NCBI: GENERALIDADES DEL REPOSITORIO Y BREVE DESCRIPCIÓN DE RECURSOS APLICABLES AL ESTUDIO DE LA FITOGENÉTICA

NCBI: REPOSITORY OVERVIEW AND BRIEF DESCRIPTION OF RESOURCES APPLICABLE TO THE STUDY OF PLANT GENETICS

Oswaldo Guzmán-López[1], Celeste Ricaño-Rodríguez[2], Daniela Luis-Yong[2] y Jorge Ricaño-Rodríguez[3]*

[1]Universidad Veracruzana (UV), Facultad de Ciencias Químicas, Coatzacoalcos, Veracruz, México. [2]UV, Centro de Investigación en Micología Aplicada, Xalapa, Veracruz, México. [3]UV, Centro de EcoAlfabetización y Diálogo de Saberes, Campus USBI, Xalapa, Veracruz, México.

**Autor de correspondencia* (jricano@uv.mx)

RESUMEN

El Centro Nacional para la Información Biotecnológica (NCBI) proporciona acceso a un conjunto de recursos computacionales que permiten el conocimiento de sistemas biológicos; además, en éste se incluyen diversos repositorios de secuencias genómicas y proteicas, citas y resúmenes científicos referentes a las ciencias naturales y de la salud. En virtud de lo anterior, el objetivo de este trabajo fue llevar a cabo una recopilación de generalidades del NCBI, tomando como base de organización del documento los siguientes elementos: el sistema Entrez, fuentes de información y actualización de literatura, incluyendo la base de datos de taxonomía, gestión de metadatos y expresión genética, gestión de colecciones de secuencias nucleotídicas, procesamiento de secuencias genómicas y proteicas. De manera complementaria, se describen algunas herramientas aplicables al estudio de la fitogenética que han apoyado el ensamble de genomas completos de plantas; además, se discuten de manera breve algunos resultados de investigaciones derivados del uso de recursos bioinformáticos como anotaciones genómicas, generación y organización de metadatos, estudios de expresión genética incluyendo transcriptómica, búsqueda de alineaciones nucleotídicas locales y globales, así como estructuración y modelación tridimensional de proteínas, entre otros.

Palabras clave: BLAST, GenBank, NCBI, PubMed, recursos fitogenéticos.

SUMMARY

The National Center for Biotechnology Information (NCBI) provides access to a set of computational resources that allow knowledge of biological systems; in addition, it includes various repositories of genomic and protein sequences, citations and scientific abstracts concerning the natural and health sciences. This work aimed to carry out a compilation of NCBI generalities, taking the following elements as the organizational basis of the document: the Entrez system, sources of information and updating of literature, including the taxonomy database, management of metadata and genetic expression, management of nucleotide sequence collections, processing of genomic and protein sequences. In addition, some useful tools applicable to the study of plant genetics that have supported the assembly of complete plant genomes are described, briefly discussing some research results derived from the use of bioinformatics resources such as genomic annotations, generation and organization of metadata, gene expression studies including transcriptomics, search for local and global nucleotide alignments, as well as structuring and three-dimensional modeling of proteins, among others.

Index words: BLAST, GenBank, NCBI, plant genetic resources, PubMed.

INTRODUCCIÓN

El Centro Nacional para la Información Biotecnológica (NCBI; National Center for Biotechnology Information) es un repositorio que pertenece a la Biblioteca Nacional de Medicina de los Estados Unidos de América (NLM; National Library of Medicine). Éste fue creado en 1988 para desarrollar sistemas de información en materia de biología molecular (Jenuth, 2000; Sayers *et al*., 2020) y mantiene la base de datos de GenBank, que contiene secuencias de ácidos nucleicos y proteínas (Benson *et al*., 2012) que a su vez, recibe información mediante la colaboración del Banco de Datos de ADN de Japón (DDBJ; DNA Data Bank of Japan) (Ogasawara *et al*., 2020) y la base de datos de secuencias de nucleótidos del Laboratorio Europeo de Biología Molecular (EMBL-EBI; The European Bioinformatics Institute). Por otra parte, el NCBI proporciona distintos sistemas de recuperación de información y recursos computacionales para el análisis de muchos otros tipos de sistemas biológicos, que se complementa con contribuciones de la comunidad científica en general.

Las bases de datos biológicas-moleculares a menudo contienen relaciones entre registros basadas en inferencias computacionales de similitud; con menos frecuencia, registran explícitamente datos que derivan de manera experimental (*e.g.* enlaces entre secuencias consideradas homólogas en bases de datos de proteínas y nucleótidos, así como genes que interactúan en una ruta metabólica) (Benson *et al*., 2012; Geer *et al*., 2010). Debido a la gran cantidad disponible de estos repositorios, la integración de herramientas para estudios moleculares de uso común, por ejemplo el portal de reagrupación de recursos denominado

Recibido: 27 de febrero de 2022
Aceptado: 27 de septiembre de 2022

DOI: https://doi.org/10.35196/rfm.2023.1.63

"Entrez" perteneciente al NCBI, se realiza caso por caso (Maglott *et al*., 2005; Sayers *et al*., 2009).

A través de lo anterior se lleva a cabo un intercambio de información para todo el conjunto de repositorios, integrando sus datos en las bibliotecas existentes, como es el caso de PubMed, cuya base de datos comprende más de 33 millones de citas de literatura biomédica de Medline, revistas de ciencias biológicas y libros en línea, además de que las citas suelen incluir enlaces a contenido de texto completo y sitios web de editores. Asimismo, PubChem (otra base de datos digna de mencionar) es considerada la colección más grande del mundo de información en materia de química de libre acceso. De esta manera, centralizar y vincular las bases de datos de biosistemas aumenta potencialmente su utilidad.

Tomando en cuenta lo anterior, es importante mencionar que el estudio de los recursos fitogenéticos no se excluye de la mayoría de las disciplinas bioinformáticas, pues las herramientas disponibles en el NCBI son capaces de llevar a cabo estudios *in silico*, como por ejemplo la identificación de regiones genómicas conservadas (marcadores moleculares) para la clasificación de especímenes vegetales (Jorrin-Novo, 2020; Morgante y Salamini, 2003; Rensink y Buell, 2005; Ricaño-Rodríguez *et al*., 2019); o bien, evaluar la calidad de la secuenciación de genomas en especies de plantas utilizando etiquetas de secuencia expresada (Shangguan *et al*., 2013).

El Cuadro 1 muestra la estadística de registros en el NCBI hasta septiembre de 2022, relacionados con el campo de la fitogenética. La Figura 1 resume las principales herramientas disponibles en el repositorio para el estudio genético-molecular de las especies (incluyendo el reino vegetal) (NCBI Resource Coordinators, 2015; 2016; 2018). Cabe destacar que el acceso a los recursos e información del NCBI es en línea y disponible a la comunidad científica.

A la luz de las consideraciones anteriores, este trabajo tiene como objetivo recopilar algunas de las características generales más destacables del NCBI, tomando como base organizacional del documento elementos como el sistema Entrez, fuentes de información y actualización de literatura incluyendo la base de datos de taxonomía, gestión de metadatos y expresión genética, gestión de colecciones de secuencias nucleotídicas, procesamiento de secuencias genómicas y proteicas. Asimismo, se presentan algunos ejemplos de investigaciones en materia de fitogenética que emplearían herramientas computacionales pertenecientes al repositorio, las cuales incluyen anotaciones genómicas derivadas de secuenciaciones parciales y masivas, generación y organización de metadatos, estudios de expresión genética y transcriptómica, búsquedas de alineaciones nucleotídicas locales y globales para la identificación molecular de las especies, así como estructuración y modelación en 3D de proteínas, por mencionar algunos ejemplos. De la misma manera, se citan tutoriales enfocados a la ejecución de estos algoritmos, los cuales se encuentran disponibles de manera detallada en la página de inicio correspondiente del NCBI, al igual que su respectivo manual del usuario.

El sistema Entrez; reagrupación de bases de datos y accesos conjuntos

Entrez es un sistema integrado de reagrupación de bases de datos que brinda acceso a un conjunto diverso de 35 repositorios, los cuales contienen más de 3,000 millones de registros (Sayers *et al*., 2020; Trumbly, 2022). Los enlaces al portal web generalizado se proporcionan en la página de búsqueda global de Entrez. Por mencionar algunos ejemplos representativos de especies vegetales, tan sólo para el caso de *Solanum lycopersicum* y *Arabidopsis thaliana* se encuentra información registrada en 25 y 28 repositorios, respectivamente. Actualmente existe un manual de ayuda que permite conocer las utilidades de programación del sistema en mención.

Hasta septiembre de 2022, este motor de búsqueda reagrupó más de 1,153,725 documentos académicos relacionados con plantas a través de PubMed. En el apartado de genómica (Genomes) existen por lo menos 2,845 proyectos de ensamblajes (Assembly) de genomas de distintas especies de plantas y microorganismos relacionados metabólicamente, *e.g. Oryza officinalis* (GCA_008326285.1)*, Secale cereale* (GCA_902687465.1) y *Prunus yedoensis* (GCA_005406145.1).

Asimismo, en PubMed se encuentran diversas publicaciones referentes a secuenciaciones de genomas completos y parciales (incluyendo shotguns, fragmentación y reagrupación de ADN), al igual que la generación de líneas transgénicas de especies vegetales resistentes al estrés abiótico, que en conjunto son de interés biotecnológico, agrícola y económico-social (*e.g. A. thaliana*, *Vitis vinifera*, *Sorghum bicolor*, *Zea mays*, *Glycine max*, *Theobroma cacao*, *Phoenix dactylifera*, *Solanum tuberosum*, *Brassica rapa*, *Cannabis sativa*, *Prunus persica*, *Citrus aurantifolia* y *Allium sativum*) (Michael y Jackson, 2013; Ricroch *et al*. 2022).

En este sentido, existe un repositorio independiente al NCBI denominado PLAZA que funge como punto de acceso para la genómica comparativa de plantas, ya que éste centraliza los datos genómicos producidos por diferentes iniciativas de secuenciación (Van Bel *et al*., 2022). En la Figura 2 se muestra una línea de tiempo que

Cuadro 1. Estadística de registros (Entrez) en el NCBI relacionados con recursos fitogenéticos. Actualizada a septiembre 2022.

Base de datos	Registros	Descripción
Literatura		
Catálogo NLM	14,933	Índice de colecciones NLM
Libros	9,833	Libros e informes
PubMed Central	875,509	Artículos de revistas de texto completo
PubMed	1,153,559	Resúmenes/citas científicos y médicos
Genes		
Perfiles de expresión génica	3,545,349	Expresión génica y perfiles moleculares
Genes	7,294,554	Recopilación sobre *loci* genéticos
Colecciones de secuencias de ADN	72,865	Conjuntos de secuencias de estudios filogenéticos y de poblaciones
Grupos de datos de expresión génica	225,929	Estudios sobre genómica funcional
Proteínas		
Dominios conservados	2,806	Dominios proteicos conservados
Grupos de proteínas idénticas	18,526,704	Secuencias de proteínas agrupadas por identidad
Estructuras	7,898	Estructuras biomoleculares *in silico*
Proteínas	30,107,259	Secuencias proteicas
Genomas		
Biocolecciones	16	Colecciones físicas de biodepositorios
Genomas	1,192	Proyectos de secuenciación de genomas de organismos
Ensamblajes	2,468	Información sobre ensamblajes genómicos
Bioproyectos	102,976	Proyectos biológicos que proporcionan datos a NCBI
Biomuestras	1,387,820	Descripciones de materiales de origen biológico
Nucleótidos	93,102,533	Secuencias de ADN y ARN
Archivos de lectura de secuencia (SRA)	1,570,076	Archivo de lectura de secuencias de ADN y ARN de alto rendimiento

reagrupa algunas de las principales especies vegetales cuyos genomas han sido secuenciados, con base en el número de lecturas (*contig*), hasta septiembre de 2022. Cabe destacar que se resaltan los genomas de los cultivos alimenticios, frutales o agroindustriales más importantes; asimismo, el incremento representativo de especies estudiadas se debe en gran medida a los recursos bioinformáticos disponibles en la actualidad, como la secuenciación de nueva generación, los genotipados por secuenciación, así como el desarrollo de herramientas y una mayor disponibilidad de bases de datos a través del NCBI.

En materia de fitogenética, también es posible encontrar más de 103,000 bioproyectos (BioProject, recopilación de datos biológicos relacionados con una sola iniciativa provenientes de una organización o consorcio, lo que proporciona a los usuarios un lugar único para encontrar enlaces a los diversos tipos de datos generados); igualmente, se han registrado cientos de miles de biomuestras (BioSample, base de datos que contiene descripciones de materiales de origen biológico utilizados en ensayos experimentales), incluyendo archivos de lectura de secuencia (SRA: Sequence Read Archive, repositorio de datos de secuenciación de alto rendimiento, cuyo acervo acepta información de todas las ramas de la vida, así como estudios metagenómicos y ambientales) (Cuadro 1).

También, hasta septiembre de 2022, se encuentran depositadas en el NCBI más de 93 millones de secuencias de nucleótidos representadas en su mayoría por moléculas de ADN, ARNm y miRNA, así como algunas biocolecciones [Biocollections, conjunto de metadatos para colecciones de cultivos, museos, herbarios y otras colecciones de historia

Figura 1. Principales herramientas disponibles en el NCBI para el estudio genético-molecular de las especies, incluyendo el Reino Plantae. El repositorio BioSystems se retiró del NCBI en marzo de 2022. La herramienta de Protein incluye visores gráficos. (NCBI Resource Coordinators, 2015; 2016; 2018).

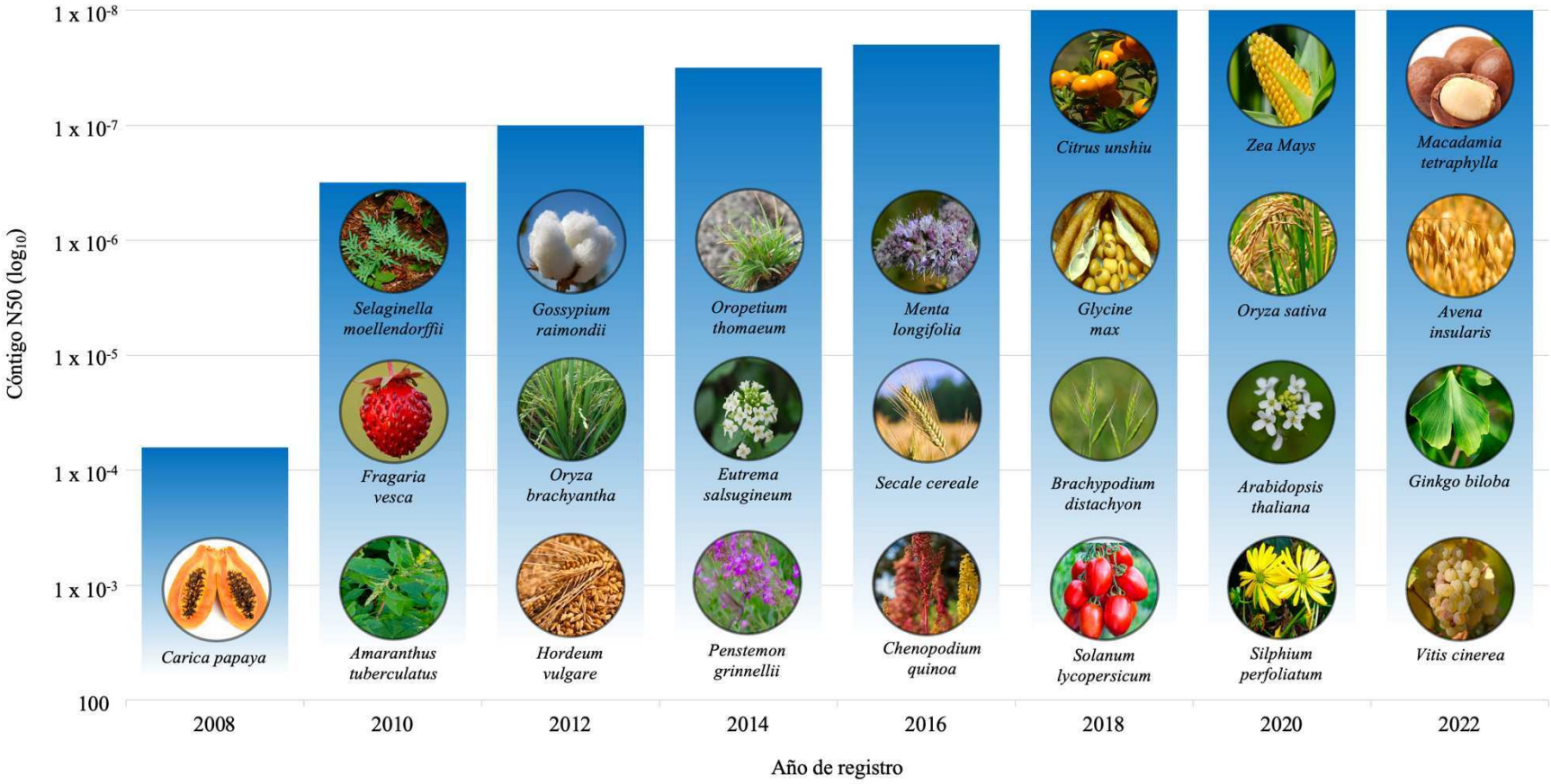

Figura 2. Línea de tiempo representativa (hasta septiembre de 2022) de las principales especies de origen vegetal cuyos genomas han sido secuenciados y ensamblados, y que fueron depositados en el NCBI. La gráfica muestra el año de registro (a partir de 2008), así como el número en aumento de lecturas (*contig*) N50 (log_{10}) en cada secuenciación.

natural, incluidos los códigos de colección e institución de Darwin Core (Wieczorek *et al*., 2012), al igual que fórmulas de URL para mapear identificaciones de especímenes en páginas web dentro del sitio respectivo].

En el Cuadro 2 se muestran algunas de las biocolecciones globales más importantes de plantas y microorganismos directamente relacionados, que han sido depositadas en el NCBI. De la misma manera, existen millones de perfiles de expresión génica (Gene Expression Omnibus), grupos de datos de expresión génica (PopSets, colección de secuencias de ADN relacionadas con estudios poblacionales, filogenéticos, mutacionales y de ecosistemas), al igual que secuencias de proteínas caracterizadas en su mayoría hipotéticamente (NCBI Resource Coordinators, 2015; 2016; 2018; Sayers *et al*., 2020) (Cuadro 1).

Fuentes de datos y actualización de literatura del NCBI

Las bases de datos y recursos del NCBI están organizados en siete áreas conceptuales: 1) literatura, 2) genomas, 3) variación, 4) salud, genes y expresión génica, 5) nucleótidos y proteínas, 6) moléculas pequeñas y 7) ensayos biológicos, las cuales dan origen a los 35 repositorios respectivos. Cada área conceptual comienza con un capítulo de descripción general que proporciona un marco contextual para los recursos discutidos bajo ese concepto. La descripción general va seguida de capítulos separados que cubren bases de datos o recursos individuales. Para más información respecto al apartado en mención es posible consultar una guía de literatura respectiva del NCBI (NCBI Resource Coordinators, 2015; 2016; 2018).

El repositorio genera esquemas de colaboración que

Cuadro 2. Estadística de registros de biocolecciones en el NCBI relacionadas con recursos fitogenéticos (septiembre 2022).

Institución	Tipo de colección	Calificador	ID
Colección de Plantas con Semillas Parásitas del Oeste de Virginia	Herbario	Comprobante de muestra	5612
Instituto de Plantas Medicinales ACECR de Irán	Herbario	Comprobante de muestra	8725
Centro de Investigación de Plantas Medicinales de la India	Herbario	Comprobante de muestra	8053
Colección Internacional de Microorganismos de Plantas de Nueva Zelanda	Cepario	Cepario	3871
Instituto Central de Plantas Medicinales y Aromáticas de la India	Herbario	Comprobante de muestra	4439
Instituto de Investigaciones de Plantas Medicinales, Departamento de Ciencias Médicas de Tailandia	Herbario	Comprobante de muestra	7374
Instituto de Investigación de Plantas Raras del río Yangtze, provincia de Hubei, China	Herbario	Comprobante de muestra	9122
Herbario de Plantas Vasculares Kathryn Kalmbach, Jardín Botánico de Denver	Herbario	Comprobante de muestra	4578
Instituto de Investigación de Plantas Medicinales y Aromáticas de toda Rusia	Herbario	Comprobante de muestra	4219
Laboratorio de Yunnan para la Conservación de Plantas Forestales Raras, Amenazadas y Endémicas	Herbario	Comprobante de muestra	7323
Universidad Estatal Mecynikov de Odessa, Departamento de Morfología y Sistemática de Plantas	Herbario	Comprobante de muestra	4925
Estación de Investigación de Plantas Medicinales del Instituto de Agroecología y Economía de los Recursos Naturales de Ucrania	Herbario	Comprobante de muestra	8983
Instituto de Bioquímica y Fisiología de Plantas y Microorganismos de la Academia Rusa de Ciencias	Cepario	Cepario	7877
Instituto Botánico de la Academia de Ciencias de Tayikistán, Departamento de Flora y Sistemática de Plantas Superiores	Herbario	Comprobante de muestra	4343
Colección de Plantas del Museo de Historia Natural de Londres	Herbario	Comprobante de muestra	107,858
Colección de Plantas Vasculares del Museo de Historia Natural de París	Herbario	Comprobante de muestra	107,341

involucra a investigadores de todo el mundo, además de consorcios científicos particulares y gubernamentales, incluyendo a miles de entidades académicas que pertenecen principalmente a universidades públicas; a manera de ejemplo, el NCBI colabora con el repositorio ENA (European Nucleotide Archive del EMBL) (Amid *et al*., 2020), el cual es una plataforma abierta y compatible para la gestión, intercambio e integración de archivos, así como difusión de datos de secuencias de nucleótidos.

Algunos ejemplos de otros repositorios externos colaboradores que también recopilan información sobre biomoléculas de origen vegetal son los siguientes: UniProt (recurso de libre acceso de secuencias de proteínas e información funcional), RNAcentral (colección de secuencias de ncRNA), EBIMgnify (base de datos metagenómicos e interacciones con microbiomas) y ArrayExpress (almacenamiento de datos de experimentos de genómica funcional de alto rendimiento).

Todos los recursos anteriores se basan en los servicios y contenido de ENA; igualmente, el NCBI colabora con el Banco de Datos de ADN de Japón y la Colaboración Internacional de Bases de Datos de Secuencias de Nucleótidos, los cuales cubren un gran espectro de lecturas sin procesar a través de alineaciones y ensamblajes, al igual que anotaciones funcionales. Estos repositorios son enriquecidos con información contextual vinculada a muestras biológicas y diseños experimentales mediante la International Nucleotide Sequence Database Collection (INSDC) (Arita *et al*., 2021).

De manera complementaria, existen otros recursos informáticos importantes que gestionan colecciones de literatura dentro de las bases de datos del NCBI; por ejemplo, Bookshelf (acceso gratuito en línea a libros y documentos sobre ciencias de la vida y atención médica), MeSH (Medical Subject Headings, tesauro de vocabulario controlado por NLM utilizado para indexar artículos para PubMed) y PCM (PubMed Central, archivo de texto gratuito de literatura de revistas en materia de biomedicina y ciencias de la vida perteneciente a la Biblioteca Nacional de Medicina y el Instituto Nacional de la Salud) (Sayers *et al*., 2020).

Base de datos taxonómica

La base de datos de taxonomía del NCBI, denominada Taxonomy, comprende un conjunto de clasificaciones y nomenclaturas curadas (organización e integración) de todos los organismos que forman parte de las bases de datos de secuencias públicas. Ésto representa actualmente alrededor del 10 % de las especies descritas en el planeta. El repositorio se encuentra estandarizado mediante la INSDC, que comprende GenBank, ENA (EMBL) y bases de datos del DDBJ. Cabe resaltar que Taxonomy es un centro de organización central para muchos de los recursos del NCBI que proporciona un medio para agrupar elementos dentro de otros dominios del sitio web (Federhen, 2012). La base de datos taxonómica sirve como un importante punto de entrada al sistema Entrez para aquellos usuarios que desean conocer la información disponible sobre un taxón en particular; desde el nivel de especie hasta género, familia, orden o superior, según los niveles de jerarquía referidos.

Muchos de los dominios de Entrez (secuencia, estructura, genes, genomas, literatura, etc.) están indexados por taxonomía en el campo de búsqueda [organismo], y estos índices apoyan vínculos recíprocos entre la taxonomía y otros dominios. Todo lo anterior se detalla en los manuales de ayuda NBK25501 y NBK54428 correspondientes (NCBI, 2010; NCBI Taxonomy Help, 2011). Las consultas de taxonomía Entrez se guardan en MyNCBI y el usuario puede registrarse para recibir actualizaciones periódicas por correo electrónico, toda vez que algo nuevo en Entrez satisfaga la consulta; por ejemplo, se puede registrar la consulta [propiedad específica] y solicitar la recepción de un correo electrónico de manera periódica, con la lista de especies que han aparecido en las bases de datos de secuencia por primera vez en la última semana.

Registro y gestión de metadatos complementarios y perfiles de expresión genética (transcriptómica)

Como se mencionó con anterioridad, el NCBI es líder en el campo de la bioinformática y se enfoca principalmente en estudios computacionales para dilucidar fenómenos biológicos (Jenuth, 2000; Sayers *et al*., 2020). Existen conjuntos de metadatos que invitan a los usuarios a complementar los registros de sus investigaciones; por ejemplo, las biocolecciones, bioproyectos y biomuestras, ésto a través de la reagrupación de datos de rutas metabólicas empleando el recurso PubChem Pathways. Por otra parte, el repositorio GEO DataSets reagrupa datos de expresión génica, así como registros originales de series y plataformas en el repositorio Gene Expression Omnibus (GEO).

Los registros de conjuntos de datos contienen recursos adicionales, incluidas herramientas para la generación de grupos (clusters) y consultas de expresiones diferenciales, *e.g.* GEO Profiles (repositorio que almacena perfiles de expresión génica individuales de datos seleccionados en GEO). Geo Profiles permite la búsqueda de perfiles específicos de interés en función de la anotación de genes o características de perfil calculadas previamente.

En la actualidad, existe un sinnúmero de antecedentes en relación con estudios genómicos en plantas, cuyas investigaciones involucraron en gran parte estudios de expresión génica que se valieron de los algoritmos anteriores; por ejemplo, Krzyszton y Kufel (2022) desarrollaron un perfilado de ARNnc (no codificante) mediante secuenciaciones de alto rendimiento en *A. thaliana*; el resultado principal demostró que la alteración de una enzima glucolítica enolasa llamada LOS2 (baja expresión del gen 2 de responsabilidad osmótica) provoca respuestas de defensa constitutiva o autoinmunidad en la especie sujeta a estudio. Los datos anteriores se encuentran registrados en el repositorio de bioproyectos (PRJNA715469) correspondiente.

Otro ejemplo es el trabajo realizado por Garighan *et al*. (2021), quienes identificaron ARN pequeños expresados diferencialmente en embriones de manzana; a través de ello, revelaron el papel potencial de un miRNA denominado Mir159-MYB durante el periodo de latencia. El proyecto se encuentra igualmente registrado (PRJNA784097) en las colecciones del NCBI.

Recientemente, Bouissil *et al*. (2022) hicieron uso de este tipo de repositorios para organizar perfiles de expresión genética derivados de factores de inducción de proteínas de defensa y resistencia de la palmera datilera a *Fusarium oxysporum* f. sp. *albedinis*, en respuesta a alginato extraído de *Bifurcaria bifurcata*. En este proyecto se incluyeron también las anotaciones funcionales de los péptidos respectivos.

Bases de datos genómicos y gestores gráficos

Hoy en día se desarrollan mejoras constantes para la búsqueda de secuencias genómicas mediante la introducción de conjuntos de datos (Datasets), un recurso que permite a los usuarios recopilar fácilmente contenido de todas las bases de datos del NCBI. Los Datasets organizan la información a través de interfaces web distintas, así como de líneas de comandos para descargarlos posteriormente en paquetes de archivos estructurados. Así, los conjuntos de datos admiten consultas de genomas y genes en una amplia gama de grupos taxonómicos. En este sentido, los usuarios que requieren utilizar datos genómicos pueden ensamblar secuencias de genomas, transcriptomas y proteínas, así como anotaciones funcionales contenidas en paquetes computacionales. Estos paquetes incluyen también informes de metadatos enriquecidos, ya que la interfaz web de Datasets permite estudiar genomas a través de reconstrucciones filogenéticas y, elegir cualquier conjunto de secuencias eucariotas y procariotas depositadas en la base de datos de NCBI Assembly. La interfaz de programación de estas aplicaciones brinda acceso también a genomas virales y respalda las búsquedas con identificadores taxonómicos o accesiones de ensamblaje (NCBI Resource Coordinators, 2015; 2016; 2018).

Para el caso de las embriofitas, cuyos especímenes se clasifican como plantas verdes antecesoras de todos los especímenes terrestres (phylum Streptophita), se han agrupado, hasta septiembre de 2022, un poco más de 7,300,000 genes relacionados con cientos de procesos metabólicos distintos, desde mecanismos fitopatológicos de defensa hasta identificación de haplotipos y polimorfismos de nucleótidos únicos (SNPs) (Jones y Dangl, 2006; Ricaño-Rodríguez *et al*., 2018; 2019).

En el último año se ha mejorado la capacidad de descarga de datos dentro del visor gráfico de las bases de datos de nucleótidos y proteínas del NCBI, a través de NCBI Sequence Viewer. Como se mencionó con anterioridad, es posible descargar secuencias de genes y datos de anotación funcional, inclusive de SNPs (https://www.ncbi.nlm.nih.gov/snp/). Los usuarios también pueden copiar cadenas cortas de datos de secuencias directamente al portapapeles y descargar segmentos más grandes en formato de archivo FASTA (formato de texto que representa secuencias de nucleótidos o aminoácidos mediante códigos de una sola letra).

Con respecto al estudio estructural genómico, incluida la traducción a proteínas, estas herramientas también proporcionan datos referentes a clasificaciones de dominios conservados (CDD), así como número de exones e intrones en secuencias nucleotídicas. En el caso de alineaciones tipo SRA o BLAST, los algoritmos identifican cualquier dato no alineado, incluidas inserciones y colas 5' y 3'. Las herramientas en mención son útiles en estudios de genotipeo por secuenciación; por ejemplo, a través del estudio de regiones sumamente conservadas de especímenes vegetales (*e.g. Theobroma cacao* L. y *Aegilops umbellulata*) se determinan variaciones genotípicas a nivel polimórfico, así como el mapeo de genes de resistencia a fitopatógenos, como la roya (Edae *et al*., 2016; Lima *et al*., 2009; Ricaño-Rodríguez *et al*., 2019); ésto permite a los investigadores dilucidar más a fondo las características genéticas distintivas de grupos de individuos que pertenecen a un mismo phylum.

Visor de datos genómicos

El navegador de datos genómicos insignia del NCBI, denominado Genome Data Viewer (GDV), es una interfaz que integra un motor gráfico con un sólido algoritmo

de búsqueda/recuperación/análisis de datos. Para respaldar mejor los análisis y solventar las necesidades de investigación de cada usuario en particular, se han lanzado numerosas mejoras de los aspectos técnicos de dicha herramienta (NCBI Resource Coordinators, 2018).

El genoma de *Zea mays* (B73) es uno de los conjuntos de datos más completos registrados en GDV. A través del visor gráfico se observan reconstrucciones filogenéticas diversas de esta especie. Mediante los hipervínculos disponibles relacionados con el proyecto anterior (RefSeq accession GCF_902167141.1) se puede estudiar la mayoría de sus genes y transcritos correspondientes (incluyendo anotaciones), así como consultar a detalle su exoma y *loci* respectivos. Estos estudios se complementan con la herramienta RefSeq (conjunto completo, integrado, no redundante y bien anotado de secuencias de referencia que incluyen genómica, transcriptómica y proteómica). Las mejoras adicionales del visor gráfico también permiten la navegación agrupando por ensamblajes y regiones cromosómicas.

Gestión y procesamiento de secuencias genómicas

En el NCBI existe un recurso informático denominado "Gene" que incluye datos sobre perfiles de expresión genética de diversas especies, éstos van desde representaciones gráficas de la expresión de cada gen integrada en su propia página de informe completo, hasta la descarga de conjuntos de secuencias nucleotídicas para su procesamiento. Gene integra información de una amplia gama de especies y un solo registro incluye nomenclatura, secuencias de referencia, mapas, vías metabólicas, variaciones y fenotipos, así como enlaces a recursos específicos del genoma y *loci* de especímenes de todos los reinos (Brown *et al*., 2015).

Los perfiles de expresión son complementos útiles de funciones genéticas ya caracterizadas, y también medios potenciales para estudiar la función de genes descubiertos de manera reciente (Ozsolak y Milos, 2011). Constantemente se realizan actualizaciones para generar resúmenes de texto que acompañen al perfil de expresión de cada gen en cuestión, y que estos datos sean indexados dentro del sistema de consulta de Entrez. Los perfiles de expresión se calculan a partir de alineaciones de secuencias de ARN generadas por la canalización de cada genoma anotado dentro del NCBI. Dicho proceso selecciona conjuntos de datos disponibles públicamente en formatos SRA (Fagerberg *et al*., 2014).

Después de alinear las lecturas de una muestra con la secuencia genómica, para cada gen se calcula la cobertura de lectura (en comparación con todos los exones anotados), y se normaliza a todas las lecturas alineadas con el genoma de referencia, empleándolas para derivar lecturas por kb por cada millón de éstas situadas a través de la secuencia diana. Los datos de las réplicas biológicas dentro del mismo proyecto de SRA se promedian y se presentan con su respectiva desviación estándar. Es importante mencionar que los niveles de expresión de diferentes proyectos de SRA se informan de manera independiente.

Colecciones de secuencias nucleotídicas y reconstrucciones de grupos homólogos

Existen dos herramientas básicas en el NCBI denominadas HomoloGene y Nucleotide que emplean un sistema automatizado para construir grupos de homología putativa, a partir de conjuntos de genes completos de una amplia gama de especies eucariotas; por ejemplo, a través de los recursos anteriores Hosmani *et al.* (2019) estudiaron las características del genoma nuclear del tomate (*S. lycopersicum*) mediante la secuenciación de fragmentos digeridos con *Eco*RI submetilados. Por otra parte, el "proyecto genoma de la naranja dulce de China" (The draft genome of sweet orange) (*Citrus sinensis*) (Xu *et al*., 2013) consistió en secuenciar los cromosomas de esta especie (bioproyecto PRJNA86123) y estudiarlos posteriormente a través de HomoloGene y Nucleotide.

En la actualidad se pretende que los usuarios tengan acceso a una colección de secuencias genómicas provenientes de varias fuentes, incluidas GenBank, RefSeq, TPA, Third Party Anotation (base de datos diseñada para capturar resultados experimentales o inferenciales por parte de terceros que derivaron de los datos primarios de GenBank) y PDB, Protein Data Bank (Berman *et al*., 2000).

Herramienta básica de búsqueda de alineación local (BLAST)

La herramienta básica de búsqueda de alineación local (BLAST, Basic Local Alignment Search Tool) es un algoritmo informático de comparación y alineación de secuencias nucleotídicas (*i.e.* ADN, ARN y aminoácidos) que se utiliza a través de una interfaz web, o bien, como una herramienta independiente para comparar y analizar la consulta de otro usuario con una base de datos de secuencias previas (Altschul *et al*., 1990; 1997).

BLAST funciona bajo un enfoque heurístico que identifica coincidencias entre dos secuencias, e intenta iniciar alineaciones desde estos puntos de partida. Además de realizar alineaciones, BLAST proporciona

información estadística sobre cada trabajo llevado a cabo; por ejemplo, sugiere un valor "esperado" o tasa de falsos positivos, y de manera general, este recurso es empleado para la búsqueda de probables genes homólogos. Dicha herramienta usa el algoritmo "Smith-Waterman" (programación dinámica optimizada con matrices de substitución) (Smith y Waterman, 1981) para realizar las alineaciones correspondientes. BLAST es una herramienta utilizada en múltiples campos de las ciencias naturales, incluyendo el estudio de la fitogenética.

En la página de inicio de BLAST se enumeran las diferentes búsquedas por naturaleza realizadas, *i.e.* nucleótidos, transcritos y genomas. Cuando se envía una consulta al servidor del NCBI, ya sea como una secuencia en formato FASTA o como un identificador de secuencia, *e.g.* número de acceso de GenBank, la búsqueda se envía al servidor BLAST y se devuelve un identificador de solicitud denominado RID. La consulta y los resultados correspondientes se almacenan en un formato estructurado hasta 24 h después de que se emite un RID. Este identificador reagrupa la búsqueda y permite ver los resultados en varios formatos, que incluyen el informe BLAST, una tabla de resultados simplificada, archivos XML (Extensible Markup Languaje, Lenguaje de Mercado Extensible) y ASN.1 (Abstract Syntax Notation one, Notación Sintáctica Abstracta uno) (Madden, 2002).

Las nuevas bases de datos BLAST están disponibles en un sitio FTP (File Transfer Protocol, Protocolo de Transferencia de Archivos) del NCBI, así como en los proveedores de nube de GCP (Google Cloud Plataform, Plataforma de Nube de Google) y AWS (Amazon Web Services, Servicios de Nube de Amazon). En este sentido, los tres sitios ofrecen las mismas 23 bases de datos y éstas poseen desde una colección de betacoronavirus, hasta genomas representativos de Ref-Seq pertenecientes a organismos eucariotas, procariotas y otras familias de virus.

De manera paralela se han desarrollado herramientas complementarias como Primer-BLAST (Ye *et al*., 2012), la cual diseña cebadores comunes para grupos de secuencias de ADN muy similares; ésto permite que los investigadores realicen tareas específicas, como amplificar múltiples variantes de transcripción para un solo gen, o detectar un grupo de cepas bacterianas filogenéticamente relacionadas (McGinnis y Madden, 2004); asimismo, el apartado de análisis de secuencias del NCBI cuenta con un conjunto de recursos sumamente importantes que complementan las tareas anteriores, *i.e.* BLAST (Stand alone), FTP. BLAST Databases, BLAST: microbial Genomes, BLAST RefSeqGene, COBALT, Genome Remaping Services, Genome Workbench, Multiple Sequence Alignment Viewer, Open Reading Frame Finder (ORF Finder), ProSplign y Tree viewer (NCBI Resource Coordinators, 2015; 2016; 2018).

Bases de datos de proteínas y herramientas para estudios proteómicos

En 2014 el NCBI introdujo el "Informe de Proteínas Idénticas" en su base de datos para corroborar las relaciones entre las secuencias de aminoácidos existentes (número de acceso de proteínas no redundantes), así como el conjunto de secuencias de dominios conservados de nucleótidos individuales (NCBI Resource Coordinators, 2015). En la actualidad, estos informes se han mejorado y recopilado en un nuevo recurso denominado Identical Protein Groups; Grupos de Proteínas Idénticas. Este recurso, abreviado IPG, incluye todas las secuencias de proteínas registradas en el NCBI y PDB, con enlaces a secuencias de transcripción de nucleótidos de GenBank y RefSeq.

Por otra parte, hace dos décadas se creó un consorcio denominado UniProt, el cual surge de la colaboración entre el Instituto Suizo de Bioinformática (SIB), el Instituto Europeo de Bioinformática (EBI) y el Recurso de Información sobre Proteínas (PIR). Swiss-Prot, junto con TrEMBL (traducción automática de EMBL), se unieron con PIR para generar la "UniProt Knowledgebase" (UniProtKB), el catálogo de proteínas más importante del mundo. En materia de proteómica vegetal existen antecedentes de proyectos que emplearon las bases de datos anteriores, incluyendo desde estudios de fitopatología hasta aspectos nutrimentales concernientes a los seres humanos (Shewry y Lucas, 1997; Yagami, 2002).

Recientemente, el NCBI lanzó una versión actualizada (2.19.0) de un software denominado iCn3D (Wang *et al*., 2020), una interfaz gráfica de estructuración molecular tridimensional (3D) que se ejecuta directamente en los navegadores web. Las vistas interactivas de iCn3D están integradas en las páginas de resumen de la estructura de la base de datos de modelado molecular (Molecular Modeling DataBase) del NCBI, e iCn3D visualiza los resultados de las comparaciones de estructuras 3D calculadas por un algoritmo llamado VAST+ (Vector Alignment Search Tool), que consiste en un software que se enfoca principalmente en la búsqueda de estructuras macromoleculares que tengan una unidad biológica similar, en lugar de aquellas que son similares a nivel de una molécula proteica individual o con dominios tridimensionales, así como las alineaciones de secuencias con estructuras por pares calculadas mediante BLAST. iCn3D muestra simultáneamente estructuras 3D, esquemas de interacción 2D, alineaciones y secuencias de proteínas/nucleótidos, así como anotaciones de sitios funcionales y huellas de dominios conservados.

Por otra parte, la base de datos Structure alberga más de 7,890 estructuras tridimensionales de macromoléculas de plantas, su conformación computacional proporciona una gran cantidad de información sobre la función biológica e historia evolutiva de las especies, y se pueden usar para examinar relaciones secuencia-estructura-función, así como interacciones y sitios activos moleculares. Entre las estructuras más interesantes en materia de proteómica vegetal se pueden mencionar ejemplos como proteínas relacionadas con fotosistemas (Amunts *et al*., 2010), supercomplejos de cloroplastos PSI-NDH (Shen *et al*., 2022) y proteínas ribosomales L30e (Halic *et al*., 2005). Asimismo, es posible observar estructuras cristalográficas de ribosomas mitocondriales, hidrolasas, glucanasas, fosfatasas, fitocromos y aldehído deshidrogenasas, entre otras (NCBI Resource Coordinators, 2015; 2016; 2018).

En la Figura 3 se muestra un ejemplo de una estructura molecular compuesta por dos elementos de una proteína Argonauta tipo 1 de *A. thaliana* (Argonaute protein, Gen *AGO1*), la cual es visualizada con la mayoría de sus elementos constitutivos a través de la interfaz del algoritmo Structure. Las proteínas argonautas desempeñan un papel importante en la regulación génica en el núcleo celular; debido a ello, estas biomoléculas son capaces de interferir en procesos epigenéticos durante la transcripción (Arribas-Hernández *et al*., 2016).

Otras características recientemente agregadas a esta base de datos incluyen la visualización extendida de redes de interacción 2D entre proteínas y ligandos, u otras proteínas; también se lleva a cabo la visualización de potenciales electrostáticos calculados por el método Delphi (Li *et al*., 2012) y la visualización de la ubicación de bicapas de membrana en relación con estructuras de proteínas transmembranales (Lomize *et al*., 2012). iCn3D está disponible en https://github.com/ncbi/icn3d y sus características más novedosas actualizadas se pueden consultar en la página de la galería correspondiente. La herramienta anterior suele utilizarse de manera sinérgica con la base de datos de dominios conservados, la cual es un recurso imprescindible para la anotación de unidades funcionales en proteínas, su colección de modelos de dominios incluye un conjunto seleccionado por el NCBI, que utiliza estructuras 3D para proporcionar información sobre las relaciones de secuencia/estructura/función (Geer *et al*., 2002).

CONCLUSIONES

Uno de los objetivos principales del NCBI es el desarrollo constante de nuevas tecnologías de la información, lo cual conlleva a un mejor entendimiento del origen y consecuentes procesos genético-moleculares de las especies, incluyendo las plantas; asimismo, permite la descripción de nuevas hipótesis como un reto a resolver para las generaciones presentes y futuras. Dentro del grupo de las plantas (Reino: Plantae), diversas bases de datos y herramientas del repositorio en mención que en su mayoría han sido previamente descritas permiten una comprensión más profunda de la naturaleza de miles de especímenes de interés agro-biotecnológico, medicinal, industrial y hasta biocultural para la sociedad, así como para la comunidad científica, lo que incrementa su potencial de aprovechamiento. Teniendo en cuenta tales consideraciones, es importante destacar la necesidad de una mayor incursión en el tema en el cual se utilicen las herramientas bioinformáticas disponibles para el desarrollo de las ciencias ómicas de manera prioritaria, ya que este universo de conocimiento se actualiza de manera constante debido a la diaria generación de datos y recursos bioinformáticos, principalmente aplicados al campo de las ciencias naturales.

AGRADECIMIENTOS

A los revisores de este manuscrito, al igual que a todos aquellos colegas que realizaron una lectura previa. Sin duda cada uno de sus comentarios y sugerencias permitieron la obtención de un mejor documento.

BIBLIOGRAFÍA

Altschul S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389-3402, https://doi.org/10.1093/nar/25.17.3389

Altschul S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215:403-410, https://doi.org/10.1016/S0022-2836(05)80360-2

Amid C., B. T. F. Alako, V. B. Kadhirvelu, T. Burdett, J. Burgin, J. Fan, … and G. Cochrane (2020) The European Nucleotide Archive in 2019. *Nucleic Acids Research* 48:D70-D76, https://doi.org/10.1093/nar/gkz1063

Amunts A., H. Toporik, A. Borovikova and N. Nelson (2010) Structure, determination and improved model of plant photosystem I. *Journal of Biological Chemistry* 285:3478-3486, https://doi.org/10.1074/jbc.M109.072645

Arita M., I. Karsch-Mizrachi and G. Cochrane (2021) The international nucleotide sequence database collaboration. *Nucleic Acids Research* 49:D121-D124, https://doi.org/10.1093/nar/gkaa967

Arribas-Hernández L., L. J. Kielpinski and P. Brodersen (2016) mRNA decay of most Arabidopsis miRNA targets requires slicer activity of AGO1. *Plant Physiology* 171:2620-2632, https://doi.org/10.1104/pp.16.00231

Benson D. A., I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell and E. W. Sayers (2012) GenBank. *Nucleic Acids Research* 40:D48-D53, https://doi.org/10.1093/nar/gkr1202

Berman H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, … and P. E. Bourne (2000) The protein data bank. *Nucleic Acids Research* 28:235-242, https://doi.org/10.1093/nar/28.1.235

Bouissil S., C. Guérin, J. Roche, P. Dubessay, Z. El Alaoui-Talibi, G. Pierre, … and C. El Modafar (2022) Induction of defense gene expression and the resistance of date palm to *Fusarium oxysporum* f.

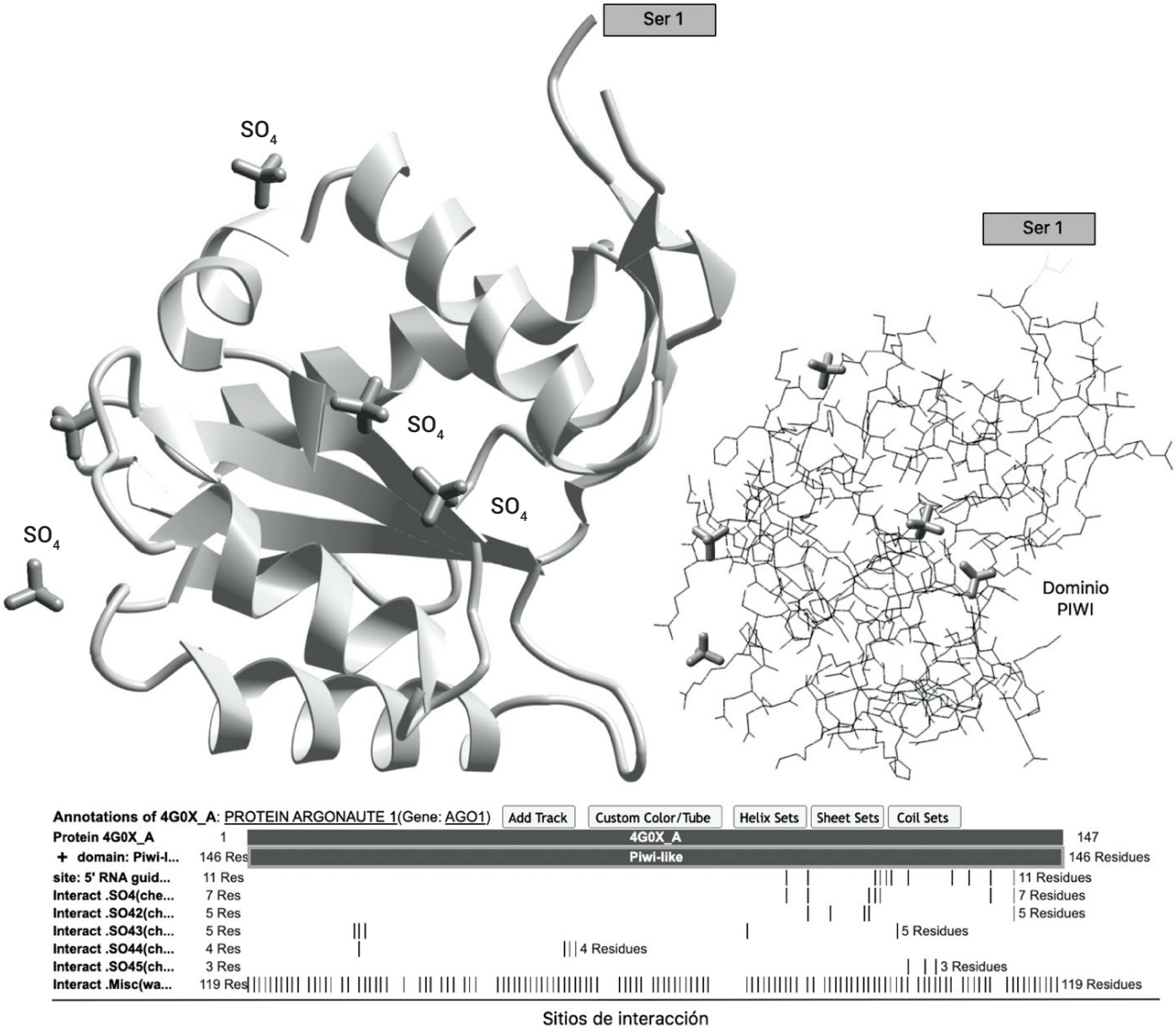

Figura 3. Estructura cristalográfica de la proteína argonauta Ago1 dominio medio (MID) de *Arabidopsis thaliana* observada a través de la interfaz del algoritmo Structure del NCBI. La estructura izquierda se encuentra representada por 147 residuos de aminoácidos, iniciando con la serina. Se muestran siete sitios de interacción (SO_4, agua y sitios de anclaje a hebras de ARN 5´ guías). La estructura derecha representa el dominio tipo PIWI de la proteína (dominio encontrado en proteínas involucradas en el silenciamiento de ARN). Este fenómeno se refiere a un grupo de mecanismos de silenciamiento de genes relacionados mediados por moléculas cortas de ARN, que incluyen siARN, miARN y ARN guía relacionados con heterocromatina (Madej *et al.,* 2014; NCBI Resource Coordinators, 2015; 2016; 2018).

sp. *albedinis* in response to alginate extracted from *Bifurcaria bifurcata*. *Marine Drugs* 20:88, https://doi.org/10.3390/md20020088

Brown G. R., V. Hem, K. S. Katz, M. Ovetzky, C. Wallin, O. Ermolaeva, … and T. D. Murphy (2015) Gene: a gene-centered information source at NCBI. *Nucleic Acids Research* 43:D36-D42, https://doi.org/10.1093/nar/gku1055

Edae E. A., P. O. Olivera, Y. Jin, J. A. Poland and M. N. Rouse (2016) Genotype-by-sequencing facilitates genetic mapping of a stem rust resistance locus in *Aegilops umbellulata*, a wild relative of cultivated wheat. *BMC Genomics* 17:1039, https://doi.org/10.1186/s12864-016-3370-2

Fagerberg L., B. M. Hallstrom, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, … and M. Uhlén (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics* 13:397-406, https://doi.org/10.1074/mcp.M113.035600

Federhen S. (2012) The NCBI taxonomy database. *Nucleic Acids Research* 40:D136-D143, https://doi.org/10.1093/nar/gkr1178

Garighan J., E. Dvorak, J. Estevan, K. Loridon, B. Huettel, G. Sarah, … and

F. Andrés (2021) The identification of small RNAs differentially expressed in apple buds reveals a potential role of the miR159-MYB regulatory module during dormancy. *Plants* 10:2665, https://doi.org/10.3390/plants10122665

Geer L. Y., M. Domrachev, D. J. Lipman and S. H. Bryant (2002) CDART: protein homology by domain architecture. *Genome Research* 12:1619-1623, https://doi.org/10.1101/gr.278202

Geer L. Y., A. Marchler-Bauer, R. C. Geer, L. Han, J. He, S. He, … and S. H. Bryant (2010) The NCBI BioSystems database. *Nucleic Acids Research* 38:D492-D496, https://doi.org/10.1093/nar/gkp858

Halic M., T. Becker, J. Frank, C. M. Spahn and R. Beckmann (2005) Localization and dynamic behavior of ribosomal protein L30e. *Nature Structural & Molecular Biology* 12:467-468, http://doi.org/10.1038/nsmb933

Hosmani P. S., M. Flores-González, H. van de Geest, F. Maumus, L. V. Bakker, E. Schijlen, … and S. Saha (2019) An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *bioRxiv* 1:767764, https://doi.org/10.1101/767764

Jenuth J. P. (2000) The NCBI: *In*: Bioinformatics Methods and Protocols. S. Misener and S. A. Krawetz (eds.). Humana Press. Totowa, New Jersey, USA. pp:301-312, https://doi.org/10.1385/1-59259-192-2:301

Jones J. D. G. and J. L. Dangl (2006) The plant immune system. *Nature* 444:323-329, https://doi.org/10.1038/nature05286

Jorrin-Novo J. V. (2020) What is new in (plant) proteomics methods and protocols: the 2015-2019 quinquennium. *In*: Plant Proteomics. Methods in Molecular Biology. J. Jorrin-Novo, L. Valledor, M. Castillejo and M. D. Rey (eds.). Humana Press. New York, USA. pp:1-10, https://doi.org/10.1007/978-1-0716-0528-8_1

Krzyszton M. and J. Kufel (2022) Analysis of mRNA-derived siRNAs in mutants of mRNA maturation and surveillance pathways in *Arabidopsis thaliana*. *Scientific Reports* 12:1474, https://doi.org/10.1038/s41598-022-05574-4

Li L., C. Li, S. Sarkar, J. Zhang, S. Witham, Z. Zhang, … and E. Alexov (2012) DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC Biophysics* 5:9, https://doi.org/10.1186/2046-1682-5-9

Lima L., K. Gramacho, N. Carels, R. Novais, F. Gaiotto, U. Lopes, … and F. Micheli (2009) Single nucleotide polymorphisms from *Theobroma cacao* expressed sequence tags associated with witches' broom disease in cacao. *Genetics and Molecular Research* 8:799-808, https://doi.org/10.4238/vol8-3gmr603

Lomize M. A., I. D. Pogozheva, H. Joo, H. I. Mosberg and A. L. Lomize (2012) OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Research* 40:D370-D376, https://doi.org/10.1093/nar/gkr703

Madden T. L. (2002) The BLAST sequence analysis tool: *In*: The NCBI Handbook. J. McEntyre (ed.). National Center for Biotechnology Information. Bethesda, Maryland, USA. pp:11-17.

Madej T., C. J. Lanczycki, D. Zhang, P. A. Thiessen, R. C. Geer, A. Marchler-Bauer and S. H. Bryant (2014) MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Research* 42:D297-303, https://doi.org/10.1093/nar/gkt1208

Maglott D., J. Ostell, K. D. Pruitt and T. Tatusova (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* 33:D54-D58, https://doi.org/10.1093/nar/gki031

McGinnis S. and T. L. Madden (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research* 32:W20-W25, https://doi.org/10.1093/nar/gkh435

Michael T. P. and S. Jackson (2013) The first 50 plant genomes. *The Plant Genome* 6:1-7, https://doi.org/10.3835/plantgenome2013.03.0001in

Morgante M. and F. Salamini (2003) From plant genomics to breeding practice. *Current Opinion in Biotechnology* 14:214-219, https://doi.org/10.1016/S0958-1669(03)00028-4

NCBI, National Center for Biotechnology Information (2010) Entrez Programming Utilities Help. National Center for Biotechnology Information. Bethesda, Maryland, USA. http://www.ncbi.nlm.nih.gov/books/NBK25501/ (February 2022).

NCBI, National Center for Biotechnology Information (2011) Taxonomy help. Frequently asked questions. National Center for Biotechnology Information, Bethesda, Maryland, USA, https://www.ncbi.nlm.nih.gov/books/NBK54428/ (Febrary 2022).

NCBI Resource Coordinators (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 42:D7-D17, https://pubmed.ncbi.nlm.nih.gov/25398906/

NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 44:D7-D19, https://doi.org/10.1093/nar/gkv1290

NCBI Resource Coordinators (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 46:D8-D13, https://doi.org/10.1093/nar/gkx1095

Ogasawara O., Y. Kodama, J. Mashima, T. Kosuge and T. Fujisawa (2020) DDBJ database updates and computational infrastructure enhancement. *Nucleic Acids Research* 48:D45-D50, https://doi.org/10.1093/nar/gkz982

Ozsolak F. and P. M. Milos (2011) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* 12:87-98, https://doi.org/10.1038/nrg2934

Rensink W. A. and C. R. Buell (2005) Microarray expression profiling resources for plant genomics. *Trends in Plant Science* 10:603-609, https://doi.org/10.1016/j.tplants.2005.10.003

Ricaño-Rodríguez J., E. Hipólito-Romero, J. M. Ramos-Prado y E. Cocoletzi-Vásquez (2019) Genotipado por secuenciación de variedades nativas de *Theobroma cacao* (Malvaceae) de los Estados de Tabasco y Chiapas, México. *Botanical Sciences* 97:381-397, https://doi.org/10.17129/botsci.2258

Ricaño-Rodríguez J., J. M. Ramos-Prado, E. Cocoletzi-Vásquez y E. Hipólito-Romero (2018) El estudio genómico del cacao; breve recopilación de sus bases conceptuales. *Agroproductividad* 11:29-35, https://doi.org/10.32854/agrop.v11i9.1211

Ricroch A. E., J. Martin-Laffon, B. Rault, V. C. Pallares and M. Kuntz (2022) Next biotechnological plants for addressing global challenges: the contribution of transgenesis and new breeding techniques. *New Biotechnology* 66:25-35, https://doi.org/10.1016/j.nbt.2021.09.001

Sayers E. W., J. Beck, J. R. Brister, E. E. Bolton, K. Canese, D. C. Comeau, … and J. Ostell (2020) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 48:D9-D16, https://doi.org/10.1093/nar/gkz899

Sayers E. W., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, … and J. Ye (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 37:D5-D15, https://doi.org/10.1093/nar/gkn741

Shangguan L., J. Han, E. Kayesh, X. Sun, C. Zhang, T. Pervaiz, … and J. Fang (2013) Evaluation of genome sequencing quality in selected plant species using expressed sequence tags. *PLoS ONE* 8:e69890, https://doi.org/10.1371/journal.pone.0069890

Shen L., K. Tang, W. Wang, C. Wang, H. Wu, Z. Mao, … and X. Zhang (2022) Architecture of the chloroplast PSI-NDH supercomplex in *Hordeum vulgare*. *Nature* 601:649-654, https://doi.org/10.1038/s41586-021-04277-6

Shewry P. R. and J. A. Lucas (1997) Plant proteins that confer resistance to pests and pathogens. *Advanvces in Botanical Research* 26:135-170, https://doi.org/10.1016/S0065-2296(08)60120-2

Smith T. F. and M. S. Waterman (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* 147:195-197, https://doi.org/10.1016/0022-2836(81)90087-5

Trumbly R. J. (2022) Accessing genomic databases. *In*: Molecular Analyses. S. O. Rogers (ed.). CRC Press. Boca Raton, USA. pp:126-138.

Van Bel M., F. Silvestri, E. M. Weitz, L. Kreft, A. Botzki, F. Coopens and K. Vandepoele (2022) PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Research* 50:D1468-D1474, https://doi.org/10.1093/nar/gkab1024

Wang J., P. Youkharibache, D. Zhang, C. J. Lanczycki, R. C. Geer, T. Madej, … and A. Marchler-Bauer (2020) iCn3D, a web-based 3D viewer for sharing 1D/2D/3D representations of biomolecular structures. *BMC Bioinformatics* 36:131-135, https://doi.org/10.1093/bioinformatics/btz502

Wieczorek J., D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, … and D.

Vieglais (2012) Darwin core: an evolving community-developed biodiversity data standard. *PLoS ONE* 7:e29715, https://doi.org/10.1371/journal.pone.0029715

Xu Q., L. L. Chen, X. Ruan, D. Chen, A. Zhu, C. Chen, … and Y. Ruan (2013) The draft genome of sweet orange (*Citrus sinensis*). *Nature Genetics* 45:59-66, https://doi.org/10.1038/ng.2472

Yagami T. (2002) Allergies to cross-reactive plant proteins. *International Archives of Allergy and Immunology* 128:271-279, https://doi.org/10.1159/000063859

Ye J., G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen and T. L. Madden (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13:134, https://doi.org/10.1186/1471-2105-13-134