Nociones básicas para el análisis bioinformático de proteínas

Esther Menéndez Gutiérrez, Raúl Rivas González

{esthermenendez, raulrg}@usal.es

Departamento de Microbiología y Genética, Universidad de Salamanca

Palabras Clave: educación bioinformática, software gratuito, proteína, ExPASy, aminoácido, NCBI, EMBL

Resumen

Actualmente existe una capacidad aguda de recopilar, compartir y analizar datos lo que ha dado lugar a un creciente aumento en el interés de la comunidad científica por la bioinformática, debido probablemente a las enormes posibilidades de aplicabilidad que ofrece esta disciplina. Introducir a los alumnos, docentes e investigadores en éste área es una necesidad. Con éste fin, hemos diseñado un curso/taller básico en el que se inicia al alumno en el análisis bioinformático de proteínas. Este curso se basa en el empleo de las bases de datos de proteínas y en la predicción de las estructuras proteicas mediante estudio de dominios proteicos y modelización de proteínas aplicándolo a sencillos ejemplos prácticos que permitan afianzar los conocimientos de los alumnos

Introducción

La biología es una disciplina basada en el conocimiento, por lo que ahora más que nunca, apoyándose en las diversas técnicas integrales que se han desarrollado en los últimos años, produce datos a un ritmo inexorable. Por ejemplo, en junio de 2013, GenBank, que es el principal repositorio de secuencias genéticas, acumuló aproximadamente 164 millones de registros de secuencias de genes y aproximadamente 110 millones de entradas de secuencias del genoma (http://www.ncbi.nlm.nih.gov/genbank/statistics). Además de esto, Uniprot Knowledgebase (Uniprot), que es la principal base de datos en cuanto a función y secuencia primaria de proteínas, cuenta con más de 13,5 millones de entradas desde enero de 2011 (2).

Con la acumulación de tales volúmenes de datos, el uso de los ordenadores se ha convertido en un imperativo para su curación e interpretación. Esto ha dado lugar a un nuevo campo de investigación, a saber, la bioinformática, donde la ciencia informática y la estadística son aplicadas para analizar los datos biológicos con el fin de acelerar,

mejorar y diversificar la investigación biológica. Así, la Bioinformática, esencialmente se puede definir como la aplicación de los recursos informáticos a los datos biológicos, independientemente de que el tipo de datos sean secuencias de ADN, secuencias de proteínas así como dominios y estructuras de proteínas.

Nunca antes los científicos han tenido la capacidad de recopilar, compartir y analizar datos como la que tienen hoy. Por tanto, las Bases de datos Biológicas (bioDBs) desempeñan un papel cada vez más importante en la era post-genoma. Con esta explosión de datos moleculares y capacidades biotecnológicas, las industrias farmacéuticas y de la salud dependen de profesionales con conocimientos de tecnología bioinformática, para aprovechar plenamente estos recursos y traducir los datos moleculares y celulares en los nuevos descubrimientos genéticos y terapéuticos, así como desarrollar nuevas biotecnologías que afecten positivamente a la salud humana. En definitiva, la bioinformática se ha establecido como una importante disciplina científica, dando lugar a un cambio de paradigma en diversas disciplinas como la medicina molecular, la genómica comparativa, la evolución molecular, las aplicaciones de los genomas microbianos, el descubrimiento de fármacos y la biotecnología en cualquiera de sus ramas. A este respecto, está ampliamente reconocido que la Bioinformática o la Biología Computacional es un campo multidisciplinar que requiere de un entrenamiento preciso para adquirir competencias básicas (1).

En este sentido, los futuros egresados han de ser entrenados en el empleo y uso de las estructuras de datos biológicos, ya que la colección de bases de datos que existen en el planeta no dejan de crecer a un ritmo exponencial, acumulando una cantidad ingente de datos sobre secuencias génicas o estructura y actividad de proteínas. Los alumnos deben ser capaces de entender e integrar los sistemas bioinformáticos con los datos biológicos aplicando los programas pertinentes que les permitan afrontar el reto que se les plantea la investigación actual.

Por esta razón, y como objetivo principal, este curso ha sido diseñado para animar a los estudiantes a desarrollar habilidades fuera de su disciplina formal, aumentando sus conocimientos en el manejo de programas para el análisis bioinformático de proteínas, provocando que la inmensa y diversa información de la que disponemos sea menos compleja y sí más comprensible y accesible. Para ello les hemos informado y adiestrado sobre las herramientas de software adecuadas que les permita lograr una

interoperabilidad para mejorar el acceso y uso de los datos disponibles de diversos sistemas.

Contenidos y metodología

Para alcanzar nuestro objetivo, el desarrollo del curso se ha distribuido en una serie de bloques temáticos, comenzando por ofrecer al alumno unas nociones básicas sobre bioinformática y proteómica y terminando con la modelización y predicción de estructuras proteicas. Además, les hemos informado de aplicaciones y programas informáticos que podrán utilizar, con el fin que el alumno, al final del curso pueda trabajar con secuencias aminoacídicas con autonomía y seguridad. Los bloques temáticos se distribuyen de la siguiente manera:

- Bloque 1. 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.
- Bloque 3. Alineamiento múltiple de secuencias proteicas.
- Bolque 4. Predicción de dominios proteicos.
- Bloque 5. Modelización y predicción de estructuras proteicas.

Dentro de estos bloques temáticos, se ofrece al alumno la posibilidad de realizar una serie de ejercicios prácticos propuestos específicamente para cada bloque, siguiendo una lógica similar a la de los análisis bioinformáticos llevados a cabo por un bioinformático, aunque siempre adaptados a las necesidades del alumno. Dentro de cada bloque temático se desarrollan una serie de contenidos específicos, que se detallan en los parrafos siguientes.

En el primer bloque se ofrece una introducción general a la bioinformática y se refrescan conceptos básicos como por ejemplo, la composición del código genético, el dogma de la biología molecular, la estructura de las proteínas y su composición aminoacídica, la definición de marcos de lectura (ORF-*open reading frame*) y dominios proteicos. Además, se introducen conceptos de biología molecular básica que apoyen la introducción de los análisis que se van a llevar a cabo.

En este bloque, el alumno podrá conocer las ventajas de trabajar con secuencias aminoacídicas frente a trabajar con secuencias nucleotídicas. En el transcurso de la explicación, se ofrecen distintas herramientas para trasladar información de una secuencia nucleotídica a una secuencia aminoacídica, empleando el código genético y siempre trabajando con secuencias de ambos tipos en formato FASTA. Para esta tarea, se presentan dos posiblidades al alumno. La primera es utilizar aplicaciones on line

gratuitas, como es la herramienta Translate, englobada en los recursos que nos ofrece el ExPASy, el portal de recursos bioinformáticos del Instituto Suizo de Bioinformática (SIB, www.expasy.org). Esta utilidad nos devuelve 6 marcos de lectura para una misma secuencia de ADN que se corresponden a todas las posibles proteínas que se puedan formar con una diferencia de +/- 3 nucleótidos. En este punto y según como sea 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, teniendo en cuenta que todas las proteínas comienzan por una Metionina (ATG) y terminan en un codón de parada (TAA, TAG ó TGA). La segunda opción es utilizar softwares de edición de secuencias, que sirven tanto para secuencias de ADN como para secuencias proteicas, como el programa BioEdit (Hall, 1999), que es gratuito y muy sencillo de utilizar. De entre las múltiples utilidades de las que dispone, también posee una herramienta para trasladar secuencias nucleotídicas a secuencias de proteínas, ofreciendo al usuario todas las posibles opciones. Aunque en principio es similar a la herramienta Translate del ExPASy, es menos intuitivo y la experiencia con los alumnos no expertos nos hace inclinarnos hacia una apuesta por la herramienta online.

En el segundo bloque, los alumnos practican y trabajan con las secuencias protéicas obtenidas en la parte práctica del bloque anterior, enfrentando dichas secuencias contra las bases de datos disponibles. Algunos ejemplos son la Protein Data Bank in Europe (PDBe) (www.ebi.ac.uk/pdbe), que es una base de datos de estructuras 3D de proteínas dependiente del Intituto Europeo de Bioinformática (EMBL-EBI), el GenBank (http://ww.ncbi.nih.gov/) base de datos general dependiente del National Center for Biotechnology Information (NCBI) y UniProtKB (Universal Protein Knowledgebase: http://www.uniprot.org/help/uniprotkb) y Swiss-Prot especializadas en secuencias de proteínas dependientes del Instituto Suizo de Bioinformática (SIB).

En este curso, nos hemos centrado 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 (3). Además, como recurso preferente en análisis de proteínas vamos a utilizar UniProtKB, que es la fuente universal de proteínas, un repositorio central de datos de proteínas creado por la combinación de: UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef), y la UniProt Archive (UniParc). Esta base de datos también posee la herramienta BLAST, con un funcionamiento similar que en GeneBank. UniProt es una colaboración entre el European Bioinformatics Institute (EBI), el Swiss Institute of Bioinformatics (SIB) y el Protein

Information Resource (PIR). Entre las tres instituciones, trabajan cerca de 150 personas en distintas tareas como la curación de las bases de datos, el desarrollo de software y el soporte técnico. Esto la ha convertido en el líder mundial en el almacenamiento de información sobre proteínas.

Ambas bases de datos son perfectamente válidas aunque el alumno ha comprobado 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, un estandarizado del uso de palabras clave para describir las funciones, poseer una herramienta interactiva de búsqueda y ofrecer la posibilidad de exportar los resultados en forma de tabla, aunque como inconvenientes hay que resaltar que sólo funciona para secuencias proteicas y que su base de datos es menor que la del NCBI.

Del mismo modo, 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 y 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 estan capacitados para enfrentar la secuencia editada por ellos mismos en protein-BLAST e interpretar los datos, ya que la aplicación nos devuelve los alineamientos más relacionados con la secuencia aminoacídica dada.

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. En los bloques tercero, cuarto y quinto, se introducen diversos programas para realizar análisis de dominios proteicos y conservación de dichos dominios, a la par que modelización de proteinas, es decir, predicción de sus estructuras terciarias y cuaternarias.

Para los alineamientos múltiples, se utiliza la herramienta MUSCLE (Edgar, 2004), disponible online entre los recursos que nos ofrece el Laboratorio Europeo de Biología Molecular (EMBL). Para trabajar este punto, les facilitamos un supuesto práctico en el que se les ofrece una serie de secuencias aminoacídicas correspondientes a una serie de proteínas homologas y un grupo de organismos concretos. La finalidad es que realicen un alineamiento múltiple y puedan observar el grado de conservación de las proteínas dentro del grupo de organismos seleccionados.

Posteriormente, se realiza con una o varias de las proteinas homologas una predicción de dominios proteicos, utilizando la herramienta InterProScan (4, 5), también disponible online entre los recursos que nos facilita el Laboratorio Europeo de Biología Molecular (EMBL). Esta aplicación nos ofrece mucha información sobre el tipo de proteína, incluyendo el artículo científico donde se publicó, si es el caso, y/o si posee péptido señal, imprescindible en algunas proteinas para su correcta localización.

Finalmente se mostró al alumno una herramienta para la predicción y modelización de esctructuras proteicas denominada Phyre2 (Protein Homology/analogY Recognition Engine V 2.0) (6) que depende del Structural Bioinformatics Group, Imperial College, London. En ella, además de más información (alineamientos, dominios, etc.), se crea, a partir de alineamientos con las bases de datos disponibles, un modelo de estructura de la proteina seleccionada (Figura 1).

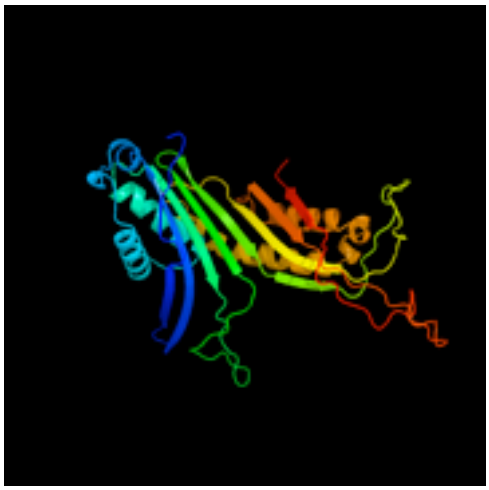

Figura 1. Predicción de la estructura de la uricasa de Aspergillus flavus, realizada con la aplicación Phyre2.

Resultados

El presente curso de nociones básicas para el análisis bioinformático de proteínas ha tenido como objetivo principal iniciar en el manejo de secuencias proteicas a alumnos y profesionales de diferentes áreas científicas, en especial, a los pertenecientes a la Facultad de Farmacia, que carezcan de conocimientos previos en este área. 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 aspectos. 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 58% de asistentes). Sin embargo, las encuestas muestran que la participación fue heterogénea ya que también hubo participantes de otras Facultades, como la Facultad de Ciencias Agrarias y Ambientales (21%), la Facultad de Biología (11%) y de la Facultad de Químicas (5%), así como una participación de profesionales adscritos a centros de Investigación (5%). De entre los alumnos, un 37% son estudiantes de Grado, un 16% estudiantes de posgrado, un 16% PDI y un 31% de otra procedencia. Además, han asistido mayoritariamente mujeres, un 74% frente a un 26% de hombres. El rango de edades de los asistentes ha sido muy variado, existiendo un marcado porcentaje de gente mayor de 40 años (21%), correspondiente sobre todo a personal técnico (técnicos de laboratorio y/o auxiliares de investigación) y trabajadores que posiblemente busquen renovar y/o ampliar sus conocimientos (31%). La difusión ha funcionado a la perfección, principalmente mediante la herramienta Eventum (47%), que ha sido la manera más eficiente de difusión, aunque parece ser que el boca a boca (37%) también ha funcionado correctamente en este sentido.

En general, el curso/taller propuesto ha tenido muy buena acogida entre los destinatarios posibles dentro de las áreas científicas prioritarias. La satisfacción global ha sido elevada ya que un 100% de los alumnos la valoran por encima de 7 puntos en una escala del 1 al 10, lo cual nos anima a repetir el curso en futuras ediciones. Además, el 100% de los asistentes puntuaron como muy bueno o excelente algunos aspectos, como son, la organización de contenidos, la calidad de los ponentes y principalmente la utilidad de las herramientas propuestas en el desarrollo de su futura actividad docente, investigadora o estudiantil.

Es necesario mencionar que un aspecto que algunos de los alumnos asistentes han valorado negativamente es la duración del curso (un 15%), ya que lo consideran demasiado corto por lo en ediciones futuras podríamos proponer una duración superior con una ampliación de conceptos, como por ejemplo, introducción a las herramientas

bioinformáticas para realizar filogenias, o al menos 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 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 lista de espera de interesados. Aunque existe la posibilidad de realizar muchas mejoras de cara a un futuro o próximas ediciones del curso, el alumnado ha expresado por medio de unas sencillas encuestas su satisfacción total con el curso. Además, es destacable el interés del alumnado, que ha sido excelente, siendo muy participativos y resultando un curso muy dinámico. También podemos destacar el interés suscitado en los profesionales técnicos, que optan por realizar el curso para renovar y/o ampliar sus conocimientos lo que nos permite suponer que la formación en esta área es muy demandada.

En definitiva, podemos asegurar 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 del programa de innovación docente donde se engloba este curso/taller, EducaFarma 2.0.

Referencias

1. Fetrow, J.S., John, D.J. Bioinformatics and computing curriculum: a new model for interdisciplinary courses. 2006. ACM SIGCSE Bulletin 38, 185–189.
2. Magrane, M., and Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. 2011. Database, 2011, bar009.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403-10.
4. Zdobnov E.M. and Apweiler R. InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 2001, 17(9): 847-8.
5. 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.
6. Kelley LA and Sternberg MJE. Protein structure prediction on the web: a case study using the Phyre Server. 2009. Nature Protocols 4, 363 – 371.