Instrucciones de la primera tarea de Bioinformática

J. Ignacio Lucas Lledó

21/3/2021

Durante la primera parte de la asignatura de Bioinformática habéis aprendido sobre algunas herramientas básicas, como los formatos más conocidos de texto plano, las bases de datos, los alineamientos y el BLAST. En las sesiones prácticas hemos usado R, con pequeñas incursiones en la línea de comandos, y siempre en un entorno de computación común, proporcionado por MyBinder. Este es el enlace al entorno de trabajo:

Siguiendo el estilo de las prácticas, la tarea que se os pide para evaluar esta primera parte de la asignatura consiste fundamentalmente en editar o adaptar un script previamente existente. No se trata, pues, de programar *ex novo*, sino de demostrar la comprensión de lo que está ya programado así como la capacidad de adaptarlo para un análisis similar pero con otros datos de partida.

El ejemplo de análisis que podéis utilizar como plantilla es el cuaderno `Ejemplo.ipynb`. En él hay descrito y ejecutado un análisis de la secuencia aminoacídica de la protoclorofílida reductasa C dependiente de luz de *Arabidopsis thaliana* (LPOR), que ya vimos en la práctica 7. En este ejemplo se usa **blastp** con la finalidad de explorar la distribución taxonómica de las proteínas homólogas a LPOR.

La tarea que os propongo es aplicar el mismo análisis a una proteína de partida diferente. Tenéis tres proteínas disponibles en sendos archivos FASTA: `CHRM1.fas`, `CHRNA3.fas` y `CHRNA7.fas`. Se trata de proteínas que participan en neuroreceptores de acetilcolina, y que según Viscardi *et al* (2021) aparecieron durante la evolución temprana de los animales, en el linaje del último antepasado común entre cordados y cnidarios. El apartado **Contexto evolutivo de las receptores de acetilcolina** os da algunas pistas más que pueden ayudar a redactar una pequeña introducción y algunas conclusiones de vuestro trabajo.

Objetivos

En la evaluación del trabajo valoraré los aspectos siguientes, en orden de importancia:

1. Los análisis són reproducibles.
2. Los análisis demuestran la utilidad de alguna herramienta o recurso estudiado en la asignatura.
3. De los resultados se extrae alguna conclusión correcta acerca del objeto de estudio.
4. La bibliografía incluye referencias a los algoritmos y recursos utilizados.
5. Tanto la redacción como el código son claros y coherentes.
6. La longitud del trabajo no debe ser excesiva.
7. El análisis incluye algún cálculo o proceso adicional no incluído en la plantilla.
8. La discusión propone qué otros análisis podrían ayudar a resolver las preguntas pendientes.

No es necesario hacer el trabajo sobre familias de receptores de acetilcolina. Si tienes interés en alguna otra familia proteica, también será aceptable. Igualmente, los análisis propuestos en el documento `Ejemplo.ipynb` no son más que una propuesta para facilitar el trabajo. Si refieres realizar un alineamiento, o cualquier otro análisis basado en las prácticas, tu trabajo será evaluado con los mismos criterios.

Contexto evolutivo de los receptores de acetilcolina

Recientemente, Viscardi *et al.* (2021) han realizado un extenso análsis sobre los orígenes evolutivos de las familias proteicas que en humanos participan en la neurotransmisión. De 321 genes humanos, a 83 se les atribuyen funciones específicamente neuronales. Curiosamente, la mayor parte de todos estos genes se originaron muy temprano en la evolución de los eucariotas, y muy pocas proteínas con función exclusivamente neuronal han aparecido en el linaje humano después de la divergencia con el linaje de los cnidarios.

Existen dos tipos principales de receptores de acetilcolina en las neuronas de vertebrados: muscarínicos y nicotínicos. La proteína humana CHRM1 (cuya secuencia puede encontrarse en el archivo `CHRM1.fas`) es un receptor muscarínico de acetilcolina. Las proteínas CHRNA3 y CHRNA7 (también disponibles en formato fasta) son dos subunidades parálogas de receptor nicotínico de acetilcolina. En el genoma humano hay unos cuantos genes que codifican subunidades de este tipo. Viscardi *et al.* (2021) alegan que estas proteínas aparecieron en el linaje del último ancestro común entre cordados y cnidarios. Es decir, no se conocen proteínas homólogas en ctenóforos, poríferos, placozoa, ni mucho menos en hongos ni en plantas.

Formato de entrega

El trabajo debe ser entragado a través del Aula Virtual al menos como cuaderno jupyter (archivo `.ipynb`). Si se utilizan archivos adicionales (por ejemplo, fasta) no incluídos en el ambiente de computación disponible en MyBinder, entonces tanto el cuaderno *jupyter* con el código y el texto como dichos archivos adiconales deben ser incluidos en un archivo .tar o .zip. En cualquier caso, no se aceptarán archivos en formatos cerrados o privativos.

Si para realizar tu trabajo hubieras necesitado instalar algun programa adicional, será necesario que documentes qué programa es y qué versión has instalado.

Información práctica

Uso del Jupyter Notebook dentro de Jupyter Lab

La interfaz de Jupyter Lab es muy intuitiva y la hemos estado practicando. Aún así, es muy posible que se nos olvide cómo realizar ciertas acciones, como por ejemplo: eliminar o añadir una celda de texto o de código, borrar de la memoria de trabajo todos los resultados de la ejecución anterior de una o más celdas de código, etc. En este enlace encontrarás toda la documentación sobre Jupyter Lab. Para cuestiones más relacionadas con el cuaderno (el documento `.ipynb`), puedes darle un vistazo a esta documentación.

Recursos sobre R

Hay mucha información online. Yo recomiendo estos dos enlaces:

- https://cran.r-project.org/
- http://swcarpentry.github.io/r-novice-inflammation/

Sobre BLAST

He incluído en el espacio de trabajo los archivos `blastp_help.txt` y `psiblast_help.txt`, con toda la ayuda de los programas `blastp` y `psiblast`. Ahí encontrarás fácilmente, por ejemplo, los campos de información disponibles para personalizar los resultados en formato de tabla.

Ambiente de computación

El ambiente de ejecución preparado en MyBinder incluye esta vez una instalación de la base de datos Swissprot de BLAST con toda la información necesaria para utilizar la información taxonómica de las secuencias. Debido a que `blast+` es un paquete ajeno a R, no resulta todavía posible preinstalar `blast+` en MyBinder. Es necesario hacerlo manualmente al iniciar la sesión, y posiblemente cada vez que el ordenador virtual regrese a su configuración inicial por inactividad. Para facilitarlo tenéis el script `preparar_ambiente.sh` y las instrucciones de cómo ejecutarlo desde R en la plantilla `Ejemplo.ipynb`.

Los paquetes de R siguientes también serán preinstalados y por tanto estarán disponibles en el entorno de trabajo, solamebnte por si alguien los quiere usar:

- **DECIPHER**: Para alineamientos. Véase la práctica del tema 5 o la documentación del paquete.
- **taxize**: Permite usar los identificadores taxonómicos de NCBI. https://taxize.dev.

Gestión de archivos remotos

Recuerda que el entorno MyBinder se ejectua remotamente: todo tu trabajo habrá desaparecido cuando se cierre la sesión en MyBinder, intencionada o accidentalmente, si no lo has descargado en tu propio ordenador. Descargar con frecuencia tu cuaderno jupyter de trabajo te protegerá de las posibles interrupciones en la conexión al servidor.

Una alternativa es instalar Jupyter Lab en tu propio ordenador. Aquí tienes bastante información al respecto. Si usas tu propia instalación de Jupyter Lab (o simplemente de Jupyter Notebook), necesitarás descargarte la base de datos swissprot, si quieres usarla.

Consultas

No dudes en contactar conmigo si surge algún problema en la realización de esta tarea.

Bibliografía

- Lucas Henriques Viscardi, Danilo Oliveira Imparato, Maria Cátira Bortolini, Rodrigo Juliani Siqueira Dalmolin, Ionotropic Receptors as a Driving Force behind Human Synapse Establishment, Molecular Biology and Evolution, Volume 38, Issue 3, March 2021, Pages 735–744, doi:10.1093/molbev/msaa252.