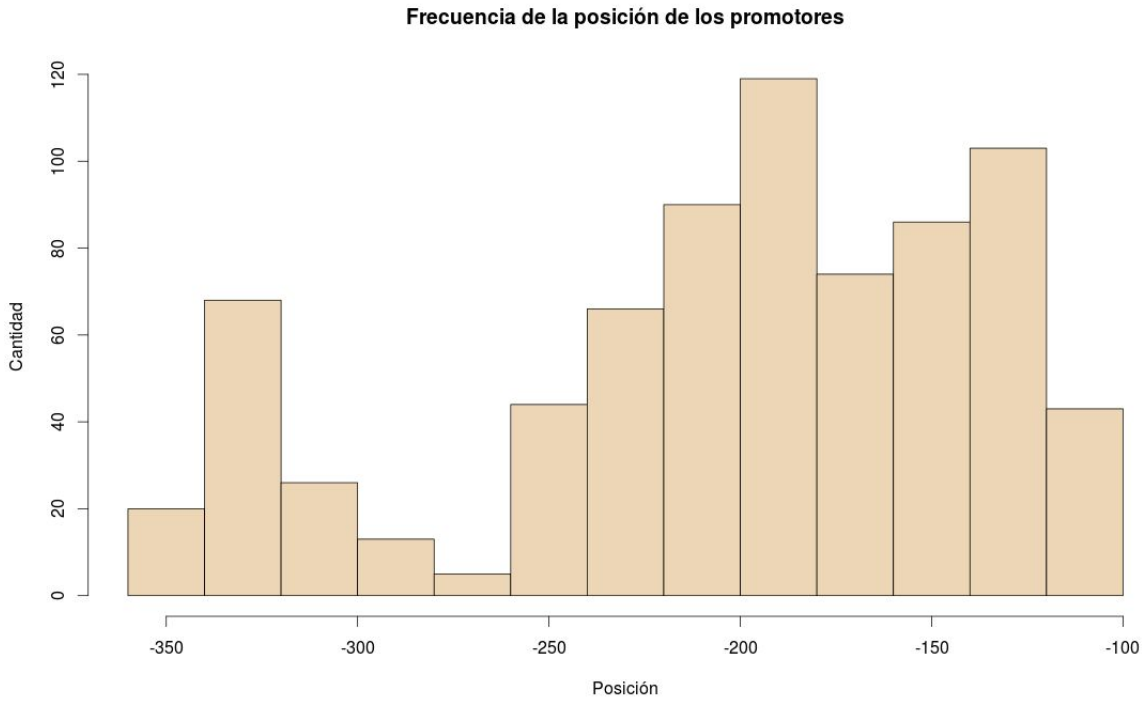
En estas gráficas se observa la distribución de los distintos valores de E1, E2 y D en las secuencias elegidas como promotoras. En E1 y D se comienza por los valores de cortes elegidos en el paso anterior. Para estas gráficas se hizo uso del lenguaje de programación R.

3) Predecir promotores en todas las secuencias del fichero K12_400_50_sites.

Los resultados de la predicción de promotores está en el archivo “resultados_finales.txt”. En la tabla la primera columna pertenece al nombre de la secuencia de donde se obtuvo el posible promotor. La segunda y la tercera indican los valores obtenidos de E1 y E2 respectivamente, la siguiente es el valor de D y las últimas dos columnas muestran el rango en donde se encuentra la secuencia promotora. En total fueron 757 las secuencias.

4) Graficar con qué frecuencia se predicen promotores en el intervalo -400,50. Con un breve comentario de los resultados es suficiente. Se les ocurre una manera de validar sus resultados, y calcular la tasa de FPs, usando RSAT::matrix-scan?

La gráfica de los promotores predichos se muestra a continuación:



En la gráfica se observa que la mayoría de los promotores se encuentra entre las regiones de -250 a -125, lo cual tiene sentido debido a que se encuentran río arriba de los genes. De igual manera, esta gráfica se hizo en R.

Nosotras pensamos que con la herramienta RSAT matrix-scan se podrían escanear las secuencias obtenidas por el predictor y compararlas con el genoma de E. coli, para verificar que estas sí se encuentran en la secuencia completa, visualizar sus posiciones y buscar los promotores que ya se han reportado, filtrando así cuáles son falsos positivos y cuales verdaderos positivos.