UOC Universitat Oberta de Catalunya

Universitat de Barcelona

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA Y BIOESTADÍSTICA

Genómica computacional

Enrique Blanco García y Jaume Sastre Tomàs

PEC2. Comparación de secuencias y predicción de genes	
Presentación y objetivos	Comparación de secuencias a nivel global y local. Búsqueda de motivos conservados en un conjunto de secuencias. Predicción computacional de un gen en una región del genoma humano.
Fecha y formato de entrega	**2 de Diciembre – 15 de Diciembre**
Criterios de corrección	Ejercicio 1 – 50% Ejercicio 2 – 50%

Presentación

Este ejercicio está diseñado para poner en práctica los conocimientos adquiridos sobre la comparación de secuencias y la anotación de genomas eucariotas. En particular, aquí nos concentraremos tanto en adquirir destreza en el alineamiento de secuencias con los programas BLAST, CLUSTAL y MEME, como en la identificación computacional de los genes en base a la predicción de sus componentes básicos: los exones. El estudiante deberá, en primer lugar, aprender a reconocer las diferencias entre distintas clases de alineamientos de secuencias y, posteriormente, efectuará la anotación de los genes de una secuencia genómica con GENEID, GENSCAN y FGENESH, evaluando después la calidad de estas predicciones mediante la comparación con anotaciones reales disponibles en bases de datos contrastadas. Junto con este enunciado, encontraréis una serie de ficheros necesarios para los distintos bloques temáticos del ejercicio. Es imprescindible la lectura de los módulos teóricos de comparación de secuencias y anotación de genomas.

Objetivos

Se pretende que los alumnos posean nociones básicas en la comparación de secuencias y la interpretación biológica de los resultados al acabar este trabajo. Por otro lado, se desea que el estudiante adquiera experiencia en el manejo de los programas de anotación genómica (predictores de genes, en este caso práctico), aprendiendo a tomar decisiones coherentes cuando existen varias soluciones alternativas. El alumno también empleará el navegador genómico UCSC para visualizar las predicciones y cuantificar las diferencias entre estas.

Formato

El informe final, como máximo, debe tener un tamaño de 20 páginas. Puede emplearse un apéndice para añadir el resto de informaciones. El nombre y apellidos del estudiante deben indicarse claramente en la primera página del informe. Es necesario incluir capturas de pantalla que expliquen el modo de alcanzar la solución de los ejercicios propuestos. La solución de esta PEC2 debe entregarse en formato PDF o Word (pueden adjuntarse aparte ficheros adicionales si fuera necesario).

Ejercicio 1. Estrategias de alineamiento [50%]

1. El programa CLUSTAL realiza alineamientos globales de dos o más secuencias. Conectaos al servidor implementado en el EBI para comparar la secuencia CDS del gen *TMEM106B* obtenida desde RefSeq (UCSC) para humano y ratón en la PEC1 anterior (hg38 y mm10, respectivamente).

http://www.ebi.ac.uk/Tools/msa/clustalo/

2. Repetid este mismo alineamiento global, utilizando ahora las respectivas proteínas de este gen en cada especie (que previamente debéis volver a recuperar de la entrada de RefSeq). Valorad el grado de homología entre estas dos secuencias.

3. El programa BLAST realiza alineamientos locales. Conectaos a BLAST, en el servidor principal del NCBI, para buscar qué versión de este programa debéis utilizar para alinear dos secuencias. Realizad ahora el alineamiento local de las dos regiones CDS del gen *TMEM106B*.

http://www.ncbi.nlm.nih.gov/blast/

4. Ahora utilizad el servidor de CLUSTAL para alinear globalmente la secuencia *genomicA.txt* y la secuencia *genomicB.txt* que encontraréis adjuntas a este enunciado.

5. Proceded ahora a efectuar el alineamiento local con BLAST de la secuencia genómica *genomicA.txt* y la secuencia *genomicB.txt* adjuntadas con el enunciado.

6. Comparad los resultados del alineamiento global y local en los dos casos anteriores (2 CDSs o las secuencias *genomicA.txt* y *genomicB.txt)*. Decidid cuál de los dos programas probados es más adecuado para cada caso en función de la estrategia empleada.

7. Unos investigadores que trabajan con el genoma del pollo (*chicken*) nos envían la secuencia adjunta *genomicC.txt*, pues sospechan que la forma ortóloga de nuestro gen *TMEM106B* está codificada en su interior. Decidid qué versión de BLAST debéis utilizar para validar esta hipótesis con la proteína humana (que tenéis de pasos previos), anotando su homóloga en esta región genómica de pollo. En caso de respuesta afirmativa, interpretad el grado de homología resultante entre ambas proteínas.

8. El programa MEME representa una familia alternativa de herramientas bioinformáticas para comparar secuencias. Definid en pocas palabras qué tipo de tarea realiza esta aplicación y cómo puede ser empleado dentro del área de estudio de la regulación génica mediante factores de transcripción:

http://meme-suite.org/
http://alternate.meme-suite.org

9. Vamos a estudiar la regulación transcripcional de nuestro gen *TMEM106B* a lo largo de la evolución. En primer lugar, empleando el navegador genómico de UCSC y las anotaciones de RefSeq, debéis extraer la región promotora del gen (seleccionad 5000 nucleótidos de longitud justo antes del inicio de transcripción del gen en cada especie) para estas especies: humano (hg38), ratón (mm10), rata (rn6) y pollo (galgal6).

10. En segundo lugar, emplead el programa MEME para comparar esas cuatro secuencias ortólogas. Buscamos los 10 mejores motivos que posean una longitud entre 5 y 15 pares de bases. Explorad qué función puede jugar el programa TOMTOM integrado dentro de la *suite* de programas MEME y efectuad una prueba con alguno de los motivos identificados.

Ejercicio 2. Anotación computacional de genes [50%]

Estamos colaborando con un laboratorio de biología molecular que sospecha que la secuencia *anonima.fa* codifica un gen humano. Este fragmento genómico está representado en el formato FASTA habitual con una cabecera inicial y la secuencia a continuación (adjunto al enunciado):

```
>human
GCCGCGGCGCCTTTGTGACGCCATCAGCCCGCGCGCCGCCGCCGCCGCCT
TCTGTGCAGTCGCGGCCCGGGCGGACGGTGGCTGGCTGCTCCGCAGCGCT
CGGCTGGCTGCAGCGGCACCGCGGGTTGCGCGGCCGGGGATGCTCCAGCG
GGCGCGATGGCCCCCGCCATGCAGCCGGCCGAGATCCAATTTGCCCAGCG
```

1. Deseamos conocer las coordenadas de los exones que constituyen el gen codificado en esta secuencia. Como primer paso de nuestro protocolo de anotación, debéis utilizar el programa GENEID para recuperar el mejor gen identificado computacionalmente en esta región del genoma humano:

GENEID:
http://genome.crg.es/geneid.html

2. Como segundo componente de nuestro *pipeline*, debéis emplear GENSCAN para recuperar el gen codificado internamente en esta secuencia humana:

GENSCAN:
http://genes.mit.edu/GENSCAN.html

3. Finalmente, como tercer componente del proceso, utilizad el programa FGENESH para identificar también la predicción de este sistema:

FGENESH:
http://www.softberry.com/

4. Para evaluar la coherencia de las predicciones obtenidas por cada programa, emplead CLUSTAL para comparar las proteínas reportadas por GENEID, GENSCAN y FGENESH. Realizad una primera interpretación de estos resultados en el contexto de este alineamiento global.

5. Finalmente, para comparar cuantitativamente los tres sistemas de predicción, rellenad la siguiente tabla con las coordenadas de todos los exones identificados dentro del mejor gen presentado por cada programa. Seleccionad dos de estos exones para realizar una búsqueda con BLASTP contra la base de datos completa de proteínas. Interpretad estos resultados para elaborar una primera anotación factible de este gen en función de estas predicciones:

	GENEID	GENSCAN	FGENESH
Exón 157-286	●	●	●
...			

6. Aprovechad BLAT para identificar en qué parte del genoma humano se encuentra *anonima.fa* (cromosoma, inicio, final, hebra). Verificad visualmente que el inicio y el final de nuestra secuencia encajan con la región correcta.

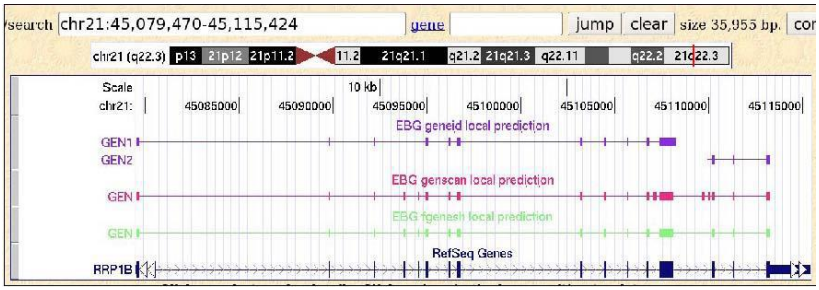
BLAT:
http://genome.ucsc.edu
http://genome.ucsc.edu/cgi-bin/hgBlat?command=start

7. Convertid manualmente nuestras predicciones de GENEID, GENSCAN y FGENESH en formato GFF para visualizarlas como Custom tracks en UCSC (será necesario adaptar las coordenadas de los exones para trasladarlos sobre el cromosoma 21):

Información para crear una Custom track:
http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#CustomTracks

Información sobre el formato GFF:
http://genome.ucsc.edu/FAQ/FAQformat.html#format3

Debéis lograr un resultado similar a la siguiente pantalla (en vuestro caso, cambiad EBG por las iniciales de vuestro nombre y apellidos):



8. Emplead el Table Browser de UCSC para calcular la correlación, dentro de la región genómica delimitada por la secuencia *anonima.fa*, entre las predicciones de (a) GENEID y GENSCAN, (b) GENEID y FGENESH, (c) GENSCAN y FGENESH. A continuación, repetid el mismo procedimiento para calcular la correlación entre cada predicción individual y el gen anotado por el consorcio RefSeq.

Table browser:
http://genome.ucsc.edu/cgi-bin/hgTables

9. Para acabar, efectuad con CLUSTAL el alineamiento múltiple global de las tres proteínas predichas por cada programa junto con la proteína real RRP1B. Analizad cuidadosamente cada sección de la proteína en busca de las mejores predicciones en ese fragmento. Con todas estas informaciones, decidid qué programa ha efectuado la mejor predicción.

10. El navegador genómico VISTA permite observar la conservación entre diversos genomas. Analizad la documentación existente sobre esta aplicación y averiguad el significado que tienen las gráficas y los colores empleados sobre cada alineamiento entre dos genomas. Posteriormente, seleccionad nuestro gen de estudio para analizar el grado de conservación que poseen los exones de éste. Razonad brevemente sobre cómo podríamos mejorar las predicciones iniciales servidas por GENEID, GENSCAN y FGENESH utilizando esta información sobre la conservación de secuencia en regiones funcionales.

VISTA:
http://pipeline.lbl.gov/