Guión para el trabajo autónomo:

Análisis de una secuencia anónima de DNA

Motivación

Frecuentemente, el resultado de un experimento viene dado por una secuencia o una región genómica. Por ejemplo, un estudio de asociación podría indicar que una región dada esta asociada a una enfermedad. Por lo tanto, el próximo paso sería caracterizar esta región, o lo que es lo mismo, anotar la secuencia que corresponde a esta región en detalle. Este análisis permitiría entender mejor la base molecular por la cual esta región se asocia con una enfermedad. De especial interés será averiguar si la secuencia contiene algún gen conocido. En caso afirmativo, será importante determinar su función, las rutas metabólicas en las que participa, los tejidos en los que se expresa y las posibles implicaciones que tiene en otras enfermedades. Aparte del análisis centrado en los genes, será también importante analizar los elementos reguladores como TFBSs (sitios de unión de los factores de transcripción), RNA no-codificante, islas CpG, la variación de secuencia (SNPs: polimorfismos de una sola base, CNVs: variación en el número de copias) y la variación epigenética (metilación del DNA). Otros análisis intresantes en este tipo de estudio podrían ser determinar el contenido en DNA repetido, el G+C de la secuencia o la composición de k-meros.

Por lo tanto, para el trabajo autónomo de esta asignatura el alumno tiene la tarea de caracterizar una secuencia anónima. Este trabajo simula un estudio de investigación "real".

Descripción

Al principio del curso se facilita al alumno una secuencia anónima de DNA con una longitud aproximada de 100 kb. La secuencia puede provenir tanto del genoma humano como de ratón. La secuencia se ha extraído de forma que contenga exactamente un gen. Además, se han escogido solo genes altamente conservados con lo que este gen estará también en otras especies. La secuencia abarca todo el gen, desde el inicio de transcripción (TSS) hasta el final de la transcripción (TES), añadiendo unos flancos a ambos extremos entre 5 y 10 kb.

La tarea del alumno será analizar y caracterizar esta secuencia, localizar la secuencia en el genoma, determinar el nombre del gen y recolectar la información relevante acerca de su función.

Nótese que las preguntas **muy importantes** se resaltan en el texto usando otro tipo de letra.

1: Análisis composicional

La composición de la secuencia permite en muchos casos obtener una primera idea acerca de la especie de que proviene, sobre los elementos funcionales que podría contener y sobre los procesos evolutivos. Para ello, **de especial interés serán las frecuencias de mono y di-nucleótidos y el contenido en G+C a lo largo de la secuencia**.

Para el análisis composicional, disponemos de una serie de programas de EMBOSS que se han presentado en clase.

- ¿Qué programa podemos usar para analizar las frecuencias de mono y dinucleótidos?
- ¿Cuál es el método más indicado para estimar la frecuencia esperada de los dinucleótidos?

- **¿Existen algunos dinucleótidos con frecuencias observadas muy distintas a las esperadas? ¿A qué mecanismos bioquímicos y evolutivos se debe este fenómeno?**
- Algunos programas permiten analizar la composición a lo largo de la secuencia. Estos programas suelen deslizar una ventana móvil a lo largo de la secuencia calculando en cada paso la característica deseada dentro de la ventana (G+C, #CpG, etc.). Estos métodos suelen ser altamente paramétricos incluyendo el tamaño de la ventana y la longitud del paso. ¿Qué efectos pueden tener estos parámetros sobre los resultados? ¿A qué regiones genómicas podrían corresponder los máximos locales en el contenido de G+C? ¿Esto nos puede dar una primera idea de en qué hebra (Watson o Crick) está localizado el gen que contiene la secuencia anónima?
- ¿Cuál es la diferencia entre un codón y un trinucleótido?

2: Contexto genómico

El objetivo aquí será situar la secuencia en su contexto genómico, es decir sus coordenadas (cromosoma, inicio y final) para un ensamblado concreto.

- ¿De qué especie proviene la secuencia?
- ¿Cuáles son las coordenadas cromosómicas de la secuencia? ¿**Qué relación hay entre las coordenadas y el ensamblado genómico?**
- ¿Por qué el programa BLAT detecta más de un alineamiento cuando mapeamos la secuencia anónima frente a un genoma de referencia?
- ¿Las coordenadas de la secuencia anónima cambian si elegimos otro ensamblado? ¿Por qué podría ocurrir esto?

3: Detección y caracterización del gen

Una vez obtenidas las coordenadas para un ensamblado concreto, podemos interrogar las bases de datos pertinentes para obtener la información acerca de la identidad del gen.

- ¿Por qué existen tantos nombres para un mismo gen? ¿Cuáles son las diferencias entre un identificador de gen, transcrito y proteína?
- ¿**Cuál es el nombre del gen que alberga la secuencia anónima y cuantos transcritos hay anotados en la base de datos de RefSeq gene**?
- ¿Cuáles son las coordenadas del gen? ¿Cuántos exones e intrones tiene?
- ¿Cuales son las longitudes de los flancos 5′ y 3′ UTR (UTR Untranslated Region), del mensajero y de los intrones (suma de todos los intrones)?
- ¿Cuál es la diferencia en el contenido en G+C entre el mensajero (mRNA) y la región codificadora del gen?
- ¿Cuál es la diferencia en G+C entre el gen y la secuencia anónima? ¿Si observamos una diferencia, que podría causarla?

4: Predicción de genes

Los genomas de la gran mayoría de las especies no están tan bien anotados como humanos y ratón. Por lo tanto, la predicción de genes es todavía importante. El objetivo de esta sección será usar uno de los métodos que hemos visto en clase para predecir genes dentro de la secuencia anónima. De especial importancia será la comparación de la predicción con la anotación obtenida en el apartado 3.

- ¿Por qué la predicción de genes es mucho más fácil en procariotas que en eucariotas?
- ¿En qué especies la detección de ORFs es especialmente relevante para la predicción de genes? ¿Por qué?
- ¿Qué diferencias se observan entre la predicción y la anotación en cuanto al número de genes y exones predichos?
- ¿Hay algún exón predicho perfectamente?

5: Detección de elementos genómicos

Existe un número alto de elementos genómicos funcionales y también DNA repetido como los transposones. El objetivo de este apartado será caracterizar la secuencia en cuanto a la existencia de TFBSs, DNA repetido e islas CpG.

- ¿Existe algún TFBS (sitio de unión de un factor de transcripción) anotado en la región promotora del gen que alberga la secuencia anónima?
- ¿Qué transposones, retro-transposones y repetido con LTR están en la secuencia anónima? ¿En qué regiones (intrones, exones, promotor) se localizan?
- ¿Contiene la secuencia anónima alguna isla CpG? ¿Qué parámetros o método se ha usado para detectar la isla?
- ¿Dónde se esperaría que se localice la isla CpG?
- ¿Qué relevancia tienen las islas CpG en la regulación de le expresión génica?
- ¿Qué propiedades composicionales tienen las islas CpG?
- ¿Qué programas conoce para la detección de islas CpG? ¿Cómo afectan los parámetros a la detección de las islas?

6: Análisis funcional del gen

Después de haber detectado el nombre del gen que contiene la secuencia anónima, el próximo paso será recolectar información acerca de las posibles funciones que tiene y de las rutas metabólicas en las que participa.

- ¿Qué funciones se han descrito para este gen?
- ¿En qué tejidos se expresa?
- ¿Se conoce la vinculación de este gen con alguna enfermedad?

- ¿El gen se encuentra regulado de forma post-transcripcional por un microRNA? ¿En caso afirmativo, que microRNAs podrían regular este gen?

7: Análisis evolutivo del gen

Finalmente, podemos usar el gen para hacer un análisis evolutivo.

- ¿Qué base de datos podemos usar para encontrar los genes homólogos?
- ¿En cuántas especies más se conoce este gen?
- ¿Un análisis filogenético con este gen confirma la filogenia establecida de las especies que contienen genes homólogos?