Hamut'ay
Revista semestral de divulgación científica
División de Investigación y Extensión Científica Tecnológica (DIECT-DUED).
Dirección Universitaria de Educación a Distancia-UNIVERSIDAD ALAS PERUANAS
http://revistas.uap.edu.pe/ojs/index.php/HAMUT/issue/archive
ISSN: 2313-7878 DOI: http://dx.doi.org/10.21503/hamu.v3i1.1000 Vol. 3(1). Enero-junio: Hamut'ay 2016. Lima-Perú

SOFTWARE LIBRE ENFOCADOS EN DIVERSOS CAMPOS DE LAS CIENCIAS BIOLÓGICAS

Free Software focused on various fields of Biological Sciences

Miguel Angel Alcalde-Alvites[1]
Universidad Nacional Mayor de San Marcos, Perú

Recibido: 13-04-2016
Aceptado: 24-06-2016

Resumen

El desarrollo de software y programas como herramientas en bioinformática continúa en auge debido a la gran cantidad de información biológica obtenida en los últimos años y que debe ser objeto de análisis. Muchos de estos software se enfocan en bases de datos biológicos, análisis de secuencias, anotación de genomas, evolución biológica computacional, medición de la biodiversidad, análisis de la expresión y regulación génica, genómica comparativa, predicción de la estructura de las proteinas, modelado de sistemas biológicos, acoplamiento molecular, entre otros. En este artículo se presenta una lista de software libre acompañados de una breve descripción de sus bondades y utilidad para conocimiento de la comunidad académica o interesada en estudios bioinformáticos en Perú. El trabajo es resultado de una búsqueda en bases de datos importantes que tienen en línea una serie de herramientas de utilidad valiosa para los estudios biológicos y que ha revolucionado el campo bioinformático; siendo este un estudio documental, se presenta el análisis de la revisión de la literatura por un periodo de once años, tiempo en el que tuvo una mayor incidencia la aplicación de la bioinformática en biología.

Palabras Clave: *Bioinformática, software libre, Biología, ciencias ómicas, tecnología.*

Abstract

The development of software and programs as tools in bioinformatics is still booming due to the large amount of biological information obtained in recent years, which should be analyzed. Many of these software are focused on biological databases, sequence analysis, annotation of genomes, computational evolutionary biology, measuring biodiversity, analysis of gene expression and regulation, comparative genomics, prediction of protein structure, modeling of biological systems, molecular docking, among others. This article presents a list of free software along with by a brief description of its benefits and applicability for the knowledge of the academic community or people interested in bioinformatics studies in Peru. The study is the result of a search in major databases, which have a number of online valuable tools, useful for biological studies, which have revolutionized the bioinformaticsfield; this being a documentary study, analysis of the literature review is presented for a period of eleven years, time in which had a higher incidence application of bioinformatics in biology.

Keywords: *Bioinformatics, Free Software, Biology, Omics Science, Technology.*

1 Facultad de Ciencias Biológicas, Escuela Académica Profesional de Genética y Biotecnología. Miembro del Grupo de investigación de fisiología vegetal y fitoquímica. (UNMSM) y del proyecto "Interactores moleculares en Carcinoma Hepatocelular. E-mail: miguelalcalde.unmsm.edu.pe@gmail.com

Introducción

La bioinformática ha crecido en los últimos años, de acuerdo con el avance de las ciencias ómicas (genómica, proteómica, metabolómica, transcriptómica, entre otras). Su información ha aumentado y, por lo tanto, sus herramientas de análisis también. La creación de software o programas que permiten dar distintos enfoques a la información obtenida se ha expandido y convertido en un sector económico y de investigación importante. Por ello, el diseño y elaboración de software en el campo bioinformático presenta un gran impacto e implica un costo importante; pero, a la larga, produce mayores beneficios como la reducción de gastos en la utilización de instrumentos costosos. En muchos casos, estos programas son gratuitos o de libre acceso para el público en general, por lo que cualquier investigador puede utilizarlos. Los campos biológicos a los que se aplican pueden ser el análisis de secuencias, la gestión de biodatos, el análisis filogenético, la expresión génica, el estudio proteómico, la interacción molecular, entre otras. Con este trabajo se busca compartir información acerca de algunos importantes software, programas o bases de datos enfocados en temas relacionados a la biología; información conocida a nivel mundial que debe ser divulgada ampliamente en nuestro país y que es necesaria para avanzar en las distintas áreas de investigación, tanto como en las ciencias biológicas y las ciencias de la información.

Método

Este artículo de revisión presenta un fundamento teórico y conceptual basado en la lectura, análisis, revisión e interpretación de páginas web, tesis y artículos científicos relacionados con la bioinformática. Es producto de la búsqueda minuciosa en importantes bases de datos, como son PubMed, Elservier, SciElo, Redalyc, NCBI, Dialnet, DOAJ, biblioteca virtual de CONCYTEC, repositorios virtuales nacionales e internacionales; así como en Google Books y Google Scholar.

Los descriptores empleados fueron:

1. bioinformática
2. software para biología,
3. uso de tecnologías para biología,
4. Software libre para biología
5. Tools for bioinformatics
6. Free software for biology
7. bioinformatics

El análisis y revisión de la literatura se realizó en el periodo comprendido entre los años 2004 al 2015, tiempo en el que tuvo una mayor incidencia la aplicación de la bioinformática en biología y ciencias relacionadas. Siendo este un estudio documental, después de realizar la exploración teórica, se organizó el artículo estructurándolo desde definiciones específicas, tipos de software y la descripción de software específicos de libre acceso para el campo de la biología.

Revisión de la literatura

Informática

La informática es el estudio de la estructura, el comportamiento y las interacciones de los sistemas computacionales naturales y artificiales. A esta ciencia le concierne la recolección, comparación y análisis de datos, tanto como el intercambio de información, datos y conocimientos; así como el uso y desarrollo de tecnologías que faciliten todo este proceso mencionado. Por lo que presenta aspectos computacionales, sociales y cognitivos, dando una noción de la transformación de la información por artefactos u organismos. Esto se observa claramente cuando se analiza el avance de las ciencias de la información, que va de la mano con la de ingeniera de sistemas de la información. Para ello, la informática está desarrollando sus propios conceptos fundamentales de la comunicación, el conocimiento, la información y la interacción, relacionándolos con fenómenos tales como el cálculo, pensamiento y lenguaje.

La informática abarca tres disciplinas académicas existentes: la inteligencia artificial, las ciencias cognitivas y las ciencias de la computación. Las ciencias cognitivas se refiere al estudio de los sistemas naturales; las ciencias de la computación se refiere al análisis de la computación y el diseño de sistemas informáticos; mientras que la inteligen-

cia artificial juega un papel de conexión y diseño de sistemas que emulan a las que se encuentran en la naturaleza. (Young, 2000)

Bioinformática

La bioinformática, como definición básica, es la aplicación de la tecnología informática para la gestión de la información biológica. Como es claro, para el análisis de toda la data biológica es necesario combinar las ciencias computacionales, la estadística, la ingeniería y la matemática; por lo que resulta ser una ciencia interdisciplinaria. Como menciona Luscombe (2001) la bioinformática tiene tres objetivos básicos. El primero es organizar los datos de una manera que permita a los investigadores acceder a la información existente y nuevos ingresos a medida que se producen, un ejemplo muy claro son las distintas bases de datos. El segundo es el desarrollo de herramientas y recursos que ayudan en el análisis de los datos, como el análisis de secuencias de DNA o de proteinas por medio de alineamientos. Y el tercer objetivo es utilizar estas herramientas para analizar los datos e interpretar los resultados de una manera biológicamente significativa. Anteriormente, los estudios biológicos examinaron los sistemas individuales en detalle, y con frecuencia los compararon con unos pocos que están relacionados. En la actualidad es posible realizar análisis globales de todos los datos disponibles con el objetivo de descubrir los principios comunes que se aplican a través de muchos sistemas y resaltar las características novedosas. Durante los últimos años del siglo XX, los avances de los estudios en genética y las nuevas tecnologías de la información permitieron el surgimiento de una disciplina que creó vínculos sólidos entre la informática y las ciencias biológicas. Así se dio inicio a la Bioinformática. Además, la información se ha incrementado de forma lineal durante los últimos 10 años, observándose dicho fenómeno en bases de datos como NCBI, KEGG, PDB y UniProt. La bioinformática ha influenciado en una gran cantidad de campos de la biología. Como señala Perezleo (2003), las especialidades médicas más influenciadas son la Genética Médica, la Bioquímica Clínica, la Farmacología, las Neurociencias, la Estadística Médica, la Inmunología, la Fisiología y la Oncología. Las principales aplicaciones se da en la gestión de datos en los laboratorios, automatización de los experimentos, ensamblaje de secuencias contiguas, predicción de dominios funcionales en secuencias génicas, alineamiento de secuencias, búsquedas en bases de datos, determinación y predicción de la estructura de macromoléculas como las proteínas, y en la corroboración de teorías de la evolución molecular y la elaboración de árboles filogenéticos.

Software libre

La definición de software libre viene de un tipo particular de software, o programa de ordenador, que permite la utilización, copia y distribución, con modificaciones o no, de forma libre. A su vez, como menciona Barahona (2011), no se debe confundir el término libre con gratuito, ya que la definición implica libertad, más no ausencia de precio. Por lo tanto, el software libre debe cumplir cuatro condiciones básicas: ejecutar el programa para cualquier finalidad, estudiar el funcionamiento del programa adaptándolo a sus necesidades, redistribuir copias del programa y, finalmente, mejorar el programa, poniendo esas mejoras a disposición del público para beneficio de todos los usuarios.

Para Sala (2014), los modelos de software libre y el acceso abierto desde su origen en los años 60 y su apogeo en los años 90, basados en el libre uso y distribución del conocimiento y la información, ambas han apoyado mucho a la ciencia. Las diferencias radican en que el software libre propone la libertad de poder acceder y circunstancialmente modificar el código fuente de los programas o software, mientras que el acceso abierto propone la disponibilidad de forma pública y gratuita de contenidos digitales de muy diversas categorías, en conjunto con la facultad de poder compartir y reutilizarlos sin restricciones o con restricciones mínimas.

Software libre para aplicaciones en Ciencias biológicas

Existe una diversidad de aplicaciones para las que se desarrollaron los software. En este caso se hablará de los orientados a la biología, que es donde entra a tallar el término bioinformática. Hay

distintas definiciones de software libre. Alcalde-Alvites (2014) lo concibe como la aplicación de las técnicas computacionales para entender, organizar y analizar la información asociada a las macromoléculas (proteínas, ácidos nucleicos, carbohidratos, lípidos, etcétera) o una ciencia que examina la estructura y función de genes y/o proteínas a través del uso de análisis computacionales, estadística y patrones de reconocimiento.

La bioinformática utiliza una infinidad de software, pero muchos de ellos son privados o particulares, por lo que no salen al público en general. Herrera (s.f.) señala algunas razones por las que se debe utilizar software libre en bioinformática. Por ejemplo: la transparencia para entender el procedimiento del programa, importante para un investigador; la economía ya que al ser generalmente gratuito permite enfocar el dinero en otros instrumentos y el respeto a los estándares que garantiza la evaluación de los datos basados en conocimientos científicos y técnicos.

Tipos de software libre para aplicaciones en biología

La variedad de software libres que existen para aplicaciones en biología, se clasifica a continuación según la temática, área o el objetivo al que van enfocados:

Bases de datos biológicas. Estas tienen por finalidad almacenar mucha y variada información, desde secuencias de genes y genomas, hasta secuencias de proteínas y proteomas, artículos y literatura relacionada, rutas metabólicas, estructuras proteicas, ontología, Single Nucleotid Polimorfism o Polimorfismo de un solo nucleótido (SNP) y otras más. A su vez, hay páginas de bases de datos especializadas en un solo punto; por ejemplo, en algún organismo, enfermedad, etcétera.

National Center of Biotechnology Information (NCBI). La base NCBI es una de las más importantes bases de datos compuestas. Abarca una variedad de información, como genes, proteinas, genomas, expresión génica, además de presentar al usuario literatura relacionada con las moléculas que está investigando. Otra ventaja importante es la posibilidad de descargar la información para analizarla en otros trabajos y, a su vez, también permite depositar información como papers y reviews publicados en revistas especializadas. La figura 1 muestra la configuración de la página de NCBI.

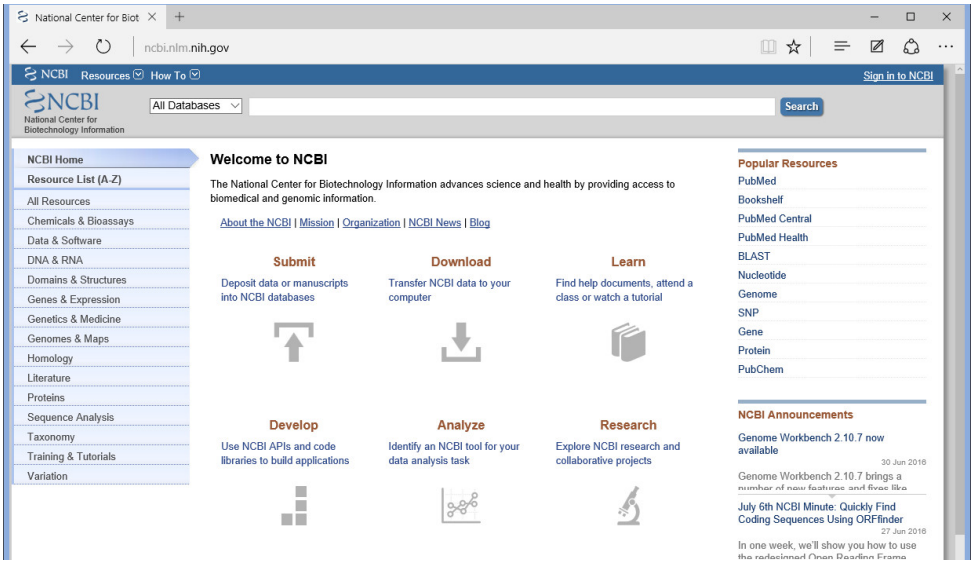

Figura1. Página principal de la base de datos NCBI.
Fuente: http://www.ncbi.nlm.nih.gov/

Uniprot. Esta base de datos se enfoca en el estudio de proteínas y proteomas. Se relaciona con la página Swiss-Prot, que sirve para realizar anotaciones de las secuencias. En esta página también se puede buscar proteínas y proteomas de interés para varios organismos; da la posibilidad de profundizar en el conocimiento de la proteína: localización celular, dominios, funciones, secuencia de aminoácidos, peso molecular, procesos biológicos relacionados, estructura, modificaciones post-traduccionales, interactómica, transcriptómica, etcétera. La figura 2 muestra la configuración de la página de Uniprot.

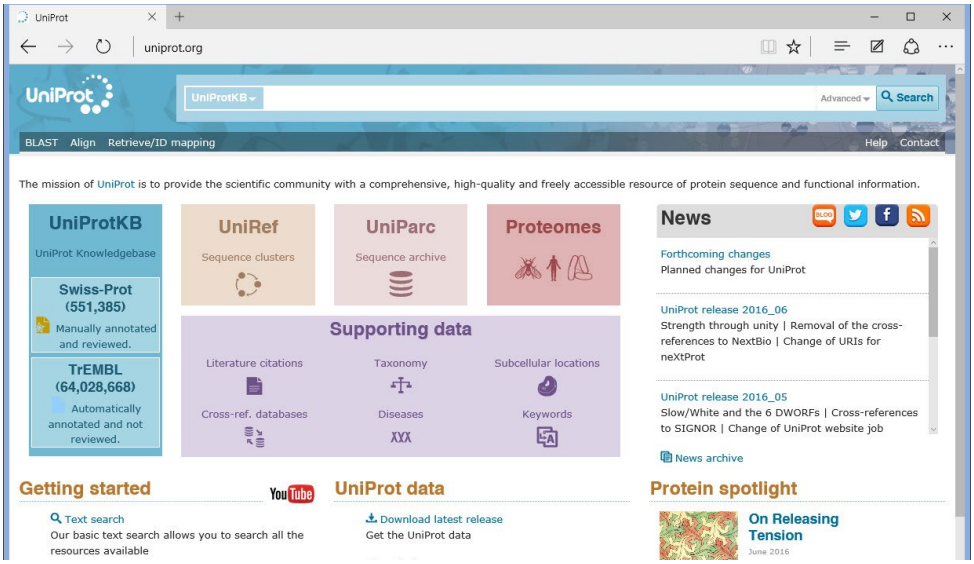

Figura 2. Página principal de la base de datos Uniprot.
Fuente: http://www.uniprot.org/

Kyoto Encyclopedia of Genes and Genomes (KEGG). Es una base de datos de información basada en genomas y proteomas, planteando el estudio de rutas metabólicas, procesamiento de información genética y ambiental, procesos celulares, enfermedades humanas, desarrollo del fármaco, etcétera. La figura 3 muestra la configuración de la página de KEGG.

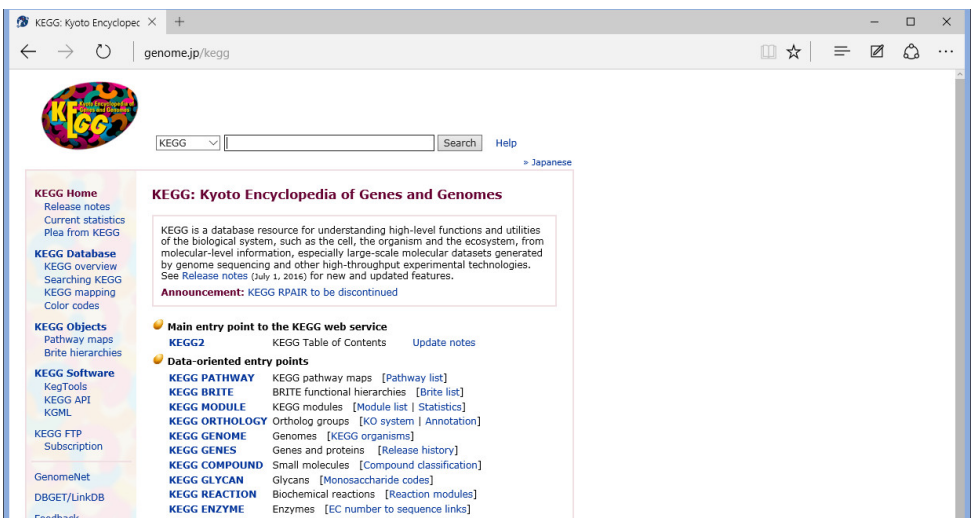

Figura 3. Página principal de la base de datos KEGG.
Fuente: http://www.genome.jp/kegg/

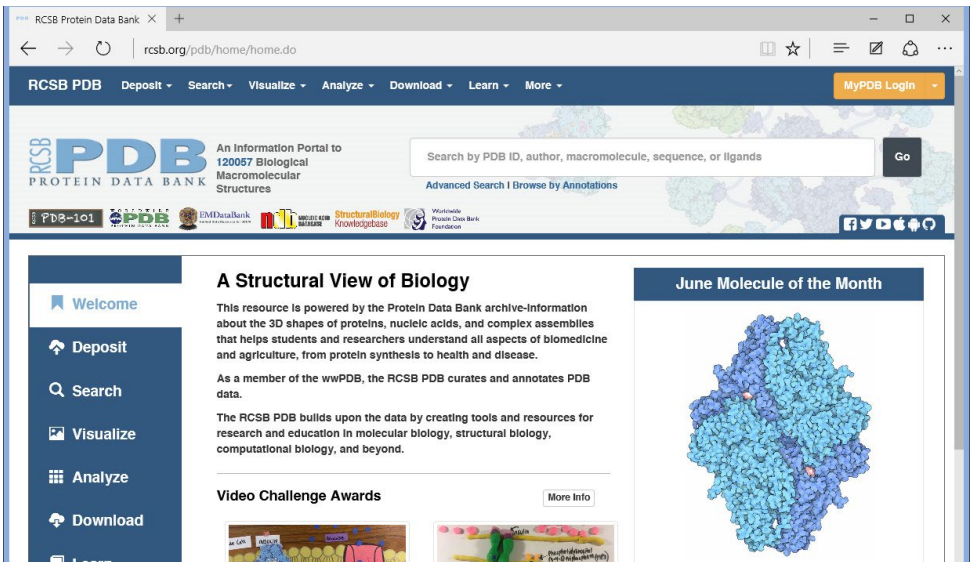

Figura 4. Página principal de la base de datos PDB.
Fuente:http://www.rcsb.org/pdb/home/home.do

Barcode of Life Data (BOLD). Como menciona Vera, Jiménez y Franco (2012, p. 200) BOLD es una plataforma libre a disposición de cualquier investigador, que sirve como regulador para garantizar que todos los registros de barcode o código de barras genético, cumplan estrictas normas que garanticen la precisión y exactitud de la identificación. La plataforma BOLD presenta tres importantes aplicaciones: la prime-

ra, es un depósito de secuencias correspondientes a la unidad básica de los estudios del "Deoxyribonucleic acid (DNA) barcode"; el segundo es una mesa de trabajo que ayuda con la gestión, el control de calidad y el análisis de los datos del "DNA barcode", y finalmente la tercera, proporciona un medio para la colaboración científica internacional permitiendo un paso a los medios de comunicación para una participación más flexible y segura. Se encuentra disponible en http://www.boldsystems.org/

Análisis de secuencias.

Los software bioinformáticos que realizan análisis de secuencias deben su origen al avance de las técnicas de secuenciación y a la gran masa de datos. Generalmente, estos programas realizan un alineamiento entre varias secuencias para análisis de regiones conservadas en filogenética, análisis mutaciones puntuales o análisis de SNP. La utilización para estas herramientas ha cambiado con el tiempo, se inició con los alineamientos en pares de secuencias en 1981 con el programa SSEARCH y posteriormente con ACANA, SPA, LALIGN o NW, para un par de años siguientes realizar el alineamiento en bases de datos en 1990 con BLAST. Más adelante, con software como HMMER, IDF, SWIPE y Genoogle paralelamente con el alineamiento múltiple por distintos software como MULTALIN, Clustal, SAM, T-coffee y Phylo; que permitió comparar muchas secuencias a la vez además de mejorarse la robustez del tamaño de la secuencia a analizar.

Basic Local Alignment Search Tool (BLAST). La base de datos NCBI realiza análisis de secuencias mediante una herramienta denominada Basic Local Alignment Search Tool (BLAST), que compara la secuencia nucleotídica o aminoacídica de interés con las secuencias de la base de datos. Luego envía las que son más similares, en base a valores de identidad y cobertura. Existen variantes como pBLAST, para el caso de comparación de secuencias de aminoácidos; nBLAST para comparar secuencias de nucleótidos, BlastX que permite ingresar una secuencia de nucleótidos que se traduce en sus seis posibles marcos de lectura y compara estas secuencias traducidas contra la base de datos de proteínas, y tBlastn, que permite comparar una secuencia de aminoácidos con la base de datos de nucleótidos. La página de la herramienta BLAST está disponible en http://blast.ncbi.nlm.nih.gov/Blast.cgi

Bioedit. Es un programa que permite editar alineamientos múltiples tanto de nucleótidos como aminoácidos. Entre sus herramientas se encuentra la búsqueda automática BLAST, alineamiento automático Clustal W, graficar y anotación de plásmidos, diseño mapas de restricción, alineamiento basado en la traducción de ácidos nucleicos, visualización y edición de cromatogramas, entre otras. La página de descarga libre está disponible en http://www.mbio.ncsu.edu/BioEdit/bioedit.html

Clustal X. Permite la elaboración alineamientos múltiples y la preparación de árboles filogenéticos. Este programa es utilizado sobre todo por la portabilidad a diferentes sistemas operativos importantes para todo investigador. Además, presenta una facilidad en el manejo de sus herramientas y flexibilidad para la visualización en la comparación de las secuencias. La página de Clustal X está disponible en http://www.clustal.org/clustal2/

Anotación de genomas

Se refiere a los procesos encargados de la anotación de genes; es decir, la búsqueda de los genes, asignando la función y otras características biológicas a una secuencia de ácido desoxirribonucleico (ADN). Estos programas se fortalecieron a base del progreso de la secuenciación masiva de genomas completos.

Artemis. El programa Artemis presenta la herramienta para anotar y navegar en los genomas, permitiendo la visualización de las características de la secuencia, datos de la secuenciación y los resultados de los análisis en el contexto de la secuencia, como también la traducción en sus seis marcos de lectura Open Reading Frame (ORF). La página de Artemis está disponible en http://www.sanger.ac.uk/science/tools/artemis

Blast2Go. Según Dias (2011), este programa permite anotar funcionalmente cDNA (Complementary DNA o ADN complementario), la cual es una hebra de ADN complementaria al ARN

mensajero y se sintetiza en una reacción que es catalizada por las enzimas de la transcriptasa inversa y ADN polimerasa, también anota las EST (Expressed Sequence Tag) que son una subsecuencia corta de una secuencia de cDNA, utilizada para identificar la transcripción de genes y, finalmente, anota CDS (coding sequences o secuencias codificantes) que son la parte del ADN o ARN de un gen, compuesto de exones, que codifica para la proteína o proteínas predichas a partir de estos datos. Al mismo tiempo, el software realiza estadísticas que permiten, por un lado, valorar la anotación realizada y, por otro, extraer información funcional. El formato de entrada es FASTA y utiliza BLAST contra la base de datos del NCBI no redundante (nr) para encontrar homólogos correspondientes con la secuencia de interés. La página de Blast2Go está disponible en https://www.blast2go.com/

Biología evolutiva computacional

Los programas enfocados en la biología evolutiva computacional estudian el origen ancestral de las especies, así como su cambio a través del tiempo. Estos software han permitido un seguimiento de la evolución de varios organismos midiendo los cambios en el DNA; conocer eventos evolutivos complejos por transferencia genética horizontal y crear modelos computacionales complejos de poblaciones para predecir el resultado del sistema a través del tiempo, realizar inferencia filogenéticas, entre otras tareas. A lo largo del tiempo se utilizaron y utilizan los que emplean sólo el método de matriz de distancias como el programa DENDRON, los que emplean sólo el método de máxima verosimilitud como el software GARLI, los que emplean sólo cálculos de distancia como el software DNAsp, los que emplean sólo el método bayesiano como el programa BEST y actualmente con mayor frecuencia se utilizan los que emplean diferentes enfoques como MEGA o DAMBE.

DAMBE. Este software permite ingresar secuencias y analizar las frecuencias de los nucleótidos que se encuentran entre los organismos analizados, mostrando gráficas de transiciones (mutaciones puntuales de la secuencias del ADN de purinas a purina o de pirimidina a pirimidina) y transversiones (mutaciones puntuales de la secuencias del ADN de purina a pirimidina o viceversa), versus la divergencia genética que ha sufrido estos organismos (las secuencias de ese marcador) a lo largo del tiempo. Este proceso se puede realizar por distintos modelos algorítmicos, que dependerán del investigador (Alcalde-Alvites, 2014). La página de DAMBE se encuentra disponible en http://dambe.bio.uottawa.ca/dambe.asp

MEGA. El programa MEGA permite realizar análisis filogenéticos, como elaborar árboles filogenéticos por medio de homologías entre secuencias; estimar tasas de evolución molecular y proveer hipótesis evolutivas. El programa brinda información importante sobre las secuencias introducidas para el análisis. Por ejemplo, mostrará sitios conservados, sitios variables, sitios informativos o sinapomorfias, que son los principales caracteres para la elaboración de los árboles filogenéticos, además de los sitios autopomórficos. Son clave para definiciones evolutivas. El software presenta también opciones para hallar distancias genéticas entre las taxas, la inferencia filogenética o la elaboración de árboles filogenéticos a partir de las distancias genéticas con un análisis bootstrap, que es un soporte estadístico para conocer el nivel de confianza con el que se está elaborando el árbol filogenético (Kumar, 2008). La página de MEGA está disponible en http://www.megasoftware.net

Mauve. El programa Mauve es un sistema útil para realizar múltiples alineaciones de genomas, en presencia de eventos evolutivos a gran escala, como la reordenación y la inversión. Las alineaciones múltiples de genomas proporcionan una base para la investigación en genómica comparativa y el estudio de la dinámica evolutiva de todo el genoma. La página de Mauve está disponible en http://darlinglab.org/mauve/mauve.html

DNA Sequence Polimorfism (DnaSP). El software DnaSP se encarga del análisis de polimorfismos de nucleótidos desde datos de alineamientos de secuencias de DNA. Este programa puede estimar una serie de medidas de variación de la secuencia de ADN, dentro de esta y entre las poblaciones (para regiones no codificantes, sitios sinónimos o no sinónimos, o en diversos tipos de posiciones de los codones), así como el desequilibrio de li-

gamiento, la recombinación, el flujo de genes y los parámetros de conversión de genes (Librado, 2009). La figura 5 muestra la página inicial del programa de DnaSP.

Figura 5. Página principal del software DnaSP 5.10. Fuente: http://www.ub.edu/dnasp/

Network. El programa Network se utiliza para el análisis intraespecífico y poblacional. Esto quiere decir que se relaciona con la filogeografía y la genética de poblaciones. El programa permite la elaboración de redes de haplotipos por medio de un conjunto de secuencias que están ordenadas en tablas. Muestra los sitios informativos y contiene datos importantes para la elaboración de los network. La página de Network está disponible en http://www.megasoftware.net

Medición de la biodiversidad

Los programas dirigidos a la medición de la biodiversidad utilizan bases de datos para recoger los nombres de las especies, así como de sus descripciones, distribuciones, información genética, estado y tamaños de las poblaciones, necesidades de su hábitat, y cómo cada organismo interactúa con otras especies. Un punto interesante es la simulación computacional que permite modelar aspectos tales como dinámica poblacional o calcular la mejora del acervo genético de una variedad o de alguna población amenazada.

Estimates. El programa Estimates es una herramienta muy útil para realizar curvas de acumulación y estimaciones de la riqueza esperada de acuerdo con modelos y métodos como CHAO 2, Jacknife y Bootstrap. Este programa toma los datos provenientes de un sistema de muestreo estandarizado; aleatoriza toda la información y realiza cálculos del número de especies observado y esperado, utilizando para ello estimadores y considerando las desviaciones estándar provenientes del proceso de aleatorización (Villareal et al., 2006). La página de Estimates está disponible en http://viceroy.eeb.uconn.edu/estimates/

Análisis de la expresión y regulación génica

Los software encargados del estudio de la expresión génica se enfocan en el estudio de los niveles de mRNA, que son medibles por una variedad de técnicas como microarrays de ADN, secuenciación de EST (Expressed Sequence Tag), análisis en serie de la expresión génica (Serial Analysis of Gene Expression - SAGE), MPSS (Massively Parallel Signature Sequencing) o diversos tipos de hibridación in situ. Los datos de expresión pueden usarse para inferir la regulación génica; podrían compararse datos de microarrays provenientes de una amplia variedad de estados de un organismo para formular hipótesis sobre los genes involucrados en cada situación o condición.

BRB array tools. Es un complemento de Excel-ins que combina R, C y Java para hacer los cálculos. Utiliza Excel para interactuar con el usuario, lo que significa que está disponible para Windows. Este programa permite la visualización y análisis estadístico de microarrays de expresión génica, el número de copias, la metilación y datos de RNA-Seq (Maestre, 2010). La página de BRB array tools está disponible en http://brb.nci.nih.gov/BRB-ArrayTools/

TM4. El programa TM4 está escrito en Java y se ejecuta en los sistemas Linux y Windows. Ofrece

capacidades de análisis de arrays, no sólo mediante el visor de varios experimentos (MeV), sino también el análisis de imágenes de arrays (Spotfinder), la normalización por separado de cada experimento (MIDAS) y un sistema de base de datos (MADAM) para almacenar los experimentos. La página de TM4 está disponible en http://www.tm4.org/

RNA seq UD. Como indica Ramirez (2015), es una plataforma bioinformática basada en Galaxy que compone las herramientas necesarias para todo el procesamiento de datos transcriptómicos en una interfaz web accesible, y que les permite a los investigadores disponer y gestionar archivos para su procesamiento, representando facilidad para enfocarse en el análisis de la información obtenida (trabajo que resulta dificultoso en algunos casos) sin necesidad de poseer conocimientos en el manejo de comandos por medio del servidor.

Su descarga está disponible en http://www.rna-seqblog.com/demo-galaxy-rna-seq-ud/

Predicción de la estructura de proteínas

Una de las más importantes aplicaciones de la bioinformática es la predicción de las estructuras de las proteínas, basándose en los diferentes niveles que se presentan: primaria, secundaria, terciaria, y cuaternaria. La estructura primaria se determina fácilmente con la secuencia de nucleótidos sobre el gen que codifica, aunque la dificultad ocurre con los tres últimos niveles. Los software plantean la heurística, las redes neuronales o el método ab initio como una forma de modelar estas proteínas; pero con el aumento de la cristalización o asignación de un espectro de masas de muchas proteínas se ha planteado la modelización por homología, es decir que, si dos proteínas presentan una misma secuencia, esta debe poseer una misma función. Por esta razón, se suele predecir la estructura de una proteína una vez conocida la estructura de una proteína homóloga.

I-TASSER (Iterative Threading ASSEmbly Refinement). El programa I-TASSER se enfoca en la predicción y organización de la estructura terciaria de las proteínas. Para ello, primero se identifica la proteína molde o de interés en PDB (protein data bank); luego se realiza el análisis por múltiples enfoques de enroscado o threading LOMETS (Local Meta-Threading-Server); posteriormente se construyen modelos completos de gran tamaño a partir de simulación de fragmentos de plantillas interactivas. Finalmente, las funciones de cada dominio se derivan de modelos 3D en la base de datos BioLip. La página de I-TASSER está disponible en http://zhanglab.ccmb.med.umich.edu/I-TASSER/

Swiss Model. Este servidor automático está basado en la predicción de estructuras terciarias y cuaternarias de proteínas por homología. La biblioteca de plantillas o moldes de proteínas se encuentran en la página principal de SWISS. Una vez descargadas las estructuras también se pueden obtener los ligandos esenciales y co-factores que permiten la construcción de modelos estructurales completos, incluyendo su estructura oligomérica (Biasini, 2014). La página de swiss-model está disponible en http://swissmodel.expasy.org/

Genómica comparativa

El núcleo del análisis comparativo del genoma es el establecimiento de la correspondencia entre genes (análisis ortólogo) o entre otras características genómicas de diferentes organismos. Esto permite conocer los procesos evolutivos responsables de la divergencia entre dos genomas o, mejor dicho, entre dos organismos.

Ensembl. Esta es una base de datos que contiene datos precalculados de genómica comparativa entre todas las especies analizadas en esta base de datos. Se puede observar análisis genómicos entre especies cercanas, no tan cercanas y lejanas, regiones de sintenia en base a alineamientos, familia de proteínas, alineamiento de proteínas, predicción de ortología/paralogía de proteínas, etcétera (Hubbard, 2002). La página de Ensembl está disponible en http://www.ensembl.org/index.html

Modelado de sistemas biológicos

El modelado de la biología de sistemas implica conocer el uso de simulaciones por ordenador de subsistemas celulares (redes de metabolitos y enzimas que comprenden el metabolismo, rutas de transducción de señales, y redes de regulación genética), tanto para analizar como para visualizar

las complejas conexiones de todos estos procesos celulares.

Cell Designer. Este programa es un editor de diagramas estructurados para la elaboración de redes génicas y bioquímicas. Las redes se dibujan en base al diagrama de proceso, con el sistema de notación gráfica propuesta por Kitano, y se almacenan utilizando los Sistemas de Biología Markup Language (SBML). CellDesigner apoya la simulación y la exploración del parámetro mediante una integración con SBML ODE Solver, simulación de núcleo SBML y COPASI. Mediante el uso de CellDesigner también se puede examinar y modificar modelos existentes SBML con referencias a las bases de datos existentes, simular y visualizar la dinámica a través de una interfaz gráfica interactiva (Funahashi, et.al, 2003). La figura 6 muestra la configuración de la página de Cell Designer, que está disponible en http://www.celldesigner.org/

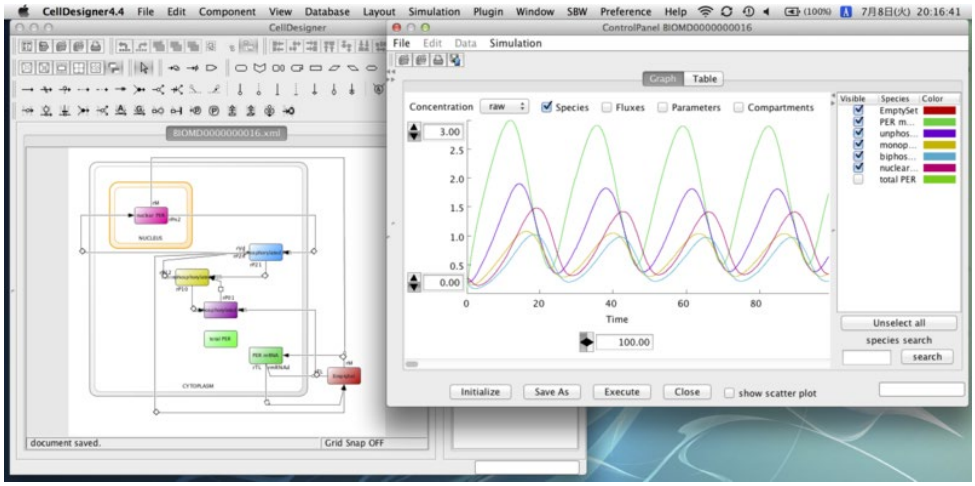

Figura 6. Página principal del software CellDesigner 4.4. Fuente: http://www.celldesigner.org/

Acoplamiento molecular

El acoplamiento o docking molecular estudia la interacción entre dos moléculas para predecir la fuerza de la asociación o la afinidad de enlace entre estas. La asociación entre moléculas biológicamente revelantes, tales como proteínas, ácidos nucleicos, carbohidratos y lípidos juega un papel central en la transducción de señales en las células. Por tanto, la orientación relativa del dúo interactuante puede afectar el tipo de señal producida. Por esta razón, el acoplamiento molecular adquiere importancia al predecir la fuerza y el tipo de señal que puede producir. Además, es utilizado para predecir la orientación del enlace de una molécula pequeña, que será candidato a fármaco, con la proteína donde ejercerán su acción, con lo que se podrá predecir la afinidad y la actividad de la molécula pequeña. Es por eso que este método tiene un rol muy importante en el diseño racional de fármacos. Dada la importancia biológica y farmacéutica, se han hecho grandes esfuerzos buscando mejorar el método usado para predecir el acoplamiento molecular.

Autodock. Este software es un conjunto de herramientas de conexión automatizadas. Está diseñado para predecir moléculas pequeñas, tales como sustratos o candidatos a fármacos, y unirse a un receptor de estructura 3D conocida. AutoDock 4 consta en realidad de dos programas principales: AutoDock, que realiza el acoplamiento del ligando a un conjunto de grids que describen la proteína diana y AutoGrid, que pre-calcula estas grids. Además de su uso para el acoplamiento, las grids de afinidad atómicas pueden ser visualizadas. Esto puede ayudar, por ejemplo, para guiar a los químicos orgánicos sintéticos a diseñar mejores aglutinantes. AutoDock Vina no requiere la elección de los tipos de átomos y mapas de la red de pre-cálculo para ellos. En lugar de ello, calcula los grid internos, para los tipos de átomos que se necesitan, y lo hace rápidamente. La página de Autodock está disponible en http://autodock.scripps.edu/

Conclusión

La bioinformática es una de las ramas que ha tenido un crecimiento constante y continuo, debido al desarrollo de software especializado para aplicación en líneas específicas relacionadas con las ciencias biológicas como son la Genética Médica,

la Bioquímica Clínica, la Farmacología, las Neurociencias, la Estadística Médica, la Inmunología, la Fisiología y la Oncología.

Una de las ventajas que brinda la descarga de cada software descrito es su libre acceso y en muchos casos, su gratuidad. Por lo tanto, los grupos bioinformáticos de los países en desarrollo, como es el caso de Perú, pueden tener acceso a potentes herramientas computacionales. Además, es posible configurar estos localmente en servidores propios, como los presentados en este artículo; por ejemplo: Clustal X, MEGA, DAMBE y NETWORK.

Si además se considera que la información biológica es en estos días de acceso libre y gratuito, la bioinformática resulta una disciplina altamente costo-efectiva y debe aprovecharse y explotarse. Por todo lo mencionado, articular un laboratorio básico de bioinformática, requiere actualmente solo de algunos espacios de trabajo de baja a mediana potencia, con una conexión a Internet de un ancho de banda adecuado, ideas innovadoras y un buen enfoque.

Es así que hay muchos ejemplos que ilustran el gran impacto de la bioinformática. Uno de ellos es el desarrollo de drogas mediante el uso de herramientas bioinformáticas que permiten disminuir el tiempo para desarrollar fármacos. Por ejemplo, fármacos que comúnmente se producen en un plazo de diez años, hoy pueden elaborarse en tres años. Esto se logra mediante el análisis del genoma del patógeno, la identificación de proteínas dianas (aquellas proteínas específicas del patógeno con participación en múltiples vías metabólicas importantes), además de realizar un screening in silico de una librería de miles de moléculas inocuas para la salud humana. Esto se puede realizar con el programa AutoDock. Una vez reducido el número de candidatos de manera significativa, se realizan recién las pruebas in vitro para identificar a los mejores candidatos. De esta manera se reducen también los gastos para todos estos estudios.

Otro enfoque que reduce los costos en la investigación es la utilización de software que predice la estructura de las proteínas, debido a que se evita el gasto realizado por los instrumentos de la resonancia magnética nuclear y la cristalografía de rayos X para representar a la proteína de interés. Asimismo, muchos de estos programas tienen una gran confiabilidad y presentan variabilidad en sus modelos de predicción, pudiendo escogerse el que mejor presente el modelo de la proteína según los distintos parámetros utilizados, para posteriores análisis. Hay que tener en claro que este modelado presenta ciertas limitaciones como una resolución de 1.5 a 3.5 Å y que el estudio depende de la similaridad que se encuentre con otras secuencias, resultando problemático si existe un bajo % de identidad o, de lo contrario, se utiliza varios secuencias moldes para alinear, generando mayor error en la resolución del modelo. Este artículo mostró programas de esta funcionalidad como Swiss-model e I-TASSER.

El uso de software libre en biología ha ido evolucionando de manera exponencial, esto debido a la minería de datos obtenida por investigadores vinculados a la biología, pero como se menciono es necesario trabajar esta información y solo puede realizarse con programas de alto nivel, diseñados por ingenieros y/o programadores conjuntamente con biólogos expertos en bioinformática. Por ello es necesario que las facultades de ingeniería de sistemas, software y/o programación establezcan una línea de investigación en bioinformática, por la importancia que se requiere en estos últimos años y el alto impacto en la investigación de ambas áreas.

Referencias bibiográficas

Alcalde-Alvites, M. (2014). Bioinformática: tecnologías de la información al servicio de la biología y otras ciencias. Hamut'ay, 1(2), 34-43.

Barahona, J. (2011). El concepto de software libre. Revista tradumàtica: traducció i tecnologies de la informació i la comunicació, (9), 5-11.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., y Schwede, T. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic acids research, gku340.

Cañedo, R., y Arencibia, J. (2004). Bioinformática: en busca de los secretos moleculares de la vida. Acimed, 12(6), 1-1.

Herrera, J. (s/n). El Software Libre en Bioinformática. Recuperado de: http://www.arareko.net/bioinformatics/free_software/index.pdf.es.pdf

Dias, D. (2011). Estrategia de solución al problema de la anotación de secuencias de ADN mediante la metodología

CommonKADS. Trabajo de fin de Master, Universidad Complutense de Madrid. Recuperado de: http://eprints.ucm.es/13062/1/TFM-IA-DanielaDiasXavier.pdf

Funahashi, A., Tanimura, N., Morohashi, M., & Kitano, H., (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks, BIOSILICO, 1 (5), 159-162. Recuperado de: http://openwetware.org/images/c/c0/Funahashi.pdf

Gilbert, D. (2004). Bioinformatics software resources. Briefings in bioinformatics, 5(3), 300-304.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., ... & Durbin, R. (2002). The Ensembl genome database project. Nucleic acids research, 30(1), 38-41.

Kumar, S., & Dudley, J. (2007). Bioinformatics software for biologists in the genomics era. Bioinformatics, 23(14), 1713-1717.

Kumar, S., Nei, M., Dudley, J., & Tamura, K. (2008). MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings in bioinformatics, 9(4), 299-306.

Librado, P. y Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452.

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. Methods of information in medicine, 40(4), 346-358.

Maestre, J. (2010). Análisis de datos de MicroArrays. Treball Final de Carrera en Enginyeria Informàtica. Universidad Politecnica de Valencia i Ibime (Informática Biomédica), 1-54.

Perezleo, L., Arencibia Jorge, R., Conill, C., Achón, G., y Araújo, J. (2003). Impacto de la bioinformática en las ciencias biomédicas. Acimed, 11(4), 0-0.

Ramirez, M., Rojas-Quintero, C. A., & Vera-Parra, N. E. (2015, June). RNA-Seq UD: A bioinformatics plattform for RNA-Seq analysis. In Information Systems and Technologies (CISTI), 2015 10th Iberian Conference on (pp. 1-5). IEEE.

Sala, H. E., & Núñez Pölcher, P. N. (2014). Software Libre y Acceso Abierto: dos formas de transferencia de tecnología. Revista Iberoamericana de Ciencia, Tecnología y Sociedad, 9(26),

Vera, S., Jiménez, P. y Franco-Lara, L. (2012) Uso de herramientas bioinformáticas en la evaluación de secuencias "dna barcode" para la identificación a nivel de especie. Revista de Facultad de ciencias básicas, 8 (2), 196-209. http://www.google.com.pe/url?url=http://www.umng.edu.co/documents/10162/3468353/ARTICULO3.pdf&rct=j&frm=1&q=&esrc=s&sa=U&ved=0ahUKEwiAyeb2oqrNAhVFLB4KHS8WAjIQFggTMAA&sig2=3nX5IZrrc5aR3gr8pQddKQ&usg=AFQjCNFdtBX7C1oF_qlYOiINf_g64Va9xQ

Villarreal, H., Álvarez, M., Córdoba, S., Escobar, F., Fagua, G., Gast, F., y Umaña, A. (2006). Métodos para el análisis de datos: una aplicación para resultados provenientes de caracterizaciones de biodiversidad. Manual de Métodos Para el Desarrollo de Inventarios de Biodiversidad. Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, Bogotá, Colombia, 185-226.

Young, K. M. (2000). Informatics for healthcare professionals. FA Davis Company.