Práctica 1: El repositorio público GenBank

Elvira Mayordomo y Francisco Merino

15/Feb/2019

1. Introducción

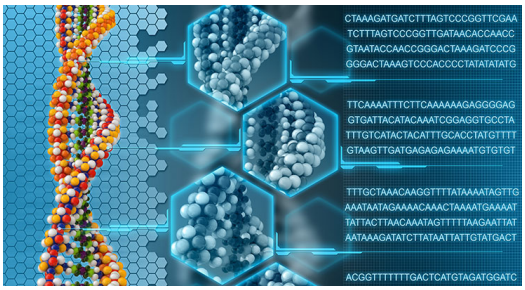
Esta primera sesión, de aproximadamente dos horas, tratará de la búsqueda de información biológica y se centrará en el repositorio publico GenBank. Al principio de la sesión dos estudiantes harán una introducción a GenBank y su uso.

2. Conceptos básicos

2.1. Información biológica

La información base que más se utiliza en bioinformática está reflejada en forma de secuencias de caracteres. Estos caracteres pueden pertenecer a distintos alfabetos, pero básicamente existen dos grandes grupos: los nucleótidos (4 caracteres) y los aminoácidos (20 caracteres). Las secuencias compuestas por el primer grupo son conocidas como DNA (o RNA), y las segundas como proteínas. Además, muchas de estas secuencias han sido ampliamente estudiadas y existe información asociada a las mismas, según cuál sea su origen, que suele ser utilizada para el diseño de algoritmos y sistemas bioinformáticos.

Existen diversos formatos estandarizados para almacenar secuencias biológicas, aunque los más extendidos son el formato FASTA y el formato PHYLIP. El primero consiste en una o más secuencias listadas de forma consecutiva, indicando el comienzo de cada una con el símbolo > seguido del identificador



http://staging2.hm-marketing-dev.com/CMSFiles/GFIA/a5/a5c1543f-92c8-40f9-a574-524999e92ce0.jpg?t=635248761813630000

de la secuencia. En la misma línea, tras un espacio en blanco, se puede añadir una descripción de la misma, aunque con frecuencia esta no suele ser incluida. En las siguientes líneas se encontrará la secuencia asociada al identificador, compuesta por una o más líneas con una longitud máxima de 80 caracteres.

```
> Seq1
AAACGATCGATCGAT
> Seq2
CGATCGCTCT-CGAT
> Seq3
CAG-TCACGATCATA
> Seq4
GCTAGCATC-CAGCT
> Seq5
TACGATCTAGCTAGT
```

```
 5 15
Seq1       AAACGATCGA TCGAT
Seq2       CGATCGCTCT C-GAT
Seq3       CAG-TCACGA TCATA
Seq4       GCTAGCAT-C CAGCT
Seq5       TACGATCTAG CTAGT
```

Diferencia entre un fichero FASTA y un fichero PHYLIP.

El formato PHYLIP es bastante diferente. En primer lugar, sólo se puede usar cuando todas las secuencias que se van a almacenar tienen longitud similar, en concreto cuando están alineadas (este concepto se explicará con detalle más adelante). En la primera línea del fichero se encuentran 2 números: el número de secuencias y su longitud máxima (separados por un espacio en blanco). En la siguiente línea, y con una longitud no superior a 10 caracteres, se encuentra el identificador de la primera secuencia y a continuación, separado por un espacio en blanco, los primeros 60 caracteres de la secuencia, en bloques de 10. En las líneas siguientes aparecen las demás secuencias que

componen el conjunto. Tras finalizar el listado, una línea en blanco indica el final del primer bloque, y a continuación aparece el segundo (si la longitud de alguna secuencia fuese superior a 60 caracteres). En esta ocasión los identificadores ya no quedan reflejados, asumiendo el mismo orden del primer bloque, y se encuentran los siguientes 60 caracteres (o menos) de cada secuencia. Esta estructura se repite hasta alcanzar el final de todas las secuencias. Esta configuración se menciona a menudo como PHYLIP-intercalado (*PHYLIP-interleaved*) y es la más usada. Existe una versión secuencial, más parecida a FASTA, donde tras el identificador de cada secuencias se encuentra la cadena que la compone de principio a fin.

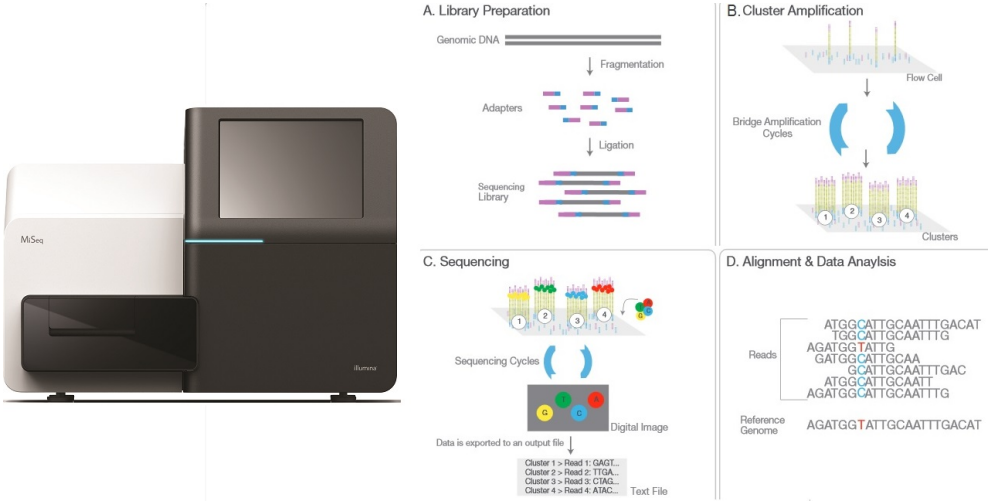
2.2. Secuenciación: de la biología a la informática

La secuenciación es el proceso por el cual una secuencia biológica se *digitaliza*, permitiendo así el uso de programas para su almacenaje y procesamiento. A grandes rasgos, la secuenciación consiste en generar muchas copias de la secuencia original, prepararlas para ser procesadas, leerlas (por distintos métodos), y, mediante diversos algoritmos, juntar los resultados para recomponer la secuencia original. Hay varios controles de calidad durante las diversas fases que aseguran una tasa de error muy pequeña, dado que resulta muy costoso repetir la secuenciación para asegurarnos de que la secuencia digital y la biológica son exactamente iguales.

2.3. Bases de datos biológicas: GenBank

Existen numerosas bases de datos creadas para almacenar secuencias biológicas y su información asociada. La más conocida, entre otros motivos por ser totalmente pública, es GenBank (`http://www.ncbi.nlm.nih.gov/genbank/`). Con el abaratamiento de los costes de secuenciación, tanto económicos como temporales, el almacenamiento de datos biológicos supone un gran problema para la comunicad bioinformática. Este ha sido uno de los campos con mayor impacto en los problemas que se asocian al término *Big Data*.

La primera tarea del alumno consistirá en acceder a GenBank y empezar a familiarizarse con el motor de búsqueda, la forma en que presenta GenBank



Combinación de: `http://blogs-images.forbes.com/matthewherper/files/2011/01/Illumina_MiSeqTM_CopyRight2011.jpg` e imagen en `http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf`

los resultados, la información detallada de cada secuencia y los distintos formatos en los que se pueden bajar las secuencias.

Nota: Se aconseja comenzar leyendo el link "help" que aparece en la esquina superior derecha (Entrez Sequences Quick Start) y dentro del mismo help buscar al menos "Entrez Nucleotide and Entrez Protein FAQs" y leer la pregunta 6 (fields in a Genbank record).

Tareas a realizar:

1. Buscar todas las secuencias almacenadas en GenBank procedentes de seres humanos.
2. A partir de la búsqueda anterior, filtrar y quedarse sólo con aquellas pertenecientes al cromosoma X.
3. Buscar todas las secuencias de DNA de referencia de mamíferos. Descargar las 10 primeras en formato FASTA.
4. Buscar todas las secuencias de proteínas pertenecientes al cromosoma Y de primates.
5. Buscar todas las secuencias de proteínas que se han incluido en lo que llevamos de año.

6. Buscar todas las secuencias de DNA mitocondrial humano con una longitud que difiera como mucho en 1 a la de la secuencia de referencia (16569bp). Descargar las 10 primeras en formato GENBANK

2.4. Caso práctico: búsqueda de secuencias conectadas con un caso criminal

El siguiente artículo

> Metzker M.L., Mindell D.P., Liu X.M., et al. (2002). Molecular evidence of HIV-1 transmission in a criminal case. *Proc Natl Acad Sci U S A.* 99(22):14292-7.

presenta un interesante caso criminal resuelto utilizando una buena dosis de bioinformática. En esta primera práctica debe leerse el resumen del artículo y encontrar y descargar las secuencias biológicas utilizadas.

2.5. Entrega

Debe entregarse una breve memoria (formato preferido PDF) sobre esta práctica incluyendo las consultas realizadas en GenBank así como los comandos para descargar las secuencias (no deben entregarse dichas secuencias pero sí conservarse para futuras prácticas). Debe enviarse por correo electrónico a `elvira@unizar.es` antes de la siguiente sesión (27 de febrero).

Agradecimiento

Jorge Álvarez Jarreta colaboró en la primera versión de esta práctica.