Bioinformática estructural

Tarea 2-Algoritmos 3D

Predicción de promotores

Carina Paola Cornejo Páramo

Claudia Saraí Reyes Ávila

INTRODUCCIÓN

Los promotores son regiones de DNA que contienen secuencias específicas de DNA que son reconocidas por factores de transcripción facilitando la unión de la RNA polimerasa para iniciar la transcripción de un gen. Los promotores están presentes tanto en organismos procariotas como en eucariotas, y deben tener propiedades especiales que permitan la separación de las cadenas de DNA para que pueda iniciar la transcripción, por lo que la secuencia de un promotor debe de conferir propiedades termodinámicas especiales que facilitan la apertura de las cadenas de DNA. El método de Kanhere, A. & Bansal, M. para la predicción de promotores se basa en que la disminución de la estabilidad del DNA es la característica más prominente de los promotores. El cambio en la estabilidad se daría como un salto brusco en los valores de energía libre de la región promotora haciéndola más propensa a la apertura. El método de Kanhere, A. & Bansal, M. tiene como ventaja que no se basa en la homología para detectar genes, pues no podría predecir genes que no tienen homólogos. Sin embargo el método aún no es eficiente en la predicción de promotores para genes de RNA y para promotores de otros organismos pues es muy eficiente en el caso de Eschericia coli y pero varía mucho con otras bacterias y no se aplica a promotores eucariotas. A pesar de que requieren mejoras, los métodos de predicción de genes basados en la detección de promotores y sitios de inicio de la transcripción son los más prometedores.

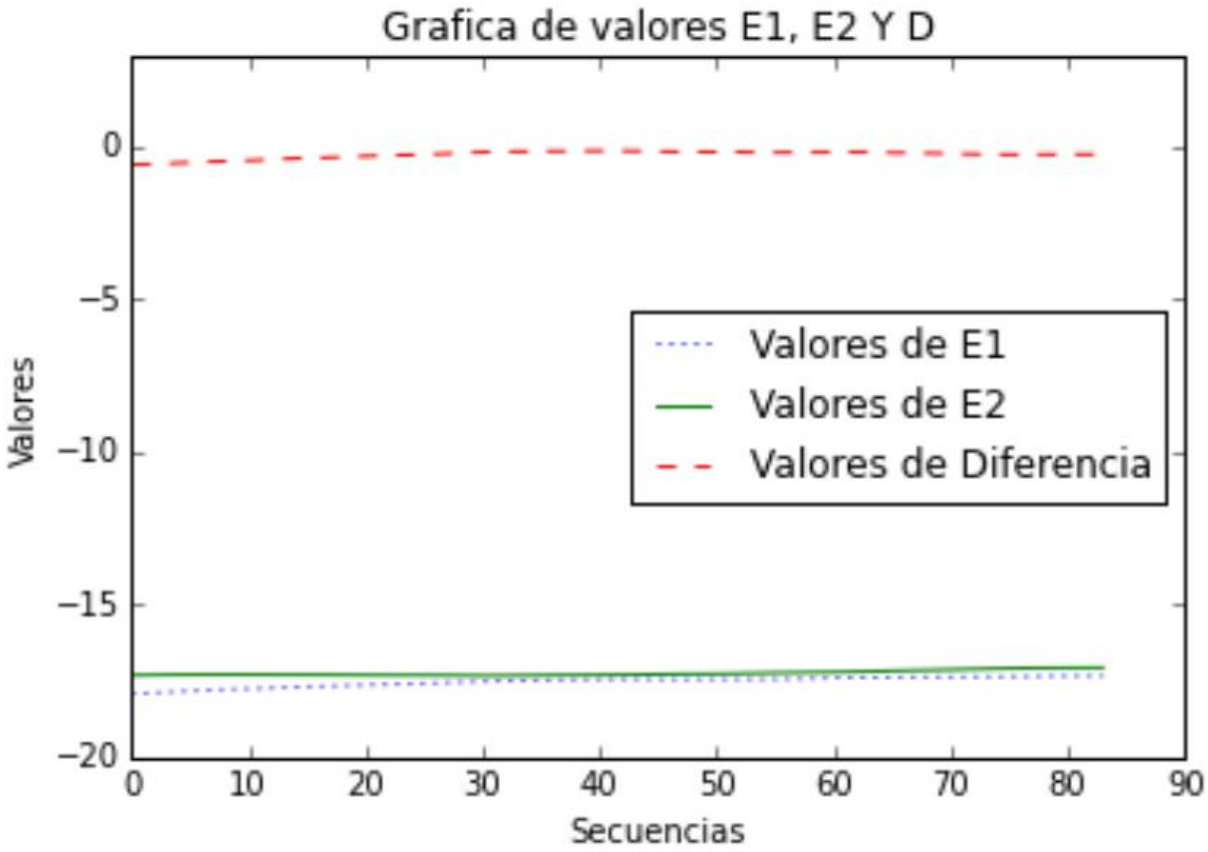
Desarrollo:

1) Completar el codigo 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. Es necesario comentar el codigo explicando sus cambios, por ejemplo con http://perldoc.perl.org/perlpod.html . Pueden usar el lenguaje de programación que quieran, siempre que haya un compilador disponible.

En el presente trabajo implementaremos el método de Kanhere, A. & Bansal, M. en lenguaje perl para predecir promotores en secuencias de 451 nucleótidos de Escherichia coli con coordenadas (-400,+50), las secuencias usadas se encuentran el el archivo K12_400_50_sites. Los valores de energía libre usados en el programa para cada dinucleótido fueron calculados experimentalmente.

El programa se encuentra en https://github.com/carinapaola/Bioinformatica_estructural/blob/master/PredictorPromotores.2.pl

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



La gráfica se generó usando el siguiente código de python

https://github.com/carinapaola/Bioinformatica_estructural/blob/master/Grafica-E1-E2-Diferencia_v3.0.ipynb

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

Para la predicción de promotores se usó el programa de perl, en el que usamos ventanas de 15 nucleótidos para calcular la energía libre de cada nucleótido, asignamos la energía libre de cada ventana al primer nucleótido de esta. Para calcular los valores de E1 consideramos los valores de delta G de 50 nucleótidos a partir de la posición n y para calcular E2 tomamos en cuenta los valores de delta G de 100 nucleótidos (desde n+99 hasta n+199). Los umbrales utilizados fueron -15.99 para E1 y 3.4 para la diferencia entre E1 y E2. Tratamos de maximizar la precisión con la que se detectan los promotores.

Consideramos que las posiciones que pasan los umbrales de E1 y diferencia que están a 25 posiciones o menos de distancia forman un mismo segmento, y que de cada segmentos la posición que representa una señal de promotor es la más cercana a la posición cero. Si estos valores se encuentran entre las posiciones -200 y 0 los consideramos como una señal de promotor, ya que éste se encontraría aproximadamente 50 nucleótidos más adelante de la señal(entre las posiciones -150 y 50), donde se da el cambio significativo de energía libre. En ocasiones encontramos más de un segmento con posiciones significativas en la misma secuencia, pero al considerar las posiciones en las que caen algunas de estas señales se consideran falsos positivos.

En las secuencias del archivo utilizado encontramos 35 promotores, el identificador de la secuencia y la posición que se considera como una señal de promotor, dado que los promotores se encontrarían ~50 nucleótidos más adelante, los promotores predichos estarían entre las posiciones -146 y -112 de las secuencias.

b0585 238

b0411 237

b0118 237

b0536 238

b0034 238

b0827 238

b0954 238

b1101 238

b0463 238

b1109 238

b0774 238

b0679 238

b0972 204

b0124 238

b1130 238

b0388 238

b0113 238

b0241 238

b1102 238

b0063 238

b0396 238

b1062 238

b0850 238

b0781 238

b0216 238

b1041 238

b0077 238

b0812 238

b0591 238

b0399 238

b0432 238

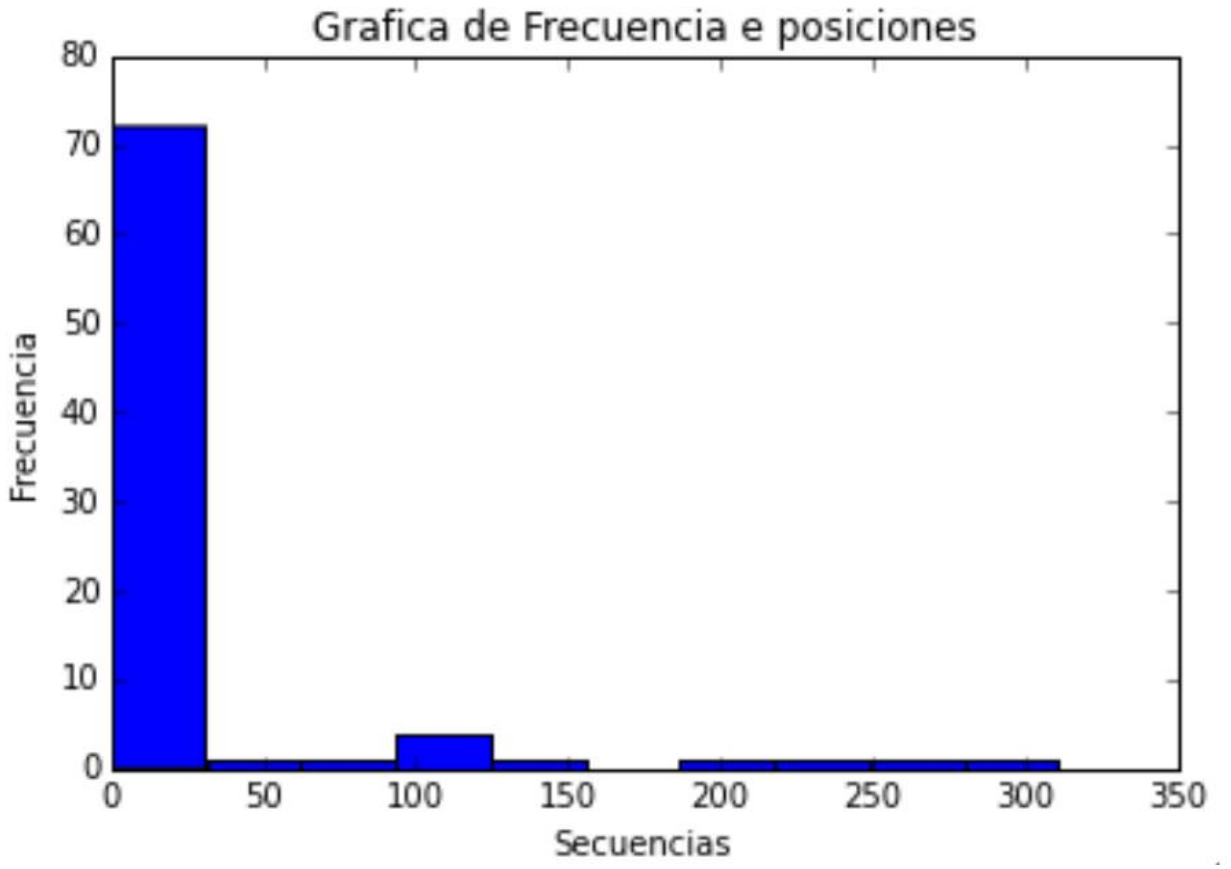
b0338 238

b0440 238

b0365 238

b0894 238

4) Graficar con qué frecuencia se predicen promotores en el intervalo -400,50. Con un breve comentario de los resultados es suficiente.



La gráfica se generó con el siguiente código de python:

https://github.com/carinapaola/Bioinformatica_estructural/blob/master/GraficaFrecuencias.ipynb

Se les ocurre una manera de validar sus resultados, y calcular la tasa de FPs, usando RSAT:matrix-scan?

Para usar matrix scan se requiere conocer el factor de transcripción para el cual se están buscando sitios de unión, pues se requiere una matriz PSSM como entrada. En regulonDB podríamos encontrar matrices PSSM. Para saber cuales matrices usar podríamos alinear las secuencias que tenemos con el genoma de E. coli K12 y buscar si hay matrices PSSM para estas regiones. Afortunadamente matrix-scan permite usar varias matrices PSSM a la vez, podríamos comparar nuestros resultados obtenidos al implementar el método de con los resultados de matrix-scan para saber cuántos falsos positivos tenemos, matrix-scan también podría darnos falsos positivos pero se pueden minimizar ajustando sus parámetros además que sus resultados son confiables porque las matrices PSSM se crean a partir de datos experimentales.