Práctica 1: Introducción a la bioinformática

Elvira Mayordomo y Jorge Álvarez

22/Feb/2017

1. Introducción

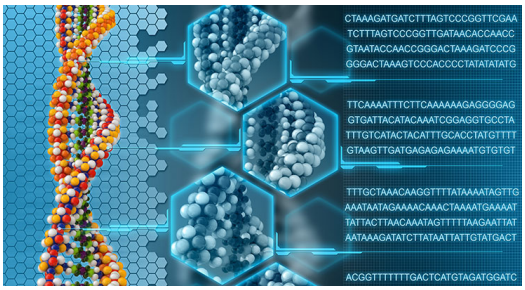
En las dos primeras sesiones de prácticas se va a repasar y profundizar en los conocimientos que se han impartido en teoría. Para ello se va a trabajar en los procesos de secuenciación, búsqueda de información biológica y alineamiento. Además, se utilizarán los conocimientos adquiridos en un caso real: estudio del DNA mitocondrial. La duración aproximada de la cada una de las sesiones será de 2 horas.

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 a las segundas, 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.



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

El primero consiste en una o más secuencias listadas de forma consecutiva, indicando el comienzo de cada una con el símbolo > seguido del identificador de la secuencia. En la misma línea, tras un espacio en blanco, se puede añadir una descripción de la misma, aunque esta no suele ser incluida con frecuencia. 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 va a almacenar tienen la misma longitud, es decir, 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, sepa-

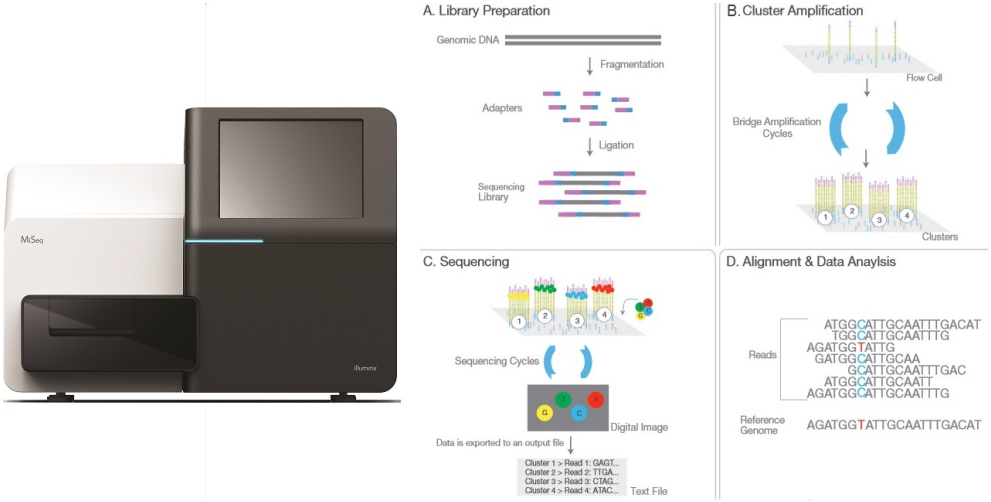
rado por un espacio en blanco, los primeros 60 caracteres de la secuencia, en bloques de 10. En las líneas siguientes aparecen las demás secuencias que componen el conjunto. Tras finalizar el listado, una línea en blanco indica el final del primer bloque, y a continuación aparece el segundo (si la longitud 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 el más usado. Existe una versión secuencial, más parecido a FASTA, donde tras el identificar 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 (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 (en algunos casos). 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*.



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`

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 los resultados, la información detallada de cada secuencia, 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 recomendadas:

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. Alineamiento

Como se ha podido observar al explorar GenBank, existen diferencias en la longitud de la misma secuencia de distintas especies. Esto es debido, principalmente, a las mutaciones que han sufrido a lo largo de los años, base de la evolución. Las mutaciones se suelen clasificar en 3 tipos: substitución, inserción y deleción, siendo más frecuentes las dos últimas (conocidas más comúnmente como *indels*). Por supuesto, nunca se puede descartar que estos errores de longitud sean debidos a un fallo en el proceso de secuenciación, pero se asume la buena praxis en esta fase.

Hay muchos estudios biológicos que requieren que el conjunto de secuencias con el que van a trabajar esté alineado, es decir, que todas las secuencias tengan la misma longitud. Para ello, se introduce en el alfabeto de la secuencia un nuevo elemento: el gap (`-`). Este símbolo se interpreta como posiciones que deberían existir y han sido eliminadas por una mutación, o como ajuste por inserciones en otras secuencias. El proceso de obtener el alineamiento óptimo tiene un coste asociado de $O(m^n)$, donde m es la longitud de las secuencias y n el número total de las mismas. Este coste es extremadamente elevado, incluso en los casos donde m y n son pequeños (dentro de los parámetros de datos reales biológicos). Por ello, las herramientas y algoritmos de alineamiento aplican heurísticas para resolver este problema.

La siguiente tarea por parte del alumno será utilizar dos herramientas muy extendidas para generar alineamientos, Mafft (`http://mafft.cbrc.jp/alignment/software/`) y Clustal Omega (`http://clustal.org/omega/`), aplicando distintas configuraciones a conjuntos de las secuencias descargadas en la tarea anterior. La idea es familiarizarse con este proceso y evaluar cómo los costes temporales varían según cambian los tamaños de entrada, así como las variaciones en los alineamientos resultantes según la configuración y herramienta utilizada.

Tareas recomendadas:

1. Escoger 10 secuencias de cada uno de los paquetes descargados en la

tarea anterior y alinearlas usando las dos herramientas en su configuración automática.

2. Estudiar algunas de las diferencias obtenidas en los alineamientos anteriores.
3. Descargar un pequeño conjunto de secuencias (no más de 10) de proteínas y repetir los puntos 1 y 2.
4. Buscar alternativas de alineamiento para mejorar la eficiencia en tiempo y/o memoria de los procesos anteriores.