Introducción al uso de herramientas bioinformáticas para el análisis de secuencias de ADN

Esther Menéndez Gutiérrez, Raúl Rivas González
{esthermenendez, raulrg}@usal.es

Departamento de Microbiología y Genética. Universidad de Salamanca

Palabras Clave: educación bioinformática, software gratuito, ADN, BioEdit, secuenciación, NCBI.

Resumen

Con el fin de introducir a los alumnos y/o profesionales universitarios en el campo de la bioinformática, hemos propuesto un curso formativo de iniciación en el uso y manejo básico, de una forma sencilla y aplicada, de herramientas bioinformáticas gratuitas y disponibles on-line. Este curso rápido y dinámico se basa en el empleo de las bases de datos y en la edición y manipulación de secuencias nucleotídicas obtenidas directamente del secuenciador para compararlas con las secuencias disponibles depositadas en las bases de datos. Además, introducimos el uso de alineamientos múltiples de secuencias para el ensamblaje de secuencias y para el diseño de primers a partir de bases conservadas

Introducción

El crecimiento de la industria de la biotecnología en los últimos años no tiene precedentes, así como los avances en modelado molecular, la caracterización de enfermedades, el descubrimiento farmacéutico, el cuidado de la salud clínica, la medicina forense y la agricultura. Estos avances impactan de forma fundamental en temas económicos y sociales en todo el mundo. No siempre ha sido así, ya que hace años las actividades de investigación y descubrimiento en las ciencias de la vida estaban limitadas normalmente a un solo gen o proteína. Sin embargo, el desarrollo de sistemas computacionales y de información (integrado con la biotecnología) ha facilitado un cambio hacia un cribado de alto rendimiento (miles de muestras por día) y hacia sistemas de detección de alto contenido (miles de puntos de datos por muestra). Este cambio ha

sido debido principalmente al desarrollo de la bioinformática, que ha sido el factor clave sobre el que se ha apoyado el sistema para desarrollarse y expandirse.

Podemos definir la Bioinformática como la aplicación de tecnología informática para la gestión y el análisis de datos biológicos. Por otra parte, El Instituto Nacional de Salud (NIH) de Estados Unidos utiliza una definición más completa y define a la bioinformática como la investigación, el desarrollo o aplicación de herramientas computacionales y enfoques informáticos para ampliar el uso de datos médicos, de salud o de comportamiento biológico, incluidas las competencias de adquirir, almacenar, organizar, archivar, analizar o visualizar estos datos (Instituto Nacional de la Salud, 2000). Por lo tanto, no podemos obviar que la bioinformática se ha convertido en una disciplina excepcional, ya que permite a los científicos descifrar y gestionar las enormes cantidades de datos que se generan a diario o que ya están disponibles. En este sentido, Allen (1) informó de que el objetivo final de las investigaciones biológicas, es el desarrollo del conocimiento que permitirá enfoques predictivos a todas las ciencias de la vida. Así, las técnicas de adquisición de datos y la gestión y análisis bioinformáticos son claves, ya que proporcionan una lente a través del la cual visualizar un conjunto de datos a gran escala.

Desde nuestro punto de vista, el objetivo de la bioinformática es presentar una representación completa de un organismo, para lo cual es necesario un avance computacional que prediga sistemas de gran complejidad, tales como los procesos de interacción celular de todo el organismo. Según Degrave et al. (3), una combinación fortuita de factores hacen que la bioinformática tenga un especial interés para los investigadores en todo el mundo, ya que ha facilitado el desarrollo de numerosos enfoques en múltiples áreas de investigación, que incluyen la detección de enfermedades, control y diagnóstico, descubrimiento y desarrollo de fármacos, mapeo genético, epidemiología o modelado de imágenes biomédicas y de los ecosistemas entre otros.

En consecuencia, existe una necesidad obvia de suministrar a los estudiantes de bio-ciencias de ciertas habilidades bioinformáticas genéricas que les permitan adquirir unas competencias básicas. Por esta razón, en el mes de febrero de 2014 organizamos un curso de introducción al uso de herramientas sencillas para el análisis de secuencias de ADN ya que, el acercamiento contemporáneo a la resolución de problemas biológicos incluye el manejo de métodos y programas para el análisis de un largo número de genes y proteínas simultáneamente. El incremento exponencial de genomas completos secuenciados, ha revalorizado el alcance de la bioinformática y por ello, pretendemos proveer a los alumnos del Grado en Farmacia así como a cualquier otra persona que lo requiera, de métodos conceptuales y prácticos para el análisis rápido y eficaz de secuencias obtenidas de una entidad biológica tan importante como es el ADN.

El objetivo principal que pretendemos conseguir es dotar a los alumnos de conocimientos en conceptos básicos de bioinformática, los cuales les permitan manejar las distintas herramientas bioinformáticas que están a su disposición, tanto online como en software informático. Además, perseguimos que al término de este curso los alumnos sean capaces de realizar búsquedas en bases de datos, imprescindibles en el desarrollo de cualquier investigación científica.

Contenidos y Metodología

Para alcanzar los objetivos planteados, el desarrollo del curso se ha distribuido en una serie de bloques temáticos, comenzando por las nociones básicas en bioinformática y terminando con el uso avanzado de algunas aplicaciones y programas propuestos de acceso libre y gratuito. Los bloques temáticos se detallan a continuación: Bloque 1. Introducción a la bioinformática; Bloque 2. Formatos de secuencias nucleotídicas. Edición de secuencias; Bloque 3. Bases de datos. Comparación de secuencias; Bloque 4. Alineamiento múltiple de secuencias nucleotídicas; Bloque 5. Diseño de cebadores/ oligos/primers sobre una secuencia conocida o sobre una secuencia consenso.

Dentro de estos bloques temáticos, el alumno tiene la oportunidad de realizar una serie de ejercicios prácticos propuestos específicamente para cada apartado, siguiendo una lógica similar a la de los análisis bioinformáticos llevados a cabo a nivel profesional, aunque siempre adaptados a las necesidades del alumno. Dentro de cada bloque temático se desarrollan una serie de contenidos específicos.

En el primer bloque temático mostramos una visión sencilla y resumida de cómo surgió la bioinformática respondiendo a las necesidades de la ciencia actual, además de una serie de conceptos biológicos, como el dogma de la biología molecular o el fundamento de la secuenciación automática, creando el contexto donde se van a englobar los análisis bioinformáticos posteriores.

La bioinformática se aplica actualmente en casi todos los campos científicos. Hoy en día, gracias a ella se pueden procesar las ingentes cantidades de datos generados por las técnicas de secuenciación, tanto secuenciación automática como pirosecuenciación. Los datos obtenidos en secuenciación están depositados y disponibles en las bases de datos que son públicas y accesibles. Durante el curso mostramos como ha aumentado dicha información depositada en las principales bases de datos a nivel mundial y cómo ha aumentado también el número de usuarios de dichas bases. Por ejemplo, el NCBI tiene actualmente una media de 2,5 millones de usuarios al día. Este aumento ha ido en consonancia con el aumento exponencial del número de secuencias depositadas en GeneBank en las últimas dos décadas.

Los datos generados y depositados en estas bases de datos corresponden a secuencias de nucleótidos y de aminoácidos, según sean ADN o proteínas. Este curso se ha dedicado por completo al manejo y edición de secuencias nucleotídicas, que será abordado en el siguiente bloque.

En el bloque número dos se presentan al alumno las formas en las que podemos representar las secuencias nucleotídicas, que son sucesiones de letras representando la estructura primaria de una molécula real o hipotética de ADN. Estas secuencias se pueden representar en varios formatos como son texto plano, en formato GeneBank, formato EMBL… y en formato FASTA (Figura 1), que es el formato que va a ser utilizado a lo largo de este curso por ser uno de los más habituales.

```
>gi|30578071|emb|AJ561043.1| Rhizobium leguminosarum bv. trifolii celc2 gene for cellulase
C2

ATGAGGCGGTGGCGCGCGCTTCTGCTGGCGGCCTCTGTCGCGGTTGCACCGGGCCTGCCGGCTACCGCGCAGCAGGCGATGATTAATGCCGA
CGCCTGGTCGGCCTACAAGGCGAAGTTTCTCGATCCGAGCGGCCGCATCGTTGACAACGGCAACGGCAACATCAGCCACAGCGAAGGGCAGG
GCTACGGCCTGCTGCTCGCCTATCTCTCGGCAAGCCCCGCCGATTTCGAGCAGATCTGGTATTTTACCCGCACCGAGCTGCTGCTGCGCGAC
GACGGCCTGGCCGTTTGGAAATGGGATCCGAACGTCAAGCCGCACGTGGCCGACACCAACAATGCCACCGACGGCGACATGCTGATCGCCTA
TGCTTTGGCGCTTGCCGGCACGGCATGGAAACGTGAAGATTATATCCTCGCTGCCTCCCGCATGGCGCAGGCGCTGCTTGCCGAAACCGTCG
GCAGCTCGCAGGGCCGCACCTTGCTGATGCCGGGAACCGAAGGGTTTACCGGCAGCGACCGCGACGATGGTCCGGTCGTCAACCCGTCCTAC
TGGATTTATGAGGCGATCCCGGTGATGGCAGCGCTCGCGCCGTCGGATGCCTGGAAAAAACTGTCGGACGATGGCGTGGAACTGTTGAAGAC
GATGCAGTTCGGCCCGCGCAAGCTTCCTGCCGAATGGGTGAGCCTGCACGACAAGCCGCGGCCGGCAGAGGGGTTCGACGCCGAATTCGGCT
ACAACGCCATCCGCATCCCGCTATATCTCGCCCGCGGCGGCATCACCGATAAGGCACTGCTCGTCCGCCTGCAAAAGGGGATGTCGCAAGAC
GGCGTTCCCGCCACGATCGATCTGACCACCGGCCGGCCGAAGACCGTGCTGTCGGACCCCGGTTATCGAATTGTTAACGATGTTGTGGCCTG
TGTTGTCGATGGGACCAGGCTGCCGAGTTCGGCGCTGCAGTTTGCCCCTGCGCTCTATTATCCGTCCACCCTTCAACTGCTGGGGCTGGCCT
ATATCGGGGAGAAGCATCCGGAGTGTCTGTGA
```

Figura 1. Secuencia nucleotídica en formato FASTA (Fuente: NCBI).

Dentro de este bloque también mostramos a los alumnos los diferentes softwares y/o aplicaciones que necesitan para editar las secuencias que habrán de ser obtenidas mediante secuenciación. Existen muchos programas bioinformáticos de edición de secuencias que están disponibles *on line* y gratuitamente para su utilización por cualquier tipo de usuario, algunos ejemplos son *Chromas* (Technelysium Pty Ltd), *4Peaks* (Nucleobytes Inc.) y *BioEdit* (5). Este último software gratuito es con el que se propone trabajar a los alumnos de este curso. Además, se propone el uso de otros tipos de software, más completos pero de pago como son los paquetes *DNAstar* y *Geneious*.

Con el programa *BioEdit* se abren las secuencias recibidas, que deben estar en formato FASTA. El programa permite abrir los cromatogramas correspondientes a cada secuencia para su edición y corrección de posibles errores cometidos en la lectura. Como apoyo, se introduce el concepto de cromatograma y las distintas posibilidades que se pueden encontrar, siempre apoyados con ejemplos prácticos reales. Los alumnos realizan un análisis visual de los cromatogramas que sirve para descubrir fácilmente algunos de los problemas comunes que pueden tener las secuencias. Aprenden a diferenciar cuando un cromatograma es de buena calidad (muestra picos únicos y separados y tiene poco

ruido de fondo) de cromatogramas que tienen una serie de irregularidades y con ello los posibles problemas en la secuencia que ha originado este tipo de cromatogramas, como por ejemplo una mala calidad del ADN secuenciado o la presencia de contaminantes.

En el tercer bloque temático, con las secuencias editadas obtenidas en la práctica anterior se procede a enfrentarlas a algunas de las bases de datos disponibles. Estas bases de datos son las siguientes: DNA Database of Japan (DDBJ) (http://www.ddbj.nig.ac.jp/), del Instituto Nacional de Genética de Japón; EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/) del European Bioinformatics Institute (EBI) y GenBank (http://www.ncbi.nih.gov/) del National Center for Biotechnology Information (NCBI). Además, existen muchas otras bases de datos propias de instituciones, algunas de ellas específicas para proteínas, para organismos, etc. De entre estas últimas, EzBioCloud para organismos procariotas, TAIR para *Arabidopsis* y SGD para *Saccharomyces cerevisiae*, son algunas de las más utilizadas.

En este curso, vamos a centrarnos en GeneBank, que es la colección anotada de secuencias del NCBI, que a su vez contiene otras bases de datos como PubMed, Gene, EST, SNP, Structure y su recurso "estrella", BLAST (2).

También en este bloque, se introduce a los alumnos del curso práctico a BLAST (Basic Local Alingment Search Tool), que es la herramienta más importante, fiable y flexible en estudios bioinformáticos que nos permite seleccionar una secuencia (query) y realizar alineamientos de pares de secuencias con todas las secuencias de la base de datos entera (target). En este punto los alumnos están capacitados para enfrentar la secuencia editada por ellos mismos y, con los parámetros adecuados poder también por ellos mismos, interpretar los datos ya que, la aplicación nos devuelve los alineamientos más relacionados con la secuencia dada.

Con el conocimiento obtenido hasta este punto, los alumnos pueden trabajar con secuencias nucleotídicas de una manera autónoma. Aún así los autores de este curso/trabajo creemos necesario ofrecer al alumno otras herramientas para proseguir con los análisis bioinformáticos. Por ello, en los bloques cuarto y quinto utilizamos una serie de programas online con el fin de llegar un paso más allá en las posibilidades que nos brinda la bioinformática ya que, en condiciones de trabajo real los investigadores disponen de multitud de secuencias con las que trabajar a la misma vez y que persiguen finalidades diferentes como por ejemplo, ensamblaje de genes y genomas o el diseño de primers a partir de secuencias conservadas.

En este bloque los alumnos realizan alineamientos múltiples y diseño de primers en un supuesto práctico. Al inicio de este bloque, se ponen en conocimiento de los participantes una serie de herramientas online adecuadas para la realización de los

alineamientos múltiples de secuencias, además de proporcionarles la opción de realizarlos mediante el software BioEdit. El supuesto práctico se realiza utilizando la herramienta MUSCLE (4) (Figura 2), disponible online entre los recursos que nos ofrece el Laboratorio Europeo de Biología Molecular (EMBL).

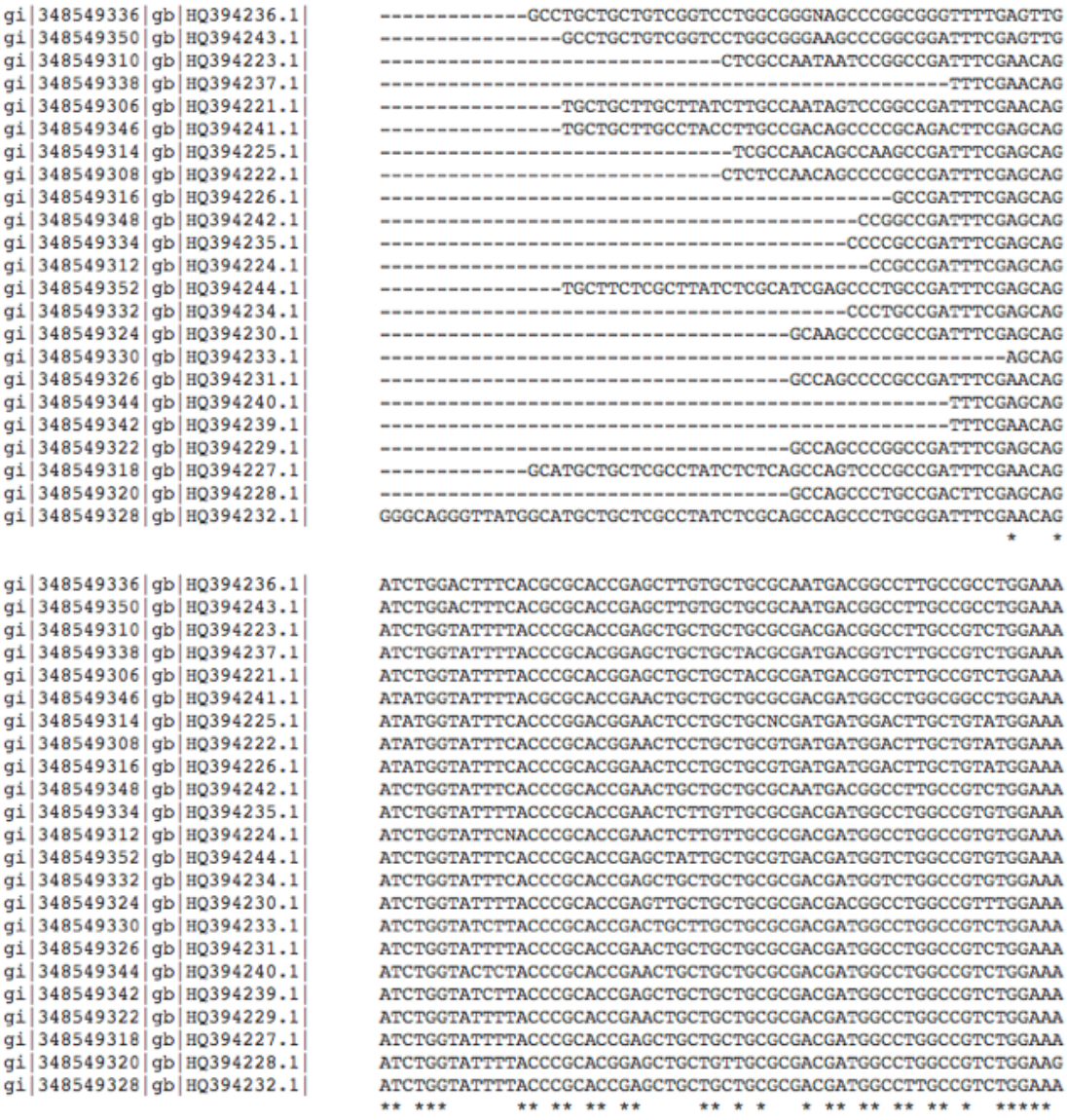

Figura 2. Alineamiento múltiple de secuencias con la aplicación onlie MUSCLE (www.ebi.ac.uk/Tools/msa/muscle/)

Además, si el alumno decide comparar con otro programa, pudo utilizar el programa BioEdit para realizar los alineamientos múltiples, la herramienta que viene incluida en el programa es ClustalW, formato en el que la aplicación MUSCLE también va a ofrecer los datos resultantes. Los alumnos realizan ejemplos prácticos de ensamblaje de secuencias del gen ribosómico 16S, marcador taxonómico en bacterias, muy sencillo de

realizar y entender. También realizaron alineamientos múltiples de secuencias para ver regiones conservadas de genes pertenecientes, por ejemplo, a especies de un mismo género, familia o incluso de un *Phyllum*, lo que es muy útil cuando queremos crear primers universales que es el último apartado que vamos a introducir en este curso.

Para un buen diseño de cebadores, iniciadores de la reacción que lleva a cabo la polimerasa en la PCR, se deben tener en consideración distintos aspectos que ponemos en conocimietno de los alumnos como son, la longitud de esos cebadores (entre 18 y 24 bases), un alto contenido en G+C y temperaturas de anillamiento cercanas entre los pares de cebadores, además de evitar regiones con potencialidad para formar estructuras secundarias internas.

Por último, ponemos a disponibilidad del alumno diversos programas de diseño de cebadores como son *Primer3* (6, 7) y algunos otros que las distintas marcas de biología molecular ponen a disposición de los usuarios.

Resultados

El presente curso de bioinformática básica ha tenido como objetivo principal iniciar en el manejo de secuencias de ADN a alumnos y profesionales de diferentes áreas científicas, en especial, a los pertenecientes a la Facultad de Farmacia, que carezcan de conocimientos previos en bioinformática. Además, el curso se engloba dentro del Proyecto de Innovación Docente EducaFarma 2.0 en el que se pretende ofrecer cursos/talleres o seminarios prácticos impartidos por profesores del centro con los propios recursos del centro.

Para tener conocimiento de la percepción de los alumnos que reciben este curso práctico así como de información adicional que nos permita mejorar la impartición de este tipo de cursos, se realizó una encuesta de satisfacción en la que se preguntó a los alumnos sobre su formación, nivel de satisfacción en distintos aspectos como son la calidad de los ponentes y la satisfacción general con el curso, entre otros.

En este sentido, este curso/taller, dirigido sobre todo a los profesionales y alumnos de la Facultad de Farmacia de la Universidad de Salamanca, tuvo participación también de personas pertenecientes a otras Facultades, como la Facultad de Ciencias Agrarias y Ambientales o la Facultad de Biología. De entre los alumnos un 71% son estudiantes de Grado, un 14% estudiantes de posgrado y un 2% de otra procedencia. Además, han asistido mayoritariamente mujeres, un 71% frente a un 29% de hombres. La difusión ha funcionado a la perfección, sobre todo por la web de la Facultad de Farmacia y Eventum (28%), aunque parece ser que el boca a boca ha sido la manera de difusión más efectiva (36%).

En cuanto a la calidad del curso y de los ponentes, y utilizando una escala del 1 al 5, siendo 1 muy malo y 5 excelente, el 100% de los alumnos evaluaron el curso como muy bueno y/o excelente. Cabe destacar que un 57% de los alumnos consideran que las herramientas de las que se ofrece el conocimiento en este curso son excelentes y de mucha utilidad. Además un 64% cree que la organización del curso ha sido muy buena y un 21% excelente.

En general la satisfacción global con el curso ha sido muy buena (en una escala del 1 al 10, el 100% de los alumnos están entre 8 y 10 puntos), aunque la duración del curso es lo que origina mayor discordancia entre el alumnado, incluso un 7% esta en desacuerdo, opinando que se debería dividir el curso en varias sesiones y aumentar los contenidos, incluyendo otros tipos de análisis e incluso ahondando en los contenidos que ya se ofrecen.

Conclusión

Como conclusión general, el curso/taller propuesto ha sido un éxito de acogida ya que se cubrieron todas las plazas disponibles en un corto periodo de tiempo e incluso se formó una pequeña lista de espera de interesados. Además, aunque se puedan realizar muchas mejoras de cara a un futuro o próximas ediciones del curso, la satisfacción tanto del alumnado como de los ponentes ha sido excelente por lo que existe una relación directa entre los objetivos perseguidos y la evaluación obtenida por lo que consideramos que el modelo utilizado es claramente positivo y apto para utilizarse en futuras ediciones.

Referencias

1. Allen, G.K. Bioinformatics: new technology models for research, education, and services. Educause Centre for Applied Research Bulletin. 2005; 8, 1–9.
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403-10.
3. Degrave,W., Leite,L., Huynh,C.H. FIOCRUZ distance-learning website (www.dbbm.fiocruz.br/helpdesk/). Oswaldo Cruz Foundation, Oswaldo Cruz Institute, 2001. Dept. of Biochemistry and Molecular Biology, Río de Janeiro, Brazil.
4. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 2004; 32 (5): 1792-1797.
5. Hall, T.A. BIOEDIT: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series, 1999; 41: 95-98.
6. Koressaar T and Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics 2007; 23(10):1289-91.
7. National Institute of Health, 2000. NIH Working Definition of Bioinformatics and Computational Biology <http://www.bisti.nih.gov/CompuBioDef.pdf>.
8. Untergrasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3 - new capabilities and interfaces. Nucleic Acids Research 2012; 40(15):e115.