BIOINFORMÁTICA PARA PRINCIPIANTES I: ANÁLISIS DE SECUENCIAS NUCLEOTÍDICAS (DNA)

Esther Menéndez Gutiérrez, Raúl Rivas González

Departamento de Microbiología y Genética
Universidad de Salamanca.

Palabras clave: educación, software gratuito, ADN, BioEdit, secuenciación, NCBI, FASTA.

Resumen

Por segundo año consecutivo y con el fin de introducir a los alumnos y/o profesionales universitarios en el novedoso campo de la bioinformática, se ha realizado un curso formativo de iniciación en el uso y manejo básico de datos y herramientas bioinformáticas gratuitas, introduciendo a los alumnos en este campo de una forma sencilla y aplicada mediante el uso de ejemplos y casos prácticos. Este curso se basa en el empleo de las bases de datos, además de 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, se introducen conceptos y usos adicionales, como el uso de alineamientos múltiples de secuencias para el ensamblaje de secuencias y para el diseño de primers a partir de bases conservadas. Todo ello ha dado lugar a un curso dinámico y muy práctico, en el que los alumnos adquieren las competencias básicas necesarias para trabajar en este campo.

Introducción

En los últimos años la industria biotecnológica, así como los avances en modelado molecular, la caracterización de enfermedades, el descubrimiento farmacéutico, el cuidado de la salud clínica, la medicina forense y otras muchas áreas biosanitarias, han demostrado un continuo desarrollo, avanzando a pasos agigantados en un entorno global. Estos avances han impactado en la economía global y en la sociedad, aunque no siempre ha sido así, ya que hasta hace pocos años existían limitaciones en cuestiones de investigación y desarrollo. Sin embargo, la llegada de nuevas tecnologías, como la aplicación de las ciencias computacionales a la biología y la biotecnología, ha facilitado el rápido desarrollo de este campo. Principalmente el desarrollo de la bioinformática, ha sido el factor clave para estos avances.

La bioinformática se define como la aplicación de la informática para la gestión y el análisis de datos biológicos disponibles. Una definición más concreta y completa es la que encontramos en el Instituto Nacional de Salud de Estados Unidos (NIH), ya que dice que la bioinformática se define como la investigación, el desarrollo o aplicación de herramientas computacionales y enfoques informáticos para ampliar el uso de datos de índole biológica, incluidas las competencias de adquirir, almacenar, organizar, archivar, analizar o visualizar estos datos. Es indiscutible que la bioinformática se ha convertido en una disciplina imprescindible, ya que permite descifrar y gestionar las enormes

cantidades de datos generados a diario o que ya se encuentran disponibles. En el año 2004 se publicó un informe en el que se expone el objetivo final de las investigaciones biológicas, que es el desarrollo del conocimiento que permitirá dar un enfoque predictivo a todas y cada una de las ciencias de la vida (Allen, 2004). De esta forma, las técnicas de adquisición de datos y la gestión y análisis bioinformáticos se perfilan como elementos claves, ya que ayudan a visualizar un conjunto de datos a gran escala.

Por tanto, estamos de acuerdo con que uno de los objetivos de la bioinformática es obtener 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 y colaboradores, una combinación fortuita de factores hacen que la bioinformática tenga un especial interés para los investigadores en todo el mundo, ya que facilita el desarrollo de múltiples áreas de investigación, como la detección temprana de enfermedades, control y diagnóstico de enfermedades, descubrimiento y desarrollo de nuevos y antiguos fármacos, epidemiología y/o modelado de imágenes, entre otros (Degrave et al., 2001).

Debido a ello, existe la necesidad de suministrar y/o actualizar conocimientos en bioinformática general a los estudiantes de bio-ciencias. Por esta razón, creímos necesario volver a organizar, en febrero de 2015, el curso de introducción al análisis de secuencias de ADN, incluyendo el manejo de métodos y programas para el análisis de un largo número de secuencias génicas de una manera simultánea. El incremento exponencial del número de genomas completos secuenciados, ha revalorizado el alcance de la bioinformática y por ello, este curso va a beneficiar al alumno que desee trabajar en este aspecto.

El objetivo principal de este curso práctico es dotar a los alumnos de conocimientos en conceptos básicos de bioinformática, los cuales les permitan manejar las distintas herramientas bioinformáticas de libre disposición, tanto online como software, siendo capaces de realizar búsquedas en bases de datos, imprescindibles en el desarrollo de cualquier investigación científica.

Métodos y contenido

El curso se ha distribuido en una serie de bloques temáticos sencillos que 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, se desarrollan una serie de contenidos específicos, a la vez que se realizan una serie de supuestos prácticos específicos de cada apartado que son perfectamente comparables a los que lleva a cabo un bioinformático profesional, aunque también se adaptan al nivel de cada alumno.

En el primer bloque ofrecemos una visión sencilla y resumida del "nacimiento" de esta nueva rama de la ciencia, la Bioinformática, originada respondiendo a las necesidades

actuales de la investigación cientifica, además de recordar una serie de conceptos biológicos, como por ejemplo las definiciones de ADN, el código genético, la reacción en cadena de la polimerasa o el fundamento de la secuenciación automática, creando el contexto donde se van a englobar los análisis bioinformáticos prácticos posteriores.

Hoy en día gracias a la bioinformática se pueden procesar las ingentes cantidades de datos generados por las técnicas de secuenciación, tanto automática como pirosecuenciación. Estos datos están depositados en las bases de datos, de las que la mayoría son públicas y accesibles. Durante el curso, para explicar el aumento de las secuencias disponibles y de los usuarios en las principales bases de datos a nivel global, nos servimos del ejemplo del NCBI, que tiene actualmente una media de 2,5 millones de usuarios al día, siendo GeneBank la principal base de datos que ha aumentado exponencialmente el número de secuencias depositadas en las últimas dos décadas. Los secuencias depositadas pueden ser nucleotídicas (nucleótidos/ADN) o proteicas (aminoácidos/peptidos/proteinas), aunque en este curso nos vaos a centrar en las primeras.

En el bloque siguiente se introduce al alumno a los formatos más comunes en los que están las secuencias de ADN depositadas, como son texto plano, en formato GeneBank, formato EMBL… y en formato FASTA, formato que vamos a utilizar a lo largo del curso por ser uno de los más habituales. Además, en este bloque mostramos a los alumnos los diferentes softwares y/o aplicaciones disponibles para editar secuencias, aunque nos decantaremos por herramientas de acceso libre. 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* (Hall, 1999), siendo este último con el que trabajaremos. Además, se muestran otro tipo de sofwares de pago, por si el alumno estuviese interesado en adquirir su licencia, como los paquetes *DNAstar* (DNASTAR Lasergene Inc.) y *Geneious* (Biomatters Limited).
Con el programa *BioEdit* se abren las secuencias recibidas, que deben estar en formato FASTA. Además, el programa permite abrir los cromatográmas 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 los distintos casos de cromatogramas correctos e incorrectos, siempre apoyados con ejemplos prácticos reales. Los alumnos realizan un análisis visual de los cromatogramas y aprenden a diferenciar un cromatograma de buena calidad de cromatogramas que tienen una serie de irregularidades o de calidad baja, además de cromatogramas de secuencias mezcladas o la presencia de contaminantes.

En el tercer bloque temático, utilizando las secuencias editadas obtenidas en la práctica anterior, los alumnos enfrentan las secuencias a algunas de las bases de datos disponibles, como la DNA Database of Japan (DDBJ) (http://www.ddbj.nig.ac.jp/), EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/) y GenBank (http://www.ncbi.nih.gov/). Además, se muestran muchas otras bases de datos propias de instituciones, como EzBioCloud para organismos procariotas, TAIR para *Arabidopsis* y SGD para *Saccharomyces cerevisiae*, son algunas de las más utilizadas. Aquí, nos centraremos en GeneBank, que a su vez contiene otras bases de datos como PubMed, Gene, EST, SNP, Structure y su recurso "estrella", BLAST (Altschul et al., 1990), recurso que los alumnos utilizarán seguidamente. BLAST (Basic Local Alingment Search Tool) es la herramienta más importante, fiable y flexible en estudios bioinformáticos que nos permite seleccionar una secuencia (query) y realizar alineamientos por pares de

secuencias con todas las secuencias de la base de datos, obteniendose por lo general muchas secuencias que guardan alguna similitud (target). Los alumnos estan capacitados y practicarán esta parte con las secuencias obtenidas anteriormente y serán capaces de interpretar los datos que nos ofrece esta herramienta.

En este punto, los alumnos son capaces de trabajar con secuencias nucleotídicas de una manera autónoma. Aún así, 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 como por ejemplo, ensamblaje de genes y genomas o el diseño de primers a partir de secuencias conservadas. Para ello, mostramos a los alumnos 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 (Edgar, 2004), disponible online entre los recursos que nos ofrece el Laboratorio Europeo de Biología Molecular (EMBL). Además, utilizamos el programa BioEdit para realizar los alineamientos múltiples, con la herramienta ClustalW, formato en el que la aplicación MUSCLE también ofrece los datos resultantes. Así los alumnos pueden comparar libremente, realizando los ensamblajes y alineamientos múltiples que les ofrecemos en los supuestos prácticos. Con los alineamientos múltiples, les introducimos el concepto de región o secuencia conservada que nos permita diferenciar entre géneros o especies incluso entre distintos Phyla y con ello, al diseño de primers o cebadores universales para amplificar estas regiones conservadas.

Para un buen diseño de cebadores o primers ponemos en conocimiento de los alumnos diversos programas de diseño de cebadores como son *Primer3* (Koressaar et al., 2007; Untergrasser et al., 2012) y algunos otros que las distintas marcas de biología molecular ponen a disposición de los usuarios, además de introducirles los distintos puntos clave para el diseño de un buen cebador.

Resultados

El presente curso tiene como objetivo principal iniciar en el manejo de secuencias nucleotídicos a alumnos y profesionales de diferentes áreas científicas, principalmente 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 3.0 en el que se pretende ofrecer cursos prácticos impartidos por profesores del centro con los propios recursos de los que dispone el centro.

Para saber como valoran el curso los alumnos que lo reciben así como de información adicional que nos permita mejorar en todos los aspectos, se realizaron encuestas de satisfacción a cada alumno.

En este sentido, este curso tuvo participación de personas pertenecientes a la Facultad de Farmacia (23%), a otras Facultades, como la Facultad de Ciencias Agrarias y Ambientales (30%), Facultad de Ciencias (7%), Facultad de Medicina (7%) o la Facultad de Biología (7%), además de profesionales del sector (30%). De entre los alumnos un 38% son estudiantes de Grado, un 38% estudiantes de posgrado y un 24% de otra procedencia. Además, la distribución entre sexos ha sido bastante igualitaria, siendo un 54% de mujeres frente a un 46% de hombres. La difusión ha funcionado a la perfección,

sobre todo mediante el boca a boca (69%), aunque parece que la difusión mediante la web de Eventum (15%) también ha surtido efecto, aunque menor.

En cuanto a la calidad del curso y de los ponentes, y utilizando una escala tipo Likert del 1 al 5, siendo 1 muy malo y 5 excelente, el 100% de los alumnos evaluaron el curso como bueno, muy bueno o excelente, obteniendose una nota media de satisfacción global de un 8,9 en una escala del 1 al 10. Cabe destacar que la mayoría de los alumnos consideran que las herramientas de las que se ofrece el conocimiento en este curso son excelentes y de mucha utilidad, además de contar con una muy buena organización del curso y una excelente calificación de los ponentes. Algunos alumnos han planteado la realización de un curso avanzado o al menos con una duración mayor, al igual que los profesionales del sector, que demandan la utilización de otros programas bioinformáticos de pago, con el inconveniente de que acarrearían cuantiosos gastos adicionales a la organización lo cual está en contradicción con el espíritu del programa por lo que de momento no lo contemplamos.

En general la satisfacción global tanto de los alumnos como de los ponentes con el curso ha sido muy buena, aunque se mejorarán diferentes aspectos en futuras ediciones y se valorará la inclusión de conceptos y metodologías adicionales para completar la formación basica del alumno.

Conclusión

Como conclusión general, podemos asegurar que el curso 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 lista de espera. Además, es destacable que la participación de los alumnos fue muy fluida y siempre mostraron un alto grado de participación y dinamismo por lo que, en base a esto y a las calificaciones obtenidas del curso, podemos asegurar que existe una relación directa y positiva entre los objetivos perseguidos y los resultados conseguidos, por lo que consideramos que el modelo utilizado es susceptible de utilizarse en futuras ediciones del programa EducaFarma.

Referencias

Allen, G.K. Bioinformatics: new technology models for research, education, and services. Educause Centre for Applied Research Bulletin. 2005; 8, 1–9.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403-10.

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.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 2004; 32 (5): 1792-1797.
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.

Koressaar T and Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics 2007; 23(10):1289-91.

National Institute of Health, 2000. NIH Working Definition of Bioinformatics and Computational Biology <http://www.bisti.nih.gov/CompuBioDef.pdf>.

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.