BIOINFORMÁTICA PARA PRINCIPIANTES II: INTRODUCCIÓN AL ANÁLISIS BIOINFORMÁTICO DE PROTEÍNAS

Esther Menéndez Gutiérrez, Raúl Rivas González

Departamento de Microbiología y Genética
Universidad de Salamanca.

Palabras clave: Educación, software gratuito, ProteinBank, ExPASy, aminoácido, NCBI, EMBL, MUSCLE.

Resumen

En la actualidad, existe una creciente demanda e interés de los científicos, investigadores e incluso docentes por la bioinformática, debido probablemente a las grandes posibilidades que ofrece la aplicación de esta disciplina. Debido a este interés se ha creado la necesidad de introducir a los alumnos, al igual que a los docentes e investigadores, en éste área de forma prioritaria. Por ello, se ha diseñado un curso básico para iniciar a los usuarios en el análisis bioinformático, en este caso de proteínas. Este curso se basa en el manejo de secuencias aminoacídicas, empleando las bases de datos de proteínas disponibles además de utilizar distintos softwares de predicción de las estructuras y de dominios proteicos, pudiendo realizar una modelización de proteínas. Este curso es básicamente práctico, por lo que se aplican los conocimientos básicos a sencillos ejemplos prácticos que permitan afianzar los conocimientos de los alumnos.

Introducción

En las úlltimas décadas, el desarrollo de la biología y la biotecnología ha generado una cantidad ingente de datos a un ritmo vertiginoso. De esta manera, en abril de 2015, GenBank, principal órgano de depósito de secuencias genéticas, acumuló aproximadamente 182 millones de registros de secuencias (http://www.ncbi.nlm.nih.gov/genbank/statistics). Además de esto, Uniprot KB/SwissProt, que es la principal base de datos en cuanto a función y secuencia primaria de proteínas, cuenta, a inicios de abril de 2015, con aproximadamente 600.000 entradas de secuencias de proteínas (http://web.expasy.org/docs/relnotes/relstat.html).

Con la acumulación de tales volúmenes de datos, la aparición de una nueva rama de conocimiento basada en el uso de los ordenadores para el manejo de dichos datos se ha convertido en un imperativo para la interpretación de resultados. Esta rama se denomina bioinformática, donde la informática y la estadística son aplicadas y combinadas para el analisis de los datos biológicos con el fin de acelerar, mejorar y diversificar la investigación biológica. De esta forma, se define la Bioinformática como la aplicación de los recursos informáticos a los datos biológicos, independientemente de la naturaleza de los datos, nucleotídicos o proteicos.

Hoy en dia, existe una gran capacidad de recopilar y manejar los datos biológicos, asi como de compartirlos. Las bases de datos biológicas o bioDBs desempeñan un papel cada vez más importante en la era post-genoma. Con la explosión de datos y capacidades biotecnológicas, sobre todo las industrias farmacéuticas y de la salud dependen de profesionales con conocimientos bioinformáticos, para aprovechar estos recursos, desarrollando nuevas tecnologías biológicas que sean positivas y de aplicación, sobre todo en salud humana y animal. En definitiva, la Bioinformática se ha establecido como una importante disciplina científica, estando ampliamente reconocido que la Bioinformática o la Biología Computacional es un campo multidisciplinar que requiere de un entrenamiento que englobe varias ramas de conocimiento para adquirir competencias básicas (Fetrow and John, 2006).

Debido a ello, se ha detectado la necesidad de que los alumnos, potenciales investigadores y docentes, y el personal ya establecido sean entrenados en el empleo y uso de las bases de datos, ya que las colecciones de bases de datos existentes están en continuo crecimiento, acumulando una enorme cantidad de datos sobre secuencias, estructura y actividad de proteínas. Los alumnos de este curso deben ser capaces de entender e integrar los sistemas bioinformáticos con los datos biológicos aplicando los programas pertinentes, dentro o fuera de los que se les ofrecen, que les permitan afrontar el reto que plantea la investigación actual.

Como objetivo principal, este curso está diseñado para que los estudiantes se animen a desarrollar estas habilidades, fuera de su disciplina formal, aumentando sus conocimientos en el manejo de programas para el análisis bioinformático de proteínas, provocando que la información sea menos compleja y más comprensible, además de más accesible. Para ello se les ha expuesto e introducido a las herramientas de software adecuadas, siempre de acceso libre, que les permita la utilización de todos los recursos que les ofrecen diversos sistemas.

Métodos y contenido

El curso se ha distribuido en una serie de bloques temáticos que comprenden teoría y práctica, comenzando por la impartición de nociones básicas sobre bioinformática y proteómica, seguido de unos supuestos prácticos que contribuyen a la fijación de conceptos. Además, se presentan diversas aplicaciones y programas informáticos, con el fin de que al final del curso, el alumno pueda trabajar con secuencias aminoacídicas con autonomía y seguridad. Los bloques temáticos se distribuyen de la siguiente manera:

- Bloque 1. Breve introducción a la bioinformática. Conceptos básicos sobre proteínas. Ventajas de trabajar con secuencias proteicas respecto a las nucleotídicas. Formatos de secuencias proteicas.
- Bloque 2. Bases de datos. Comparación de secuencias aminoacídicas.
- Bloque 3. Alineamiento múltiple de secuencias proteicas.
- Bolque 4. Predicción de dominios proteicos.
- Bloque 5. Modelización y predicción de estructuras proteicas.

En cada bloque temático se ofrece la posibilidad de realizar una serie de ejercicios prácticos, siguiendo una lógica similar a la de los análisis bioinformáticos llevados a cabo por un bioinformático, aunque siempre adaptados a las necesidades del alumno.

En el primer bloque comenzamos introduciendo al alumno a la bioinformática, a la vez que se recuerdan diversos conceptos referentes a la estructura proteica, los aminoácidos y las definiciones de dominio proteico y de los marcos de lectura (ORF-*open reading frame*). Además, se introducen conceptos de biología molecular básica que apoyen la introducción de los análisis que se van a llevar a cabo, poniendo ejemplos en cada caso.

Seguidamente, se expondrán las ventajas de trabajar con secuencias aminoacídicas frente a trabajar con secuencias nucleotídicas, poniendo ejemplos y utilizando distintas herramientas para convertir una secuencia nucleotídica en una secuencia aminoacídica, empleando el código genético y siempre trabajando con secuencias de ambos tipos en formato FASTA. Para esta tarea, se utilizan aplicaciones on line gratuitas englobadas en los recursos que nos ofrece el ExPASy, el portal de recursos bioinformáticos del Instituto Suizo de Bioinformática (SIB, www.expasy.org). En este punto y dependiendo de la secuencia de ADN que tomamos como partida, los alumnos tendran la capacidad de dilucidar cual es el marco de lectura con más posibilidades de ser cierto, entre los distintos que nos ofrece el programa, teniendo en cuenta que todas las proteínas comienzan por una metionina (ATG) y terminan en un codón de parada (TAA, TAG ó TGA). Además, se presenta el uso de softwares de edición de secuencias, que sirven tanto para secuencias de ADN como para secuencias proteicas, como el programa BioEdit (Hall, 1999). Aunque en principio es similar a la herramienta Translate del ExPASy, es menos intuitivo y la experiencia con los alumnos no expertos nos hace apostar por la herramienta online.

Utilizando las secuencias de la práctica anterior, en el siguiente bloque, se enfrentarán dichas secuencias contra las bases de datos disponibles como por ejemplo Protein Data Bank in Europe (PDBe) (www.ebi.ac.uk/pdbe), GenBank (http://www.ncbi.nih.gov/) y/o UniProtKB/Swiss-Prot (Universal Protein Knowledgebase: http://www.uniprot.org/help/uniprotkb). Nos centraremos en GeneBank y UniProtKB, poseyendo ambas la herramienta BLAST (Altschul et al., 1990). Ambas bases de datos son perfectamente válidas, aunque el alumno ha de ser capaz de comprobar que UniProtKB ofrece una serie de ventajas respecto al NCBI como por ejemplo, una mayor calidad de las anotaciones referidas a la función de las proteínas, uso de palabras clave estándar para describir las funciones, poseer una herramienta interactiva de búsqueda y ofrecer la posibilidad de exportar los resultados en forma de tabla, aunque también tiene algún inconveniente como que sólo funciona para secuencias proteicas y que su base de datos es menor que la del NCBI. Utilizando ambas bases de datos, los alumnos son capaces de enfrentar la secuencia editada por ellos mismos a protein-BLAST e interpretar los datos.

En este punto, el alumno ya es capaz de trabajar individualmente con las secuencias aminoacídicas disponibles, aunque creemos que es necesario ofrecer alguna herramienta más con la que puedan completar su formación básica herramientas que desarrollaremos en los bloques tercero, cuarto y quinto.

Se presentan diferentes aplicaciones y casos prácticos para la realización de alineamientos múltiples (cuando comparamos más de dos secuencias), como es la herramienta MUSCLE (Edgar, 2004), disponible online entre los recursos del Laboratorio Europeo de Biología Molecular (EMBL) y del propio Bioedit, programa con el que también podremos realizar este tipo de alineamientos. La finalidad es que realicen un alineamiento múltiple y puedan observar el grado de conservación de las

proteínas dentro del grupo de organismos seleccionados contenidos en los casos prácticos.

Posteriormente, se realiza una predicción de dominios proteicos, utilizando la herramienta InterProScan5 (Jones et al., 2014). Esta aplicación nos ofrece mucha información sobre el tipo de proteína, sus componentes y en muchos casos, el artículo científico donde se publicó.

Finalmente, utilizamos la herramienta Phyre2 (Protein Homology/analogY Recognition Engine V 2.0) (Kelley and Sternberg, 2009) para la predicción y modelización de estructuras proteicas. En ella y a partir de alineamientos con las bases de datos disponibles, se crea un modelo de estructura de la proteína seleccionada (Figura 1), además de darnos más información sobre dominios y estructuras, entre otras.

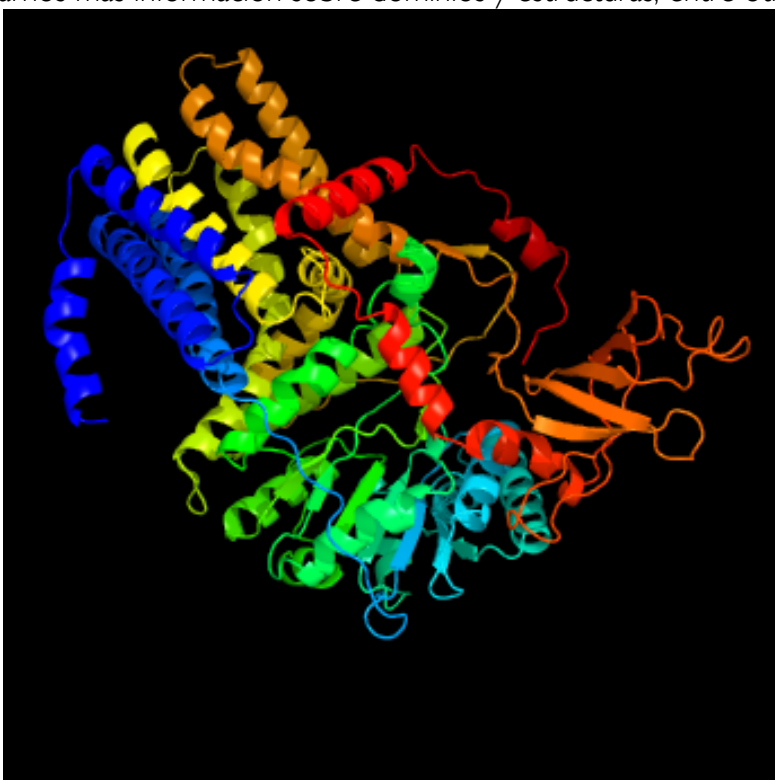

Figura 1. Predicción de la estructura de CelA, celulosa sintasa de *Rhizobium leguminosarum*, realizada con la aplicación Phyre2

Resultados

En el presente curso hemos tenido como objetivo principal la iniciación del alumno en el manejo de secuencias proteicas. Estos alumnos y profesionales provienen de diferentes áreas científicas, aunque en especial va dirigido a aquellos que no posean conocimientos previos en esta área. Además, el curso se engloba dentro del Proyecto de Innovación Docente EducaFarma 3.0 en el que se pretende ofrecer cursos/talleres o seminarios prácticos impartidos por profesores del centro, empleando sus propios recursos.

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 distintos aspectos, como su formación, su nivel de satisfacción en la

calidad de los ponentes y en general con el curso, entre otros. Algunos de estos aspectos fueron evaluados mediante una escala de valoración tipo Likert del 1 al 5, siendo 1 muy malo y 5 excelente.

Los alumnos que participaron en este curso/taller, estaban mayoritariamente vinculados a la Facultad de Farmacia de la Universidad de Salamanca (un 54% de asistentes). Sin embargo, las encuestas muestran que hubo participantes de otras Facultades, como la Facultad de Ciencias Agrarias y Ambientales (33%) y a la Facultad de Biología (13%) De entre los alumnos, un 40% son estudiantes de Grado y un 60% estudiantes de posgrado. Además, la distribución entre sexos ha sido paritaria, un 53% de mujeres y un 47% de hombres. El rango de edades de los asistentes ha estado restringido entre 19 y 31 años, siendo los menores de 20, el rango mayoritario. La difusión ha funcionado a la perfección, principalmente mediante la herramienta Eventum (27%) y el boca a boca (27%), aunque la difusión via Facebook ha sido también bastante efectiva (13%).

En general, el curso/taller propuesto ha tenido muy buena acogida entre los destinatarios posibles dentro de las áreas científicas prioritarias. La satisfacción global media ha sido elevada obteniendo una nota media de 8,20 en una escala del 1 al 10, lo cual nos anima a repetir el curso en futuras ediciones. Además, todas las notas medias de los distintos aspectos, como por ejemplo la organización de contenidos, la calidad de los ponentes y la utilidad de las herramientas propuestas en el desarrollo de su futura actividad docente, investigadora o estudiantil, han sido elevadas, todas alrededor del 4,5 en una escala del 1 al 5.

Es necesario mencionar que un aspecto que algunos de los alumnos han valorado negativamente es la duración del curso, ya que lo consideran demasiado corto y con poca profundidad por lo en ediciones futuras cabe la posibilidad de proponer una duración superior con una ampliación de conceptos al igual que dividir el curso en distintas sesiones, lo cual permita a los alumnos asimilar los contenidos de una manera más eficaz.

Conclusión

Como conclusión general, el presente curso ha sido un éxito de acogida por segunda vez consecutiva ya que se cubrieron todas las plazas disponibles en un corto periodo de tiempo e incluso se formó una larga lista de espera de interesados, aunque existe la posibilidad de realizar mejoras de cara a un futuro próximo. Además, es destacable el alto interés de los alumnos, muy participativos y dinámicos. En definitiva, podemos asegurar que existe una relación directa y positiva entre los objetivos perseguidos y la evaluación obtenida, por lo que consideramos que el modelo utilizado es susceptible de utilizarse en futuras ediciones del programa EducaFarma.

Referencias

Fetrow, J.S., John, D.J. Bioinformatics and computing curriculum: a new model for interdisciplinary courses. 2006. ACM SIGCSE Bulletin 38, 185–189.

Magrane, M., and Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. 2011. Database, 2011, bar009.

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

Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004 Aug;5 113. doi:10.1186/1471-2105-5-113.

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.

Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics, Jan 2014; doi:10.1093/bioinformatics/btu031

Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R. A new bioinformatics analysis tools framework at EMBL-EBI. 2010. Nucleic acids research 2010 Jul, 38 Suppl: W695-9.

Kelley LA and Sternberg MJE. Protein structure prediction on the web: a case study using the Phyre Server. 2009. Nature Protocols 4, 363 – 371.