Galo Ezequiel Balatti

Instituto de Física de Buenos Aires; Instituto de Química y Metabolismo del Fármaco; Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina.

Contacto: balatti@live.com / gbalatti@df.uba.ar

Nathalia M. V. Flórez-Zapata

Centro de Bioinformática y Biología Computacional de Colombia BIOS, Colombia; Universidad EIA, Colombia.

Contacto: nathalia.florez@eia.edu.co

Bioinformática para la creación y fortalecimiento de empresas de base tecnológica: conceptos y aplicaciones

Resumen

La bioinformática nace como una aproximación multidisciplinaria con el propósito de desarrollar técnicas de recolección, clasificación, almacenamiento y análisis de datos biológicos mediante la gestión de recursos computacionales. En los últimos años, ello generó un crecimiento exponencial de este tipo de datos, y el mayor desafío radica en convertir dicha información en recursos útiles para el sector académico e industrial. Así, la bioinformática ha logrado extenderse a la actividad comercial, donde empresas de base tecnológica (EBT) desarrollan y/o usufructan sus herramientas para ofrecer servicios de alto valor agregado. En el presente artículo describiremos la gama de herramientas disponibles para emprendedores tecnológicos y analizaremos el surgimiento de algunas de estas EBT.

Aunque no existe un consenso general sobre qué es la bioinformática, se la puede pensar como una disciplina que se encarga de la aplicación de técnicas computacionales para comprender y organizar la información biológica (Luscombe, Greenbaum y Gerstein, 2001). Dado que dicha información se encuentra principalmente almacenada en macromoléculas como el ADN, ARN y las proteínas, no es de extrañar que esta disciplina haya cobrado gran importancia conforme han ocurrido avances tecnológicos que permiten la obtención de información a partir de tales macromoléculas a menor costo, por ejemplo, mediante las tecnologías de secuenciación de nueva generación (NGS). No obstante, se acepta que la bioinformática nació en la década de los 50, mucho antes del auge de la secuenciación de ADN, con las contribuciones de Margaret Dayhoff en la aplicación de los primeros métodos computacionales para el análisis de las secuencias de proteínas (Gauthier, Vincent, Charette & Derome, 2018).

Si bien el origen de la bioinformática se remonta a varias décadas atrás, el interés comercial en ella es mucho más reciente. La generación masiva de datos biológicos y la necesidad de dar a estos un significado biológico aplicable ha generado la demanda de soluciones bioinformáticas y dado oportunidad a la creación de empresas biotecnológicas especializadas (Saviotti, Michelland & Catherine, 2000). En el año 2000 se pronosticaba que el mercado de la bioinformática sería comparable con la "fiebre del oro", en la que sobre todo las industrias farmacéuticas estarían interesadas en invertir su dinero para generar nuevos medicamentos a partir de la información genética derivada de la secuenciación del genoma humano y del auge de la era genómica (Howard, 2000). Durante esta época surgieron varias empresas especializadas que abarcaron distintos nichos, como el desarrollo de *software* bioinformático (p. ej. *Silicon Genetics*), la generación de plataformas de análisis como *Double Twist*, y provi-

sión de infraestructura computacional y experticia en bioinformática como *Lion BioScience* (Saviotti *et al*., 2000).

Aun cuando varias de estas primeras empresas bioinformáticas desaparecieron como resultado de su adquisición por compañías más grandes o por su incapacidad de responder a las necesidades del mercado y continuar siendo rentables (Knight, 2002), ello no implica que los emprendimientos en bioinformática sean poco lucrativos. De acuerdo con los pronósticos del mercado de la bioinformática realizados por *Markets and Markets™* (2018) se espera que para el 2023 este ascienda a los 13,5 mil millones de dólares a una tasa de crecimiento anual compuesto (CAGR, por sus siglas en inglés) del 14,5 %; las oportunidades del mercado que se destacan son la disponibilidad de datos y herramientas, el incremento en las colaboraciones e iniciativas que requieren de bioinformática, el surgimiento de nuevas tecnologías como el *blockchain* y cómputo en la nube, la integración de la inteligencia artificial y la llegada de inversión de grandes empresas del sector de las tecnologías de la información (TI). El reto es, entonces, poder encontrar un modelo de negocio para empresas en bioinformática o para EBT que utilicen a la bioinformática, que tenga en cuenta todas estas oportunidades del mercado y las materialice en productos o servicios de valor agregado que impacten en uno o varios sectores (p. ej. salud, académico, agrícola, ambiental).

Este capítulo recoge algunas de las herramientas más utilizadas en distintas aplicaciones de la bioinformática, así como un breve análisis del surgimiento de empresas de o relacionadas con la bioinformática, de manera que sea un punto de partida para todos aquellos interesados en generar un emprendimiento en esta área.

Datos biológicos: dónde encontrarlos

Como se ha mencionado, uno de los hechos que ha marcado el auge de la bioinformática ha sido el crecimiento exponencial de la información biológica generada. Según las últimas estadísticas del GeneBank —una de las principales bases de datos que alberga información de secuencias de nucleótidos—, desde 1982 al presente el número de bases depositadas en esta se ha duplicado cada 18 meses (National Center for Biotechnology Information, 2018). Las nuevas tecnologías de secuenciación han llevado a las ciencias biológicas a la era del *Big Data*, donde se proyecta que la cantidad de información generada será comparable con la de otros productores de *Big Data*, tales como la astronomía, Twitter y Youtube (Stephens *et al*., 2015). Adquirir, almacenar, distribuir y analizar tal cantidad de información supone grandes retos, como por ejemplo la necesidad de computadoras extremadamente poderosas; no obstante, también presenta grandes oportunidades (Marx, 2013). Los millones de datos depositados pueden ser utilizados para el desarrollo de nuevos productos basados en el conocimiento, y ahí es cuando la posibilidad de extraer y dar significado a esta gran cantidad de información disponible se convierte en una ventaja comparativa para las empresas del sector biotecnológico.

Las bases de datos biológicos pueden ser vistas como una fuente organizada de información biológica proveniente de experimentos y herramientas de análisis de información (Cannataro, Guzzi, Tradigo & Veltri, 2014). Por lo general se las clasifica de distintas maneras de acuerdo a los tipos de datos que pueden incluir (nucleótidos, proteínas, etc.) o según su fuente de información (primaria o secundaria). Las bases de datos *primarias* contienen la información básica de la secuencia de macromoléculas como nucleótidos y proteínas, mientras que las *secundarias* pueden contener un resumen de los resultados de los análisis del primer tipo de datos (Selzer, Marhöfer & Rohwer, 2008). También existen las *compuestas*, que pueden englobar varias bases de datos. Un ejemplo de lo anterior es Entrez, administrada por el Centro Nacional para la Información Biotecnológica (*National Center for Biotechnology Information* –NCBI-) de los Estados Unidos, que a su vez se conecta con cerca de 39 bases de datos que albergan seis categorías de información: literatura, datos clínicos, genomas, genes, proteínas y químicos (Agarwala *et al*., 2018).

En la actualidad, existen tres principales bases de datos de nucleótidos: el GeneBank, el EMBL-EBI y el DDBJ, las cuales son administradas por Estados Unidos, Europa y Japón, respectivamente. Entre 1986 y 1987 las tres establecieron una importante colaboración para estandarizar el formato de los datos que eran almacenados, definir la mínima información que debía ser depositada y facilitar el intercambio de información entre ellas (Gauthier *et al*., 2018). Dicha colaboración se conoce como la *International Nucleotide Sequence Database Collaboration* (INSDC), la cual hoy continúa vigente como una de las más importantes iniciativas en el intercambio de datos de dominio público (Karsch-Mizrachi, Takagi & Cochrane, 2018). Todos los días, la información depositada en alguna de estas bases de datos es intercambiada para asegurar la cobertura global de la misma (Benson *et al*., 2018). Este espíritu de generar iniciativas globales para hacer cambios recíprocos y estandarizar la distribución de información se ha mantenido en las ciencias biológicas y ha permitido el nacimiento de nuevas colaboraciones, como por ejemplo la *Global Alliance For Genomics and Health*,[1] la *Coordination of Standards In Metabolomics*[2] y el *International Molecular Exchange Consortium*,[3] entre otras. Lo anterior, sumado a la filosofía de muchos entes de financiación y revistas académicas de promover que los datos generados por la investigación científica sean de dominio público, ha llevado a que cada vez sea más fácil acceder a los datos biológicos, por lo que el reto sigue siendo poder analizar y obtener información relevante.

Un esfuerzo adicional que se ha hecho en busca de la estandarización e interoperabilidad (capacidad de intercambiar información) de las bases de datos es la utilización de ontologías para la descripción, organización, clasificación y contextualización de los datos biológicos generados (Schuurman & Leszczynski, 2008). Las ontologías proveen un vocabulario estandarizado que permite conceptualizar la información generada de manera que exista un conocimiento compartido de los objetos (por ejemplo, genes) que son estudiados (Stevens, Rector & Hull, 2010). Aunque existen distintas ontologías en las ciencias biológicas, una de las más utilizadas es el Gene Ontology (GO), mantenida por el *Gene Ontology Consortium*,[4] la cual ha sido utilizada para describir los genes en tres categorías: el proceso biológico en el que participan, la función molecular que desempeñan y el componente celular en el que se encuentran. Conocer la existencia de este vocabulario estandarizado es sin duda de gran ayuda para poder realizar análisis y búsquedas automatizadas.

Un aspecto fundamental a tener en cuenta en la selección de la fuente de información de datos biológicos a utilizar es si la base de datos elegida se encuentra o no *curada*. Dado que hoy en día casi cualquier persona puede depositar y compartir la información que genera, es importante que esta sea revisada y filtrada periódicamente por expertos: a este proceso se lo denomina "curación de datos". Aunque la curación de bases de datos puede demandar recursos económicos y humanos, es sumamente importante, puesto que la calidad de los datos no solo afectará las conclusiones derivadas de su análisis, sino que también hará lo propio con la posibilidad de integrarlos a datos preexistentes, e incluso con su accesibilidad y sostenibilidad (Odell, Lazo, Woodhouse, Hane & Sen, 2017). Un ejemplo de este tipo de bases curadas es la del *RefSeq*, que en 1999 empezó como un proyecto del NCBI para proveer a la comunidad de datos de genes revisados y actualizados, y que hoy alberga secuencias de referencia de eucariotas, procariotas y virus (Haft *et al*., 2018). Asimismo, las bases de datos específicas de organismos modelo como por ejemplo TAIR (Huala *et al*., 2001), Fly Base (Consortium, 2003) y MGD (Bult *et al*., 2008), que albergan información de *Arabidopsis thaliana*, *Drosophila melanogaster* y *Mus musculus*, respectivamente, también son bases de datos curadas. Finalmente, aquellas creadas con fines específicos —como PATRIC, que provee información de bacterias infecciosas (Wattam *et al*., 2013), o SILVA, especializada en secuencias de RNA ribosomal (Quast *et al*., 2012)—, son frecuentemente revisadas por expertos, por lo cual podrían catalogarse como bases de datos curadas.

La Tabla 1 presenta una selección de bases de datos biológicas muy utilizadas. La lista no es exhaustiva, por lo que sugerimos a los lectores interesados en conocer más acerca de bases de datos biológicas, consultar el número especial del año

2018 de la revista *Nucleic Acids Research* sobre bases de datos (Rigden & Fernández, 2018), así como algunas de las diversas revisiones que existen en el tema (Baxevanis & Bateman, 2015; Tang, Wang, Zhu & Zhao, 2015; Zhulin, 2015; Zou, Ma, Yu & Zhang, 2015).

Tabla 1. Bases de datos biológicas muy utilizadas

Base de datos	Descripción	Dirección web
Nucleótidos		
GenBank	Colección pública de secuencias de nucleótidos anotadas, administrada por el NCBI de los Estados Unidos.	https://www.ncbi.nlm.nih.gov/genbank
ENA EMBL-EBI	Repositorio de secuencias nucleotídicas y de procedimientos experimentales relacionados.	https://www.ebi.ac.uk/ena
DDBJ	Colección de secuencias de ADN administrada por el Instituto Nacional de Genética de Japón.	https://www.ddbj.nig.ac.jp/index-e.html
Proteínas		
Uniprot	Base de datos libre de secuencias e información funcional de proteínas.	https://www.uniprot.org
PDB	Base de datos abierta de estructuras 3D de proteínas y ácidos nucleicos, administrada por el Worldwide Protein Data Bank.	https://www.rcsb.org
Pfam	Base de datos del EMBL que recoge una colección de familias de proteínas, determinadas a partir de su información de secuencia.	https://pfam.xfam.org
Rutas metabólicas		
KEGG	Colección de bases de datos de genomas, rutas metabólicas, enfermedades y sustancias químicas. La visualización de información en la web es libre. Para la descarga de información debe pagarse suscripción.	https://www.genome.jp/kegg/

Metacyc	Base de datos curada que contiene información sobre rutas metabólicas dilucidadas experimentalmente.	https://metacyc.org
Reactome	Base de datos abierta y curada de rutas metabólicas. Contiene información sobre reacciones, proteínas y rutas metabólicas.	https://reactome.org/what-is-reactome
Genomas y metagenomas		
ENSEMBL ENSEMBL GENOMES	La base de datos ENSEMBL recoge información de secuencias y análisis de genomas de vertebrados. Por otra parte, ENSEMBL GENOMES alberga información de plantas y no vertebrados.	http://www.ensembl.org/index.html http://ensemblgenomes.org
GOLD	Base de datos de secuencias y proyectos de secuenciación de genomas y metagenomas.	https://gold.jgi.doe.gov
MGnify	Base de datos y herramientas web de análisis de datos provenientes de experimentos metagenómicos.	https://www.ebi.ac.uk/metagenomics/
MicrobiomeDB	Repositorio de datos y plataforma para el análisis de datos provenientes de microbiomas.	http://microbiomedb.org/mbio/app/
Otras		
RefSeq	Colección curada y no redundante de secuencias de referencia de ADN, transcritos y proteínas.	https://www.ncbi.nlm.nih.gov/refseq/
SRA	Repositorio de archivos de datos crudos de NGS.	https://www.ncbi.nlm.nih.gov/sra
DrugBank	Información detallada y profusa acerca de fármacos y sus interacciones moleculares	https://www.drugbank.ca/
ZINC	Colección con más de 230 millones de posibles ligandos para *docking*	http://zinc15.docking.org/

Aplicaciones bioinformáticas para el análisis de datos ómicos

Una vez descritas distintas fuentes de datos biológicos, cabe mencionar algunas de las herramientas y aplicaciones que puede tener la bioinformática en el análisis de dichos datos. Así, dentro de sus aplicaciones se destacan aquellas que giran en torno al análisis y desarrollo de herramientas para el estudio de datos provenientes de las "ómicas". Las ciencias ómicas pueden definirse como un conjunto de disciplinas que, a través de la generación masiva de datos por tecnologías de alto rendimiento, buscan el entendimiento de los sistemas vivos en términos de su composición y complejidad molecular (Palsson, 2002). Las más desarrolladas en los últimos años son aquellas asociadas al estudio de la información contenida en el ADN, ARN, proteínas y metabolitos, que se denominan genómica, transcriptómica, proteómica y metabolómica, respectivamente (Figura 1).

De acuerdo a lo anterior, una de las características de los datos biológicos es su heterogeneidad, lo cual supone que son diversas las herramientas que se han desarrollado para poder manipularlos y analizarlos. En bioinformática existen tanto herramientas libres como herramientas cuyo uso se debe pagar. Las de acceso libre ofrecen un sinnúmero de ventajas, incluyendo el ahorro de recursos financieros. No obstante, la mayoría de estas han sido desarrolladas para computadoras con sistemas operativos basados en Unix (p. ej. Linux), y en general carecen de una interfaz visual amigable para el usuario, lo cual puede ser intimidante para aquellos que no están habituados a la línea de comandos (Vincent & Charette, 2014). Por tanto, no es sorprendente que, para aquellos interesados en desarrollar habilidades en bioinformática, una de las principales recomendaciones sea familiarizarse con Unix y la línea de comandos (Dudley & Butte, 2009). Las herramientas presentadas en la Tabla 2 son principalmente libres; sin embargo, para los lectores interesados en las pagas, se sugiere consultar la revisión de Smith (Smith, 2014). Por otro lado, cabe aclarar que algunas de las herramientas aquí mencionadas y muchas otras pueden encontrarse en repositorios públicos como Bioconductor (Gentleman *et al*., 2004) o Bioconda (Grüning *et al*., 2018), por ejemplo, los cuales también pueden ser explorados por aquellos interesados en tener mayor información al respecto.

Como ya se mencionó, una de las características distintivas de los datos generados por las ciencias ómicas es su gran volumen, por lo que para el análisis de tal cantidad de información suelen re-

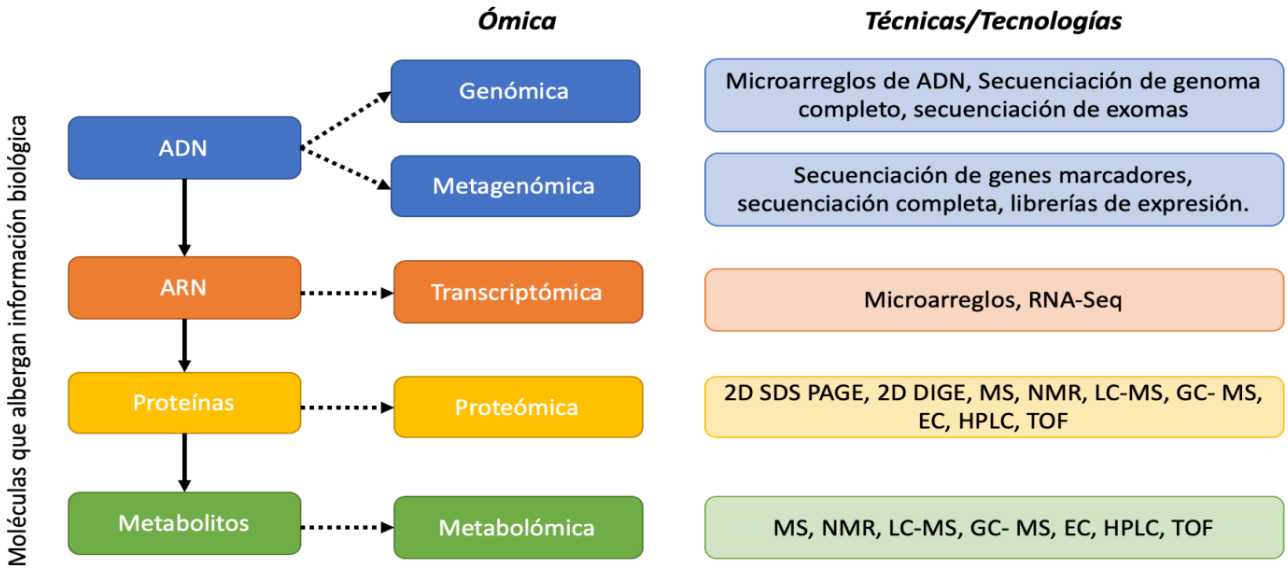

Figura 1. Fuentes de información biológica (diferenciadas por colores) y tecnologías utilizadas para su obtención en las principales ciencias ómicas.

querirse infraestructuras computacionales robustas (p. ej. *grids*, *clusters*). En general, los laboratorios de ciencias de la vida y las EBT no cuentan con dicha infraestructura, y es por esto que la computación en la nube (*cloud computing*) ha cobrado relevancia para la bioinformática en los últimos años (Balakrishnan & Soam, 2013). Dentro de las principales ventajas de este tipo de computación en bioinformática se destaca la conveniencia de no tener que invertir dinero en la adquisición y mantenimiento de la infraestructura, y la flexibilidad de poder pagar únicamente por la cantidad de recursos utilizados (Langmead & Nellore, 2018). Por ende, esta puede ser una alternativa a explorar para principiantes y emprendedores en bioinformática que no cuenten con una infraestructura computacional robusta.

Habitualmente, el primer paso en los análisis bioinformáticos es el control de calidad y preprocesamiento de los datos. Este control de calidad permite eliminar errores o el ruido que puede tener el set de datos, de manera que se reduzca el impacto negativo de estos sobre las conclusiones obtenidas (Cannataro, Guzzi, Mazza, Tradigo & Veltri, 2006; Del Fabbro, Scalabrin, Morgante & Giorgi, 2013). Algunas de las herramientas que pueden ser utilizadas para este propósito figuran en la Tabla 2. Los pasos a seguir *a posteriori* dependerán, entre otras cosas, de la pregunta que se pretenda responder con el análisis y la información previa que se tenga del organismo con el cual se está trabajando.

Cuando lo que se busca en este tipo de análisis es comprender la información contenida en el ADN de un organismo, de manera que permita hacerse una idea de qué genes están presentes, cómo estos se encuentran organizados, qué función pueden desempeñar y cómo dicho organismo ha evolucionado, la aproximación ómica más adecuada es la genómica. En general, la información del genoma es estudiada mediante la secuenciación del mismo (determinación del orden de los nucleótidos), lo cual genera pequeños fragmentos de información que tienen que ser armados a través de la bioinformática, en un proceso conocido como ensamblaje, con el propósito último de tener la representación continua y sin huecos (*gaps*) de su secuencia (Phillippy, 2017). No obstante, no siempre es posible tener un genoma completo, por lo cual se habla de borradores del genoma de distinta calidad y genomas terminados (Chain *et al*., 2009), que por su nivel de calidad delimitan el tipo de conclusiones a las que se puede arribar. Luego de armar el genoma por medio de herramientas bioinformáticas, este es anotado, lo cual supone identificar la posición de los genes en el genoma, predecir su función y su participación en distintos procesos biológicos (Stein, 2001). Finalmente, con la información de contenido y función del genoma, en muchos casos se decide compararlo con el de otros organismos (genómica comparativa), a fin de encontrar los cambios evolutivos que han ocurrido y sus consecuencias.

A diferencia de la genómica, la transcriptómica, la proteómica y la metabolómica proporcionan información global del conjunto de RNA, proteínas y metabolitos en un momento específico y bajo las condiciones particulares estudiadas. Es decir que transcriptoma, proteoma y metaboloma de un organismo pueden cambiar en respuesta a una condición fisiológica, un estímulo, una etapa de desarrollo, el tipo celular del organismo, entre otros. Aunque las herramientas utilizadas varían (Tabla 2), en términos generales los pasos de análisis de estas tres ómicas se asemejan, pues luego del preprocesamiento de los datos, se busca la identificación del transcrito, proteína o metabolito (según corresponda), la cuantificación que posibilita medir sus niveles y finalmente se realizan análisis estadísticos que permiten evaluar su representación o comportamiento diferencial en distintas muestras (De Bruin, Deelder & Palmblad, 2012; Kim, Kim, Kim, Hwang & Yoo, 2016; Yang & Kim, 2015).

Por último, corresponde resaltar que la bioinformática también ha tenido gran aplicación en la metagenómica, la cual está orientada a describir el material genético de una comunidad de microorganismos. Esta ómica ha cobrado gran importancia pues ha permitido entender cómo el microbioma humano participa en los estados de salud/enfermedad, estudiar microorganismos que no crecen en el laboratorio, dilucidar la participación de los microorganismos en el medio ambiente y explorar nuevos productos naturales provenientes de este tipo de organismos (Handelsman, Rondon, Brady,

Clardy, 2017; Martín, Miquel, Langella & Bermúdez-Humarán, 2014; Oulas *et al*., 2015; Tringe *et al,* 2005). Típicamente, el flujo de trabajo posterior al control de calidad que se sigue para analizar datos metagenómicos parte del ensamblaje y anotación de secuencias, y continúa con análisis taxonómicos y comparativos que permiten entender la composición y estructura de la comunidad (Ladoukakis, Kolisis & Chatziioannou, 2014). Existen varias herramientas que poseen flujos integrados de análisis metagenómicos; algunas de las más utilizadas figuran en la Tabla 2.

Tabla 2. Herramientas disponibles en las diferentes aplicaciones de la bioinformática

Herramienta	Descripción	Plataforma	Dirección web
Control de calidad y limpieza de secuencias			
FastQC	Permite revisar parámetros de calidad de datos de NGS.	Windows, Linux, MacOS	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trimmomatic	Permite el filtrado y cortado de secuencias obtenidas mediante la tecnología Illumina, según parámetros de calidad fijados.	Windows, Linux, MacOS	http://www.usadellab.org/cms/?page=trimmomatic
Blobology	Identificación de contaminantes provenientes de diferentes grupos taxonómicos en secuencias genómicas.	Linux, MacOS	https://github.com/blaxterlab/blobology
Genómica			
Ensamble SPAdes Velvet ABySS FALCON Canu	Diferentes programas de ensamble que permiten reconstruir la secuencia de genomas. FALCON y Canu son especiales para el ensamble de secuencias largas provenientes de las tecnologías PacBio y Oxford Nanopore	Linux, MacOS	http://cab.spbu.ru/software/spades/ https://www.ebi.ac.uk/~zerbino/velvet/ http://www.bcgsc.ca/platform/bioinfo/software/abyss https://pb-falcon.readthedocs.io/en/latest/ https://canu.readthedocs.io/en/latest/quick-start.html

QUAST	Permite evaluar la calidad de diferentes ensambles a partir de la estimación de métricas.	Linux, MacOS	http://bioinf.spbau.ru/quast
Augustus MAKER Prokka BLAST	Herramientas para identificación y predicción de genes, y para encontrar su posible función por homología (BLAST). Algunos son específicos de eucariotas (Augustus), mientras que otros están especializados en procariotas (Prokka).	Linux, MacOS, Windows (BLAST)	http://bioinf.uni-greifswald.de/augustus/ http://www.yandell-lab.org/software/maker.html http://www.vicbioinformatics.com/software.prokka.shtml https://blast.ncbi.nlm.nih.gov/Blast.cgi
Samtools Picard GATK Mauve FreeBayes	Este conjunto de herramientas permite el alineamiento, manipulación de archivos y llamado de variantes en algunos flujos de análisis de genómica comparativa.	Linux, MacOS, Windows (Mauve)	https://github.com/samtools/samtools https://broadinstitute.github.io/picard/ https://software.broadinstitute.org/gatk/ http://darlinglab.org/mauve/mauve.html https://github.com/ekg/freebayes
Metagenómica			
MG-RAST Mothur MEGAN QIIME	Flujos de trabajo y/o conjunto de herramientas para el análisis de datos metagenómicos que permiten la identificación de microorganismos y realización de pruebas estadísticas.	Linux, MacOS, Windows	https://www.mg-rast.org/ https://www.mothur.org/ http://ab.inf.uni-tuebingen.de/software/megan6/ http://qiime.org/

Transcriptómica			
BWA Bowtie2 STAR RSEM eXpress	Programas orientados al alineamiento de secuencias para su identificación, y/o cuantificación de dichos alineamientos para determinar niveles de expresión.	Linux, MacOS	https://github.com/lh3/bwa http://bowtie-bio.sourceforge.net/bowtie2/index.shtml https://github.com/alexdobin/STAR https://deweylab.github.io/RSEM/ https://pachterlab.github.io/eXpress/overview.html
edgeR DEGseq EBSeq	Estos son algunos de los paquetes disponibles en el *software* de cómputo estadístico R, que permiten identificar genes diferencialmente expresados.	Linux, MacOS, Windows	https://bioconductor.org/packages/release/bioc/html/edgeR.html https://bioconductor.org/packages/release/bioc/html/DEGseq.html https://bioconductor.org/packages/release/bioc/html/EBSeq.html
Enrichr PANTHER GOrilla	Este grupo de herramientas web apoya la definición de funciones de los genes a través de su anotación en términos GO y rutas metabólicas.	*On-line*	http://amp.pharm.mssm.edu/Enrichr/ http://www.pantherdb.org/downloads/index.jsp http://cbl-gorilla.cs.technion.ac.il
Proteómica y metabolómica			
ExPASy	Portal de recursos para el análisis de proteínas	*On-line*	https://www.expasy.org/tools/
MaxQuant	Permite el análisis de grandes sets de datos proteómicos obtenidos por espectrometría de masas.	Windows, Linux	https://maxquant.org/maxquant/
MetaboAnalyst	Conjunto de herramientas para el análisis de datos de metabolomas.	*On-line*	https://www.metaboanalyst.ca

XCMS	Paquete disponible en el *software* de cómputo estadístico R, que permite visualizar y procesar datos cromatográficos.	Linux, MacOS, Windows	https://bioconductor.org/packages/release/bioc/html/xcms.html
Edición y visualización molecular			
Avogadro	*Software* de visualización y edición molecular	Windows, MacOS, Linux	https://avogadro.cc/
SWISS-MODEL	*Software* para realizar modelado por homología	*On-line*	https://swissmodel.expasy.org/
VMD	*Software* de visualización con soporte para visualizar y analizar trayectorias de simulaciones	Windows, MacOS, Linux	https://www.ks.uiuc.edu/Research/vmd/
Diseño racional de fármacos			
McQSAR	Motor para cálculos de relación cuantitativa estructura-actividad	Windows, MacOS, Linux	http://users.abo.fi/mivainio/mcqsar/index.php
AutoDock	*Software* de *docking* molecular	Windows, MacOS, Linux	http://autodock.scripps.edu/
CHARMM-GUI	Interfaz *on-line* para la preparación de *inputs* de dinámica molecular	*On-line*	http://www.charmm-gui.org/
GROMACS NAMD	Suite de programas de simulación y análisis molecular de ME, MD, Montecarlo y otros.	Windows, MacOS, Linux	http://www.gromacs.org/ http://www.ks.uiuc.edu/Research/namd/

Bioinformática para el análisis de estructuras e interacciones moleculares

En forma análoga a lo ocurrido con las ómicas —y con la genómica en particular—, el advenimiento de las supercomputadoras y la enorme reducción en el costo de cómputo hicieron que el campo del modelado molecular creciera también en forma exponencial. El modelado molecular comprende las técnicas que permiten estudiar e interrelacionar la estructura y la función molecular mediante el uso de modelos matemáticos y de computadoras, a fin de imitar y reproducir el comportamiento de las moléculas. En su concepción más básica implica el uso de modelos físicos —como esferas y varillas plásticas— para

representar y estudiar la estructura en el nivel atómico (p. ej. los grados de libertad de una molécula). Con el advenimiento del cómputo de alta potencia, estos modelos lograron complejizarse a sistemas de miles de átomos. Así, el modelado molecular involucra desde la simple visualización de moléculas en tres dimensiones hasta los complejos modelos *ab initio* que permiten estudiar el comportamiento de electrones y simular reacciones químicas utilizando fundamentos de la mecánica cuántica (Schlick, 2010).

Para el modelado en tres dimensiones de moléculas sencillas de estructura desconocida (p. ej. ligandos o péptidos) se puede hacer uso de un *software* de edición molecular que tenga incorporado un algoritmo capaz de optimizar la geometría espacial de los átomos mediante una minimización de energía (ME). Estos programas utilizan modelos matemáticos llamados *campos de fuerzas* que describen la energía potencial del sistema. Una alternativa de libre distribución, uso sencillo y multiplataforma es el *software* Avogadro (Hanwell *et al*., 2012), que permite construir átomo por átomo un compuesto químico y luego asignarle una estructura mediante cálculos de optimización geométrica utilizando el campo de fuerza universal (UFF). Las moléculas o complejos que revisten mayor complejidad requieren técnicas más allá de la optimización geométrica: las estructuras de biomacromoléculas —como proteínas, ARN o complejos ADN-proteína— pueden obtenerse en forma experimental por medio de cristalografía de rayos X o resonancia magnética nuclear (RMN). Las estructuras obtenidas de esta forma se encuentran almacenadas en el *Protein Data Bank* (Berman *et al*., 2000). Además, proteínas o fragmentos de ellas cuya estructura se desconozca y que por la complejidad secuencial que posean no puedan ser optimizadas geométricamente *ab initio* pueden ser modeladas utilizando estructuras similares ya presentes en bases de datos. A esto se le llama *modelado por homología*, y es posible realizarlo con servicios como SWISS-MODEL (Waterhouse *et al*., 2018).

Uno de los sectores que hace una utilización profusa de este tipo de herramientas es la industria farmacéutica para el desarrollo de medicamentos. A pesar de ser un mercado dominado por grandes empresas, las *startups* están jugando un rol muy importante en la innovación: mientras las grandes compañías están disminuyendo sus niveles de innovación para concentrar sus recursos en el *marketing* y la producción de medicamentos, las pequeñas EBT farmacéuticas tienen una participación cada vez más importante en el descubrimiento de nuevas drogas. Al ocupar aquel lugar vacante, las *startups* (algunas incluso incubadas o financiadas por las grandes farmacéuticas) funcionan en simbiosis con los mayores jugadores del mercado (Dean, Zanders & Bailey, 2001; Holgersson, Phan y Hedner, 2016). También el sector público contribuye a la innovación a través de los laboratorios de investigación, con el aporte de nuevas moléculas de interés (Stevens *et al*., 2011). Así, la innovación del sector pasa ante todo por el descubrimiento y la caracterización de nuevos compuestos químicos. Históricamente este proceso se llevó adelante por medio de métodos de *screening* convencionales, que evaluaban experimentalmente grandes sets de sustancias, naturales o sintéticas, con el fin de encontrar efectos terapéuticos novedosos. Las sustancias halladas de esta manera (moléculas prototipo), con potencial para cierta área terapéutica, sufrían luego las modificaciones químicas necesarias para mantener o potenciar la actividad biológica sin comprometer la toxicidad (Gago, 1994). Si bien este método de muestreo fue el predominante durante décadas, por su naturaleza empírica su eficiencia era, a todas luces, muy baja. Se estima que solo una sustancia entre 10000 ensayadas conducía a un descubrimiento de relevancia (Sheridan & Venkataraghavan, 1987). Con el objeto de utilizar métodos más racionales para el diseño de fármacos surgió una de las primeras aproximaciones de *screening virtual (VS)*, el *Quantitative Structure-Activity Relationships* (QSAR) (Hansch & Fujita, 1964). QSAR construye modelos matemáticos que correlacionan con validez estadística descriptores fisicoquímicos de una estructura con la actividad o toxicidad de la molécula. McQSAR (Vainio & Johnson, 2005) es una implementación multiplataforma de estos algoritmos con interfaz de línea de comandos que genera soluciones utilizando los descriptores provistos por el usuario.

Otra herramienta relativa al VS de fármacos es el reconocimiento o *docking* molecular. Utilizando

una molécula blanco, un conjunto de potenciales ligandos puede ser ensayado tridimensionalmente dentro del sitio de unión de dicha molécula. A su vez, los ligandos pueden adoptar diversas orientaciones y conformaciones. Luego de realizar todas las combinaciones proteína-ligando relevantes, el algoritmo devuelve las soluciones óptimas, asignando una puntuación (*score*) basada en la energía de unión al ligando de los complejos proteína-ligando más estables (Meng, Zhang, Mezei & Cui, 2011). Así, el *docking* es una herramienta muy poderosa en el descubrimiento de fármacos a la hora de investigar cómo un ligando interacciona con un receptor de superficie, o para la evaluación de una serie de drogas diseñadas *de novo*, por ejemplo. Autodock es una *suite* de herramientas de *docking* de licencia libre muy poderosa y sencilla de utilizar. Si no se cuenta con posibles ligandos, la base de datos ZINC es una colección de libre acceso con más de 230 millones de compuestos de disponibilidad comercial para ser ensayados en moléculas blanco mediante *docking* (Sterling & Irwin, 2015).

Por último, para poseer la descripción más refinada del sistema molecular y explorar en detalle las interacciones moleculares de sus componentes entre sí (p. ej. para validar resultados de *docking*), las simulaciones permiten modelar con detalle atómico el comportamiento dinámico microscópico de prácticamente cualquier estructura química. Este tipo de simulaciones incluye tanto a las mencionadas ME utilizadas para optimización geométrica, como a las simulaciones de dinámica molecular (DM), donde las ecuaciones de mecánica clásica de Newton son resueltas en forma numérica (Tuckerman & Martyna, 2000). Debido a que es un modelo tan general, en el cual debe calcularse la energía potencial para cada par de átomos del sistema en cada instante de tiempo, su costo computacional es muy elevado. Ello resulta en una limitación respecto al tamaño de las moléculas simuladas y al tiempo de simulación total. Es por esta razón que en sistemas muy extensos se puede recurrir a métodos "de grano grueso", en los cuales los campos de fuerzas consisten en varios átomos agrupados en un solo sitio de interacción. Dentro de ellos, el campo de fuerzas MARTINI ha logrado representar con éxito sistemas como grandes complejos proteína-ADN, polímeros, nanopartículas e incluso membranas eucariotas complejas. El sitio CHARMM-GUI (Jo, Kim, Iyer & Im, 2008) permite, mediante una interfaz web sencilla, generar sistemas y archivos para realizar simulaciones de dinámica molecular en *suite* de programas de DM como GROMACS, de distribución libre y código abierto (Van Der Spoel *et al*., 2005) o NAMD (Philips *et al*., 2005), de similares características. Ambas opciones también poseen entre sus utilidades programas de análisis de las trayectorias obtenidas de la simulación. Para visualizar tales trayectorias, un *software* ampliamente extendido y de licencia gratuita es el VMD (Humphrey, Dalke & Schulten, 1996). Las simulaciones de Montecarlo (MC) son otro tipo de aproximación computacional, pero de tipo probabilístico, no relevantes a los fines de este trabajo.

Al uso de técnicas de VS en el descubrimiento de fármacos se le suma la influencia cada vez más grande de la inteligencia artificial (IA). Grandes compañías como Sanofi, Roche, Genentech, Merck, Bayer o Glaxo recurren al servicio de *startups* de IA para aplicaciones que van desde la búsqueda de blancos terapéuticos para oncología o biomarcadores para la gripe, hasta el monitoreo de pacientes (Mak & Pichika, 2018). También empresas informáticas como IBM, a través de su sistema WATSON, colabora con farmacéuticas como Pzifer o Novartis para investigar nuevas estrategias terapéuticas en oncología. Por ejemplo, WATSON ha logrado, mediante el análisis semántico de bibliografía científica, encontrar una serie de proteínas de unión a ARN que se encuentran alteradas en la esclerosis lateral amiotrófica (ELA) (Bakkar *et al*., 2018). Las técnicas de IA como el aprendizaje automatizado (*machine learning)* y su derivado, el aprendizaje profundo (*deep learning)*, pueden ser aplicadas a lo largo de todo el proceso de desarrollo de un fármaco, y, de hecho, se están utilizando en áreas como predicción de ADMET,[5] estudios de drogabilidad de receptores, búsqueda y validación de blancos, diseño *de novo* y rediseño de drogas, análisis de bibliografía, análisis de pacientes para ensayos clínicos, farmacovigilancia, entre otras (Mak & Pichika, 2018). A pesar de ser un campo de aplicación novedoso (en la actualidad no se comercializan medicamentos desarrollados mediante IA) el uso de IA promete disminuir en

forma considerable los grandes costos que implica atravesar las fases de desarrollo de un medicamento (Fleming, 2018).

La información fisicoquímica sobre drogas ya descubiertas, así como su farmacodinámica y otros datos útiles, puede encontrarse en la base de datos *drugbank* (Wishart *et al*., 2017). Otras herramientas alternativas o complementarias para el descubrimiento de fármacos no tratadas aquí, están compiladas en la página del Instituto Suizo de Bioinformática denominada *Click2Drug.*[6]

Un aspecto adicional del modelado molecular con gran potencial comercial es la impresión tridimensional (3D). La *startup* Organovo[7] se dedica al diseño y fabricación de tejidos humanos mediante impresión 3D para ser utilizadas en *screening* y ensayos de toxicidad de drogas (Vaidya, 2015). En la misma línea, n3D Biosciences[8] utiliza la misma técnica como soporte del cultivo celular tradicional, y ha logrado así reproducir con éxito modelos físicos tridimensionales de tejidos cancerosos (Jaganathan *et al.*, 2014).

EBTs en bioinformática: algunos casos de estudio

Genómica y blockchain: Nebula Genomics

Nebula Genomics[9] es una empresa de genómica aplicada a la medicina personalizada fundada por tres profesionales del área de la biomedicina y la informática. Por un lado, con solo una muestra de sangre extraída con un kit preparado para tal fin, la empresa ofrece un informe genotípico exhaustivo a fin de personalizar los tratamientos de los pacientes y conocer el riesgo de padecer posibles enfermedades, así como proporcionarle al cliente datos acerca de su genealogía. Por otra parte, ofrece un servicio de información genética para equipos de investigación y empresas de biotecnología. Dicha información genética es aportada por los mismos clientes. Nebula proporciona reportes gratuitos a cambio de la información recabada del genoma y de la respuesta que da el interesado acerca de su salud a fin de crear un perfil que pueda ser asociado a marcadores genéticos (enfermedades existentes, consumo de alcohol y tabaco, calidad de sueño, etc.). Según la empresa, la privacidad de los datos obtenidos de los usuarios se encuentra protegida por hacer uso de la tecnología *blockchain*. Esta permite prescindir de un nodo centralizado que valide el acceso a la información y conecta así al poseedor de la información con quien la compra (en general, investigadores científicos). Para la secuenciación utilizan NGS y secuencian la totalidad del genoma. El emprendimiento está financiado fundamentalmente por capitales de riesgo (*venture capitals)* y tiene un fondeo de 4,3 millones de dólares (Crunchbase, 2019).

Inteligencia Artificial aplicada: BenevolentAI

BenevolentAI es una empresa con sede en Londres fundada en 2013 por Ken Mulvany, un reconocido emprendedor del sector biotecnológico y de salud (Crunchbase, 2019). La compañía desarrolla algoritmos de IA, *machine learning* y *big data* para analizar información científica y clínica con el objetivo de hallar nuevas drogas y terapias contra enfermedades como el síndrome de Parkinson, glioblastoma o sarcopenia. Actualmente está en cartera el desarrollo de más de 20 drogas (Mulvany, 2018). Según su página web, BenevolentAI (2019) es “la única compañía de IA con capacidad de participar en todo el proceso de desarrollo de un fármaco: desde el descubrimiento temprano hasta la última etapa de ensayos clínicos”. La idea generó enormes expectativas. Tan es así que la firma se ha financiado con rondas de inversión que incluyen aportes de capitales de riesgo así como de grandes grupos de inversión como Goldman Sachs, logrando recaudar más de 207 millones de dólares (Crunchbase, 2019). Su valuación es tal (alrededor de 2 mil millones de dólares) que es considerada una *startup* de tipo “unicornio” (mote reservado a empresas cuya valuación supera los mil millones de dólares).

Computación en la nube: Seven Bridges

Seven Bridges[10] es una compañía de análisis de datos (*data analytics)* con sede en los Estados Unidos

fundada en 2009 por Deniz Kural (biólogo computacional) e Igor Bogicevic (ingeniero de *software*). La empresa ofrece una plataforma integral de análisis bioinformático con aplicación para medicina personalizada que incluye sets de datos, aplicaciones de análisis y la infraestructura de *cloud computing* integrados en una interfaz orientada al trabajo científico colaborativo. Esta compañía se encuentra en fase de *Early Stage Funding*, habiendo logrado en 2018 una ronda de inversión de tipo "B" y un fondeo total de 98 millones de dólares (Crunchbase, 2019).

Bioinformática en Latinoamérica: Neoprospecta

Neoprospecta[11] es una empresa brasileña fundada en 2011 por un grupo de doctorandos de la Universidade Federal de Rio Grande do Sul (UFRGS) que buscaban desarrollar un método sencillo de identificación de microorganismos aplicando secuenciación y algoritmos bioinformáticos. Al comienzo, y contando solo con un plan de negocios, recibieron el *Prémio para a Inovação e Emprendedorismo* del Banco Santander S. A., y al año siguiente el *Premio Iberoamericano a la Innovación y el Emprendimiento* de la Secretaría General Iberoamericana (Da Silveira, 2014). Luego lograron financiarse con inversores *ángel* en 2013, y con un capital semilla en 2014 para superar el millón de dólares de fondeo (Crunchbase, 2019). Hoy su servicio de diagnóstico microbiológico digital, que incluye el análisis bioinformático y el procesamiento de los datos genéticos de bacterias, virus, hongos y protozoos, es ofrecido a los sectores de salud, alimentos, agricultura, desarrollo farmacéutico y veterinaria (Neoprospecta, 2019).

Discusión y conclusiones

La revolución digital que comenzó a mediados del siglo XX con la proliferación de las primeras computadoras digitales y continúa en la actualidad con la popularización de internet, la hiperconectividad y la aparición de tecnologías emergentes de gran potencial como *blockchain*, comparte similares características con la explosión del sector biotecnológico, que empezó a mediados de los años 70 con los primeros experimentos en ingeniería genética y trajo aparejada la creación de la primera *startup* biotecnológica, la icónica Genentech.

Este fenómeno condujo a la creación de una gran cantidad de empresas, a tal punto que existen índices bursátiles especializados como el NBI (NASDAQ Biotech Index). Sin embargo, la biotecnología aún despierta grandes expectativas en el sector salud, propiciadas en buena medida por los avances logrados en bioinformática, sumados a una disminución abrupta de los costos y alcances de secuenciación, que suponen gran cantidad de información biológica disponible.

De esta manera, la bioinformática —que procesa enormes cantidades de información biológica gracias al soporte tecnológico de la computación— se convierte en una disciplina con gran potencial para la generación de iniciativas tecnológicas de alto valor agregado y capital intelectual. En efecto, hemos mencionado algunos casos de éxito en los cuales empresas con tan solo una idea y el conocimiento científico-tecnológico pertinente como único capital, obtienen financiamiento millonario en poco más de un año (y en ocasiones, sin sacar un solo producto al mercado). De esta manera, las herramientas bioinformáticas, que se encuentran en constante evolución desde el punto de vista académico, ofrecen un sinnúmero de oportunidades desde el punto de vista comercial.

Latinoamérica, región con dificultades a la hora de innovar debido a su inestabilidad macroeconómica, entre otras razones (Acs & Amorós, 2008), y caracterizada por ser exportadora de materias primas, posee una fuerte ventaja estratégica que reside en su gran diversidad biológica, un recurso muy apreciado en un mundo donde la biodiversidad disminuye a un ritmo alarmante. Este tipo de recursos estratégicos solo pueden explotarse mediante el conocimiento científico de alto nivel. Por tomar un caso, el centro de Bioinformática y Biología Computacional de Colombia (BIOS) cuenta con una importante infraestructura computacional, y surgió como medio para la búsqueda y análisis de secuencias genéticas de la megadiversidad colombiana con mi-

ras a su uso sostenible. La creación de BIOS se debió a una iniciativa de la empresa norteamericana Microsoft en un acuerdo con el Estado colombiano (BIOS, 2018). Otras iniciativas, como la argentina Biocódices, especializada en medicina personalizada, fueron incubadas en su totalidad por el sector público, a través de la Universidad de Buenos Aires y del Ministerio de Ciencia, Tecnología e Innovación Productiva (Biocódices, 2018). Estos casos, además de los puntualizados en el texto, indican que con el financiamiento adecuado —sea de fuentes privadas, estatales o mixtas—, la región tiene un enorme potencial para el surgimiento de emprendimientos de base tecnológica basados en bioinformática.

El acceso público a gran parte de la información biológica, así como a las herramientas que permiten su análisis, puede ser un punto de partida o bien una fuente de inspiración para cualquier emprendedor biotecnológico. El uso de estas herramientas puede complementar un servicio o producto de origen biotecnológico, ser útil en la búsqueda de nuevos nichos de explotación comercial, o bien servir de base para el desarrollo de nuevas aplicaciones bioinformáticas.

Notas

[1] Más información en la URL: https://www.ga4gh.org

[2] Más información en la URL: http://cosmos-fp7.eu

[3] Más información en la URL: http://www.imexconsortium.org

[4] Más información en la URL: http://geneontology.org

[5] Acrónimo en inglés para referirse a la absorción, distribución, metabolismo, excreción y toxicidad de un fármaco.

[6] Más información en la URL: https://www.click2drug.org

[7] Más información en la URL: https://organovo.com

[8] Más información en la URL: http://www.n3dbio.com

[9] Más información en la URL: https://www.nebula.org

[10] Más información en la URL: https://www.sevenbridges.com

[11] Más información en la URL: https://neoprospecta.com

Referencias bibliográficas

Acs, Z. J. & Amorós, J. E. (2008). Entrepreneurship and competitiveness dynamics in Latin America. *Small Business Economics*, *31*(3), 305-322 DOI: https://doi.org/10.1007/s11187-008-9133-y

Agarwala, R.; Barrett, T.; Beck, J.; Benson, D. A.; Bollin, C.; Bolton, E.; Zbicz, K. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *46*(D1), D8–D13. DOI: https://doi.org/10.1093/nar/gkx1095

Bakkar, N.; Kovalik, T.; Lorenzini, I.; Spangler, S.; Lacoste, A.; Sponaugle, K.; Bowser, R. (2018). Artificial intelli-

gence in neurodegenerative disease research: use of IBM Watson to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis. *Acta neuropathologica*, *135*(2), 227-247. DOI: https://doi.org/10.1007/s00401-017-1785-8

Balakrishnan, M. & Soam, S. (2013). Cloud Computing Technologies and its Applications in Bioinformatics. In: Roy, A. K. (Ed). *Information and Knowledge Management tools, Techniques and Practices* (pp. 93-100). New Delhi: New India Publishing Agency.

Baxevanis, A. D. & Bateman, A. (2015). The importance of biological databases in biological discovery. *Current Protocols in Bioinformatics*, *2015* (September), 1.1.1-1.1.8. DOI: https://doi.org/10.1002/0471250953.bi0101s50

Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Ostell, J.; Pruitt, K. D. & Sayers, E. W. (2018). GenBank. *Nucleic Acids Research*, *46*(D1), D41–D47. DOI: https://doi.org/10.1093/nar/gkx1094

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H. & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, *28*(1), 235-242.

Biocódices (2018). Web institucional. Recuperado (marzo de 2019) de: https://www.biocodices.com/

BIOS (2018). Web institucional. Recuperado (marzo de 2019) de: http://bios.co/

Bult, C. J.; Eppig, J. T.; Kadin, J. A.; Richardson, J. E.; Blake, J. A. & Mouse Genome Database Group. (2008). The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic acids research*, *36*(suppl_1), D724-D728. DOI: https://dx.doi.org/10.1093%2Fnar%2Fgkm961

Cannataro, M.; Guzzi, P. H.; Mazza, T.; Tradigo, G. & Veltri, P. (2006). On the Preprocessing of Mass Spectrometry Proteomics Data. In B. Apolloni, M. Marinaro, G. Nicosia & R. Tagliaferri (Eds.). *Neural Nets* (pp. 127–131). Berlin, Heidelberg: Springer Berlin Heidelberg.

Cannataro, M.; Guzzi, P. H.; Tradigo, G. & Veltri, P. (2014). Biological Databases. In N. Kasabov (Ed.). *Springer Handbook of Bio-/Neuroinformatics* (pp. 431–440). Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-642-30574-0_26

Chain, P. S. G.; Grafham, D. V.; Fulton, R. S.; Fitzgerald, M. G.; Hostetler, J.; Muzny, D. & Detter, J. C. (2009). Genomics. Genome project standards in a new era of sequencing. *Science (New York, N. Y.)*, *326*(5950), 236–237. DOI: https://doi.org/10.1126/science.1180614

Consortium, F. (2003). The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Research*, *31*(1), 172–175.

Crunchbase (2019). Web institucional. Recuperado (marzo de 2019) de: https://www.crunchbase.com/

da Silveira (2014). Multiplicação dos alvos. *Pesquisa FAPESP*. Recuperado de: http://revistapesquisa.fapesp.br/2014/12/29/multiplicacao-dos-alvos/

De Bruin, J. S.; Deelder, A. M. & Palmblad, M. (2012). Scientific Workflow Management in Proteomics. *Molecular & Cellular Proteomics*, *11*(7), M111.010595. DOI: https://doi.org/10.1074/mcp.M111.010595

Del Fabbro, C.; Scalabrin, S.; Morgante, M. & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS ONE*, *8*(12), 1–13. DOI: https://doi.org/10.1371/journal.pone.0085024

Dean, P. M.; Zanders, E. D. & Bailey, D. S. (2001). Industrial-scale, genomics-based drug design and discovery. *TRENDS in Biotechnology*, *19*(8), 288-292.

Dudley, J. T. & Butte, A. J. (2009). A quick guide for developing effective bioinformatics programming skills. *PLoS Computational Biology*, *5*(12). DOI: https://doi.org/10.1371/journal.pcbi.1000589

Fleming, N. (2018, mayo 30). How artificial intelligence is changing drug discovery. *Nature*, *557*(7706), S55-S55.

Gago, F. (1994). Métodos computacionales de modelado molecular y diseño de fármacos. *Monografías de la Real Academia Nacional de Farmacia*. Recuperado de: https://www.analesranf.com/index.php/mono/article/view/338

Gauthier, J.; Vincent, A. T.; Charette, S. J. & Derome, N. (2018). A brief history of bioinformatics. *Briefings in Bioinformatics*, (June), 1–16. DOI: https://doi.org/10.1093/bib/bby063

Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S. & Hornik, K. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, *5*(10), R80. DOI: https://doi.org/10.1186/gb-2004-5-10-r80

Grüning, B.; Dale, R.; Sjödin, A.; Chapman, B. A.; Rowe, J.; Tomkins-Tinch, C. H. & Bioconda, T. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475. DOI: https://doi.org/10.1038/s41592-018-0046-7

Haft, D. H.; Di Cuccio, M.; Badretdin, A.; Brover, V.; Chetvernin, V.; O'Neill, K.; … Pruitt, K. D. (2018). RefSeq: An update on prokaryotic genome annotation and curation. *Nucleic Acids Research*, *46*(D1), D851–D860. DOI: https://doi.org/10.1093/nar/gkx1068

Handelsman J.; Rondon, M.; Brady, S.; Clardy, J. G. R. (2017). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry and Biology*, *5*(10). DOI: https://doi.org/10.1016/S1074-5521(98)90108-9

Hansch, C. & Fujita, T. (1964). p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, *86*(8), 1616-1626.

Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E. & Hutchison, G. R. (2012). Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of cheminformatics*, *4*(1), 17. DOI: https://doi.org/10.1186/1758-2946-4-17

Holgersson, M.; Phan, T. & Hedner, T. (2016). Entrepreneurial patent management in pharmaceutical startups. *Drug discovery today*, *21*(7), 1042-1045. DOI: https://doi.org/10.1016/j.drudis.2016.02.018

Howard, K. (2000). The bioinformatics gold rush. *Scientific American*, *283*(1), 58–63.

Huala, E.; Dickerman, A. W.; García-Hernández, M.; Weems, D.; Reiser, L.; La Fond, F.; ... & Mueller, L. A. (2001). The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic acids research*, *29*(1), 102-105.

Humphrey, W.; Dalke, A. & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), 33-38. DOI: https://doi.org/10.1016/0263-7855(96)00018-5

Jaganathan, H.; Gage, J.; Leonard, F.; Srinivasan, S.; Souza, G. R.; Dave, B. & Godin, B. (2014). Three-dimensional in vitro co-culture model of breast tumor using magnetic levitation. *Scientific reports*, *4*, 6468. DOI: https://doi.org/10.1038/srep06468

Jo, S.; Kim, T.; Iyer, V. G. & Im, W. (2008). CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of computational chemistry*, *29*(11), 1859-1865. DOI: https://doi.org/10.1002/jcc.20945

Karsch-Mizrachi, I.; Takagi, T. & Cochrane, G. (2018). The international nucleotide sequence database collaboration. *Nucleic Acids Research*, *46*(D1), D48–D51. DOI: https://doi.org/10.1093/nar/gkx1097

Kim, S. J.; Kim, S. H.; Kim, J. H.; Hwang, S. & Yoo, H. J. (2016). Understanding Metabolomics in Biomedical Research. *Endocrinology and Metabolism*, *31*(1), 7. DOI: https://doi.org/10.3803/EnM.2016.31.1.7

Knight, J. (2002). Software firm falls victim to shifting bioinformatics needs. *Nature*, *416*, 357.

Ladoukakis, E.; Kolisis, F. N. & Chatziioannou, A. A. (2014). Integrative workflows for metagenomic analysis. *Frontiers in Cell and Developmental Biology*, *2*(November), 1–11. DOI: https://doi.org/10.3389/fcell.2014.00070

Langmead, B. & Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics*, *19*(4), 208.

Luscombe, N. M.; Greenbaum, D. & Gerstein, M. (2001). What is bioinformatics? An introduction and overview. *Yearbook of Medical Informatics*, 83–100. DOI: https://doi.org/10.1053/j.ro.2009.03.010

Mak, K. K. & Pichika, M. R. (2018). Artificial intelligence in drug development: present status and future prospects. *Drug discovery today*. DOI: https://doi.org/10.1038/nrg.2017.113

Markets and Markets (2018). Bioinformatics Market. Recuperado de: https://www.marketsandmarkets.com/Market-Reports/bioinformatics-39.html.

Marx, V. (2013). The big challenges of big data. *Nature*, *498*(7453), 255–260. DOI: https://doi.org/10.1038/498255a

Martín, R.; Miquel, S.; Langella, P. & Bermúdez-Humarán, L. G. (2014). The role of metagenomics in understanding the human microbiome in health and disease. *Virulence*, *5*(3), 413–423. DOI: https://doi.org/10.4161/viru.27864

Meng, X. Y.; Zhang, H. X.; Mezei, M. & Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, *7*(2), 146-157.

Mulvany, K. (19 de abril de 2018). Entrevista realizada por Paul Sandle y Ben Hirschler; editada por Alison Williams y Susan Fenton. *Euronews.* Disponible en: https://www.euronews.com/2018/04/19/ai-drug-hunter-benevolentai-worth-2-billion-after-fund-raising

Odell, S. G.; Lazo, G. R.; Woodhouse, M. R.; Hane, D. L. & Sen, T. Z. (2017). The art of curation at a biological database: Principles and application. *Current Plant Biology*, *11–12* (November), 2–11. DOI: https://doi.org/10.1016/j.cpb.2017.11.001

Oulas, A.; Pavloudi, C.; Polymenakou, P.; Pavlopoulos, G. A.; Papanikolaou, N.; Kotoulas, G. & Iliopoulos, I. (2015). Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies. *Bioinformatics and Biology Insights*, *9*. DOI: https://doi.org/10.4137/BBI.S12462

Palsson, B. (2002). In silico biology through "omics." *Nature Biotechnology*, *20*(7), 649–650. DOI: https://doi.org/10.1038/nbt0702-649

Phillippy, A. M. (2017). New advances in sequence assembly, 0–3. DOI: https://doi.org/10.1101/gr.223057.117

Phillips, J. C. (*et al.*)(2005). Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, *26*(16), 1781-1802. DOI: https://doi.org/10.1002/jcc.20289

Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P. & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41* (D1), D590—D596.DOI: https://dx.doi.org/10.1093%2Fnar%2Fgks1219

Rigden, D. J. & Fernández, X. M. (2018). The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Research*, *46*(D1), D1–D7. DOI: https://doi.org/10.1093/nar/gkx1235

Saviotti, P. P.; Michelland, S. & Catherine, D. (2000). The changing marketplace of bioinformatics. *Nature biotechnology*, *18*(12), 1247. DOI: https://doi.org/10.1038/82351

Schlick, T. (2010). *Molecular modeling and simulation: an interdisciplinary guide: an interdisciplinary guide* (Vol. 21). Madrid: Springer Science & Business Media.

Schuurman, N. & Leszczynski, A. (2008). Ontologies for bioinformatics. *Bioinformatics and Biology Insights*, *2*, 187–200.

Selzer, P. M.; Marhöfer, R. J. & Rohwer, A. (Eds.). (2008). Biological Databases. In *Applied Bioinformatics: An Introduction* (pp. 45–74). Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-540-72800-9_3

Sheridan, R. P. & Venkataraghavan, R. (1987). New methods in computer-aided drug design. *Accounts of Chemical Research*, *20*(9), 322-329.

Smith, D. R. (2014). Buying in to bioinformatics: An introduction to commercial sequence analysis software. *Briefings in Bioinformatics*, *16*(4), 700–709. DOI: https://doi.org/10.1093/bib/bbu030

Stein, L. (2001). Genome annotation: from sequence to biology. *Nature Reviews Genetics*, *2*(7), 493. DOI: https://doi.org/10.1038/35080529

Stephens, Z. D.; Lee, S. Y.; Faghri, F.; Campbell, R. H.; Zhai, C.; Efron, M. J.; … Robinson, G. E. (2015). Big data: Astronomical or genomical? *PLoS Biology*, *13*(7), 1–11. DOI: https://doi.org/10.1371/journal.pbio.1002195

Sterling, T. & Irwin, J. J. (2015). ZINC 15–ligand discovery for everyone. *Journal of chemical information and modeling*, *55*(11), 2324-2337. DOI: https://doi.org/10.1021/acs.jcim.5b00559

Stevens, R.; Rector, A. & Hull, D. (2010). What is an ontology? *Ontogenesis*. Recuperado de: http://ontogenesis.knowledgeblog.org/66

Stevens, A. J.; Jensen, J. J.; Wyller, K.; Kilgore, P. C.; Chatterjee, S. & Rohrbaugh, M. L. (2011). The role of public-sector research in the discovery of drugs and vaccines. *New England Journal of Medicine*, *364*(6), 535-541. DOI: https://doi.org/10.1056/NEJMsa1008268

Tang, B.; Wang, Y.; Zhu, J. & Zhao, W. (2015). Web resources for model organism studies. *Genomics, Proteomics and Bioinformatics*, *13*(1), 64–68. DOI: https://doi.org/10.1016/j.gpb.2015.01.003

Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics*, *6*(11), 805–814. DOI: https://doi.org/10.1038/nrg1709

Tuckerman, M. E. & Martyna, G. J. (2000). Understanding modern molecular dynamics: Techniques and applications. *The Journal of Physical Chemistry B*, *104*(2), 159–178. DOI: https://doi.org/10.1021/jp992433y

Vincent, A. T. & Charette, S. J. (2014). Freedom in bioinformatics. *Frontiers in Genetics*, *5*, 259. DOI: https://doi.org/10.3389/fgene.2014.00259

Vaidya, M. (2015). Startups tout commercially 3D-printed tissue for drug screening. *Nature Medicine News, 21*(1), 2-3.

Vainio, M. J. & Johnson, M. S. (2005). McQSAR: A Multiconformational Quantitative Structure− Activity Relationship Engine Driven by Genetic Algorithms. *Journal of chemical information and modeling*, *45*(6), 1953-1961. DOI: https://doi.org/10.1021/ci0501847

Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E. & Berendsen, H. J. (2005). GROMACS: fast, flexible, and free. *Journal of computational chemistry*, *26*(16), 1701-1718. DOI: https://doi.org/10.1002/jcc.20291

Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R. & Lepore, R. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, *46*(W1), W296-W303. DOI: https://doi.org/10.1093/nar/gky427

Wattam, A. R.; Abraham, D.; Dalay, O.; Disz, T. L.; Driscoll, T.; Gabbard, J. L. & Machi, D. (2013). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic acids research*, *42*(D1), D581-D591. DOI: https://doi.org/10.1093/nar/gkt1099

Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R. & Assempour, N. (2017). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, *46*(D1), D1074-D1082. DOI: https://doi.org/10.1093/nar/gkx1037

Yang, I. S. & Kim, S. (2015). Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. *Genomics & Informatics*, *13*(4), 119–125. DOI: https://doi.org/10.5808/GI.2015.13.4.119

Zhulin, I. B. (2015). Databases for microbiologists. *Journal of Bacteriology*, *197*(15), 2458–2467. DOI: https://doi.org/10.1128/JB.00330-15

Zou, D.; Ma, L.; Yu, J. & Zhang, Z. (2015). Biological databases for human research. *Genomics, Proteomics and Bioinformatics*, *13*(1), 55–63. DOI: https://doi.org/10.1016/j.gpb.2015.01.006