Generación de una herramienta para la búsqueda de metadata asociada a transcriptos

Rodrigo Machado[1], Sebastián Moschen [2], Sergio Alberto Gónzalez[3], Gabriela Conti [3], Mora Massaro[4], Julio Alejandro Di Rienzo[5], Lourdes Burdyn[1], Horacio Esteban Hopp[3] y Paula Fernández[3].

[1] Estación Experimental Agropecuaria INTA Concordia, Protección Vegetal, Concordia, Entre Ríos, Argentina.
machado.rodrigo@inta.gob.ar
burdyn.lourdes@inta.gob.ar

[2] Estación Experimental Agropecuaria INTA Famaillá, Tucumán, Argentina.
moschen.sebastian@inta.gob.ar

[3] Instituto de Agrobiotecnología y Biología Molecular (IABIMO), UEDD INTA CONICET, Buenos Aires, Argentina.
fernandez.pc@inta.gob.ar
gonzalez.sergio@inta.gob.ar
conti.gabriela@inta.gob.ar
hopp.esteban@inta.gob.ar

[4] Instituto de Biología y Medicina Experimental, Buenos Aires, Argentina.
mora.massaro@gmail.com

[5] Facultad de Ciencias Agropecuarias, UNC, Córdoba, Argentina
dirienzo.julio@gmail.com

Resumen. Al analizar un transcriptoma en primera instancia se debe realizar un pre-procesamiento computacional de las lecturas de transcriptos generados por secuenciadores masivos y posterior obtención de la expresión diferencial de los genes asociados a muestras control y tratadas. Los genes cuantificados deben ser caracterizados con el fin de obtener toda la información posible de la secuencia del genoma, para luego ser anotados funcionalmente. En ese sentido, el desarrollo de un buscador de metadata asociada a cada lectura de manera personalizada para cada experimento y/o especie conlleva grandes ventajas. A la hora de procesar los datos, a partir del nombre del gen o secuencia puede obtenerse, por ejemplo, la descripción de la proteína, los términos GO, vías de Uniprot, así como resúmenes sobre las categorías funcionales. En el presente trabajo, se utilizó como punto de partida una tabla de expresión diferencial obtenida a partir del procesamiento bioinformático de la colección biológica PRJNA417324 y mediante la técnica de raspado web, empleando el programa Beautiful Soup, se obtuvo metadata asociada para cada transcripto a partir de la base de datos UNIPROT.

Palabras claves: **raspado web, transcriptómica, metadata.**

1 introducción

Al realizar un análisis de transcriptómica se suele emplear dos tecnologías: microarreglos de ARN o secuenciación de ARN (RNA-seq, por sus siglas en inglés). La tecnología de microarreglos permite la detección simultánea de múltiples fragmentos de material genético hibridados sobre una micromatriz o chip. La cuantificación se realiza considerando la emisión de fluorescencia y es visualizada mediante un microscopio [1]. Por otro lado, en el RNA-seq, se utiliza la secuenciación masiva (NGS, por su acrónimo en inglés, Next Generation Sequencing) para revelar la presencia y cantidad de transcriptos, en una muestra biológica en un momento dado [2]. Cuando se realiza un RNA-seq, se producen millones de secuencias a partir de muestras complejas de ARN. Los datos crudos obtenidos mediante ambas tecnologías deben pasar por una serie de pasos *in-silico* que implican el procesamiento y normalización de los datos, la cuantificación de los niveles de expresión de cada gen y establecimiento de las diferencias en la expresión entre las diferentes condiciones experimentales [3]. En el último paso mencionado, se obtiene una matriz constituida por genes agrupados según su nivel de expresión, la cual podemos etiquetar como primer procesamiento de datos.

Luego de este primer procesamiento, los datos deben ser anotados funcionalmente para que adquieran una significancia biológica y para ello existen bases de datos biológicas que almacenan esta metadata necesaria. Estas bases de datos se encuentran enlazadas entre sí y esto se debe a la tendencia global de unificación de criterios por lo que los datos se organizan en conjunto de registros estructurados que permiten recuperar fácilmente la información. Cada registro está compuesto por un número de campos determinados que contienen datos específicos, como puede ser el nombre de un gen, nombre de la proteína, función de la proteína, proceso GO. Existen múltiples bases de datos biológicas de secuencias de nucleótidos y de proteínas: primarias o secundarias, curadas o no curadas, específicas para un organismo o generales, que se encuentran interconectadas mediante referencias cruzadas que brindan abundante información al usuario. Algunas de las más reconocidas son: UniprotKB[1], la misma contiene basta información de proteínas, cuyos datos son precisos, de alta calidad; Genbank[2], que es una colección pública de secuencias de nucleótidos anotadas, cuyos datos son redundante, pueden haber varias secuencias para un gen determinado y estas pueden ser de buena o mala mala calidad; y Refseq[3] que contiene información curada y anotada de secuencias de nucleótidos y sus productos proteicos. En este contexto, el desarrollo de una herramienta que recopile información de una base de datos de manera automatizada y se adecúe a cada experimento y/o especie sería una gran ventaja ya que reduciría los tiempos de búsqueda. A la hora de procesar los datos, a partir del nombre del gen o secuencia podría obtenerse, por ejemplo, la descripción de proteína, el código enzimático (EC), la ontología génica (GO), así como resúmenes sobre las categorías funcionales.

En este sentido, el *web scraping* o raspado web es una técnica que puede ser utilizada para extraer información de páginas web de forma automatizada. Para llevarlo a cabo, existen varios paquetes implementados en Python ampliamente utilizados como

[1] https://www.uniprot.org/uniprot/
[2] https://www.ncbi.nlm.nih.gov/genbank/
[3] https://www.ncbi.nlm.nih.gov/refseq/

Scrapy, Beautiful Soup, Curl, entre otros, que permiten el desarrollo de soluciones de scraping adaptadas la estructura de un sitio web en particular [4]. En el presente trabajo se utilizó la herramienta Beautiful Soup [5] para realizar la búsqueda personalizada de metadata asociada a muestras de transcriptoma que ya contaban con un primer procesamiento.

El objetivo de este trabajo es desarrollar un programa que utilice la metodología de escarbado web, de modo de aprovechar las ventajas que trae consigo esta tecnología, entre las que se encuentran la disminución de la carga de trabajo, el aumento en la velocidad de procesamiento, el manejo de grandes cantidades de datos y la obtención de metadata asociada a las búsquedas en un formato que luego sea procesable. Es importante aclarar que la herramienta desarrollada no es un anotador funcional como lo son Blast2GO [6], AutoFact [7], Sma3s [8], cuyos algoritmos están típicamente basados en la búsqueda de homología de secuencia y análisis predictivo. A lo que apunta el presente trabajo es a la búsqueda automatizada de múltiples entradas en una base datos y a la extracción de la información en un formato más amigable para ser previsualizado y posteriormente procesado.

2 Materiales y métodos:

El desarrollo de esta herramienta se compone de cuatro partes principales (Fig. 1).

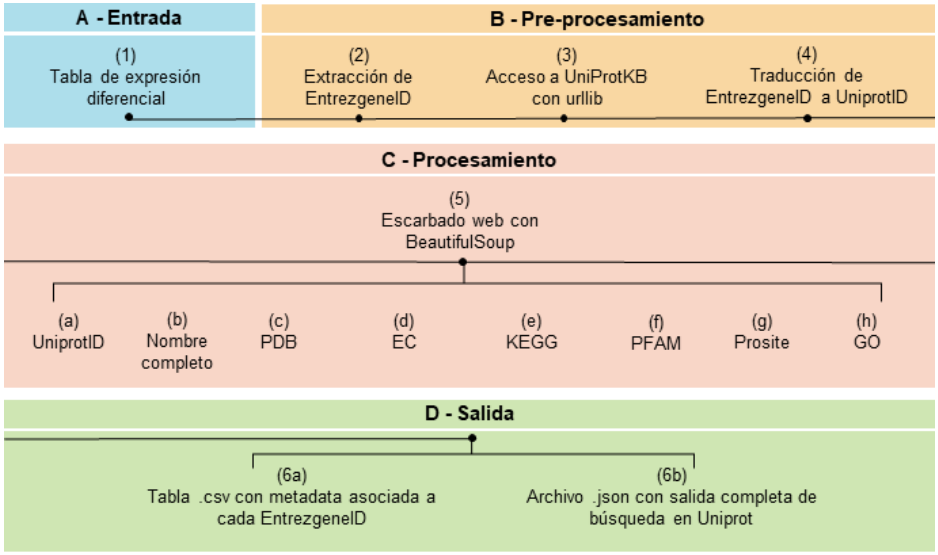

Fig. 1. Resumen gráfico de la aproximación realizada.

2.1 Entrada

Como entrada para el código se empleó una tabla con datos de expresión diferencial obtenida mediante el preanálisis de una base de datos biológica, la misma fue importada con el paquete Pandas, el cual ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales, opciones útiles para tal fin (Fig. 1.A).

La tabla de input final contenía las siguientes columnas: GeneID, LogFC, P-value y Entrez gene ID como se señala en la tabla 1.

Tabla 1. Expresión diferencial de genes.

GeneID	Log2FC	Padj	Entrez_gene_ID

Los parámetros establecidos en el input son: *GeneID*, el cual es un identificador estable de genes y otros locus, para cada especie en particular; *Log2FC* o tasa de cambio, que es una medida que describe cuánto cambia una cantidad entre una medición original y una posterior, en este caso la expresión diferencial de un gen entre una condición control y un tratamiento; *Padj* o probabilidad de un valor estadístico ajustado, que indica cuáles comparaciones entre los niveles de los factores dentro de una familia de comparaciones (pruebas de hipótesis) son significativamente diferentes; y el *Entrez gene* ID, que corresponde a un número de acceso a Entrez Gene[4] que es una base de datos de información específica de genes que se enfoca en genomas completamente secuenciados y/o bajo intenso análisis.

2.2 Pre-procesamiento

A partir del *Entrez gene* ID se parseo la base de datos UniProtKB (Fig. 1.B), que está centrada en las secuencias de las proteínas e incluye información relacionada con el papel de la proteína como su función, estructura, localización subcelular, interacciones con otras proteínas y dominios que la componen, así como un amplio rango de características relativas a la secuencia como sitios activos y modificaciones postraduccionales (Fig. 2).

[4] https://www.ncbi.nlm.nih.gov/gene

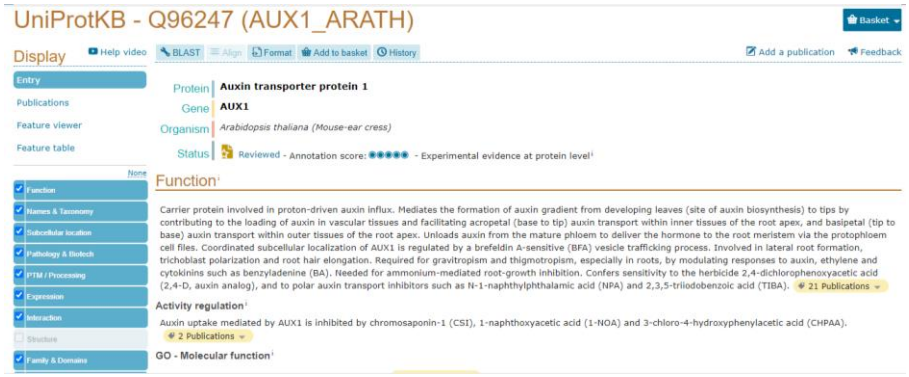

Fig. 2. Captura de la interfaz web de la base de datos UniProtKB.

A través de Urlib se accedió a UniProtKB(Fig. 1.3). Urlib es una librería estándar de Python que contiene funciones para solicitar datos a través de la web y en este caso mediante urllib.request se pudo acceder y abrir el URL correspondiente a UniprotKB[5]. Se definió una función para la búsqueda del número de acceso de UniProtKB a partir del Entrez gene ID.

2.3 Procesamiento

Mediante Beautiful Soup se realizó un raspado web de la página web señalada (Fig. 1.C). Beautiful Soup es una herramienta muy útil para analizar documentos HTML. Esta biblioteca crea un árbol con todos los elementos del documento y puede ser utilizado para extraer información, para nuestro caso particular, la metadata asociada a cada gen utilizando como búsqueda el Entrez gene ID, y obtener como salida los parámetros: nombre completo de la proteína, código de enzima (EC), número de acceso en PDB y resolución, función y proceso de ontología génica (GO), código de familia proteica (Pfam) y de PROSITE.

2.4 Salida

La salida del raspado web se plasmó en un marco de datos que finalmente contenía los identificadores de genes (*gene* IDs, número de acceso de Uniprot), nombre completo de la proteína, código EC, código PDB, proceso y función GO, código Pfam, y código de PROSITE como se muestra en la Tabla 2.

[5] https://www.uniprot.org/uploadlists/

Tabla 2. Expresión diferencial de genes.

Gene ID	Log FC	Pval	En-trez_ Gene _ID	Uni prot	Full Name	E C	K E G G	GO Proc	GO Fu nct	P fam	Pro site	P D B

El marco de datos de salida se descargó con la extensión .csv para posterior análisis. La salida de la búsqueda se guardó en un archivo .json para su posterior consulta en caso de necesidad de incorporar nueva información (Fig. 1.E).

3 Resultados

Para testear la herramienta se utilizó como archivo de entrada una tabla de resultados del análisis de expresión diferencial correspondiente a la colección biológica PRJNA417324. La misma corresponde a un ensayo de secuenciación masiva de ARN hecha con la plataforma Illumina HiSeq4000 (PE100) para 32 muestras de hojas de naranja luego de ser infectadas mediante injerto con yemas infectadas con la bacteria *Candidatus* Liberibacter *asiaticus* [9]. Para el presente estudio se tomó como punto de partida la tabla de expresión generada a partir del procesamiento de las lecturas crudas de las muestras correspondientes a 26 semanas según se señaló en Materiales y Métodos.

La tabla de entrada fue la siguiente:

GeneID	logFC	PValue	entrezgene
CICLE_v10000026mg	-0.9619840295	0.04083529074	18042740
CICLE_v10000062mg	0.93823004430	0.01468090098	18040165
CICLE_v10000069mg	-1.5377512513	8.552040372e-6	18042686
CICLE_v10000103mg	-0.78768715306	0.01399899051	18042773
CICLE_v10000120mg	1.51135267023	0.001992717208	18040840
CICLE_v10000198mg	1.6500864546	5.422522766e-5	NA
CICLE_v10000217mg	0.84226112225	0.02272469192	18043078

Por medio del escarbado web de UNIPROT se obtuvo la siguiente tabla resultante:

Tabla 4. Tabla de salida con expresión diferencial de genes a las 26 semanas del proyecto PRJNA417324. Se incluyen alguna de las anotaciones obtenidas por escarbado web de UNIPROT, entre ellas GO_process, KEGG, EC y GO_function.

GeneID	logFC	PValue	Entrez_gene_ID	Uniprot	Full Name	KEGG	GO_process	GO_function	EC
CICLE_v10000026mg	-0,96	0,04	18042740	V4SZQ7	DUF5311 domain-containing protein {ECO:0000259|Pfam:PF17238}	cic:CICLE_v10000026mg -.	nucleus IEA		
CICLE_v10000062mg	0,94	0,01	18040165	V4T6B4			integral component of membrane IEA		
CICLE_v10000069mg	-1,54	0,00	18042686	V4V007		cic:CICLE_v10000069mg -.			
CICLE_v10000103mg	-0,79	0,01	18042773	V4TDJ1	Cellulose synthase {ECO:0000256|RuleBase:RU361116}	cic:CICLE_v10000103mg -.	cell wall organization IEA	cellulose synthase (UDP-forming) activity IEA, metal ion binding IEA	2.4.1.12 {ECO:0000256|RuleBase:RU361116}
CICLE_v10000120mg	1,51	0,00	18040840	V4TB01	Lon protease homolog, mitochondrial {ECO:0000256|HAMAP-Rule:MF_03120}	cic:CICLE_v10000120mg -.	oxidation-dependent protein catabolic process IEA	ATP-dependent peptidase activity IEA, sequence-specific DNA binding IEA, ATP binding IEA, serine-type endopeptidase activity IEA	3.4.21.53 {ECO:0000256|HAMAP-Rule:MF_03120}
CICLE_v10000217mg	0,84	0,02	18043078	V4V803	Protein kinase domain-containing protein {ECO:0000259|PROSITE:PS50011}	cic:CICLE_v10000217mg -.	integral component of membrane IEA	oxidoreductase activity IEA, metal ion binding IEA	

4 Discusión y Conclusión.

Inicialmente se utilizaron diversos programas para la búsqueda de información funcional y estructural de los genes en bases de datos biológicas, entre ellos se puede nombrar a Biomart [10], Biopython [11] y Bioservices [12]. Los tres paquetes en su desarrollo informático utilizan el wrapping o encapsulamiento, que se basa en la extracción de contenido de una fuente de información en particular, en este caso una base de datos biológica, y posteriormente realizan una traducción en una forma relacional. Sin embargo, cuando se desarrolló la herramienta de búsqueda de metadata, se encontraron ciertas complicaciones a la hora de extraer la información buscada y eso estuvo asociado a: 1) la limitante en las funciones que cada paquete contaba; 2) la necesidad de utilizar varios paquetes para suplir un mismo requerimiento; 3) la elevada demanda de procesamiento que involucra utilizar diversas funciones en simultáneo.

En consecuencia, se decidió utilizar la técnica de web scraping y trabajar con UniProtKB. La misma cuenta con abundante información y datos cruzados de otras bases de datos, por lo que toda la información requerida se encuentra en la misma. Mediante el uso de la herramienta Beautiful Soup se realizó una búsqueda integral, rápida y eficiente de la metadata asociada a cada uno de los genes diferencialmente expresados en el experimento de prueba.

Es importante señalar ciertas limitantes dentro del desarrollo del trabajo, entre ellas que las búsquedas superiores a 25 entradas en ocasiones pueden generar un error por la sobrecarga del servidor UniprotKB, por lo que se debería considerar una condición time.sleep() o algo similar para evitar este evento. Actualmente se fragmentan las búsquedas en listas de 25 unidades y cada tabla resultante tarda menos de 1 minuto en generarse en una computadora convencional.

Como trabajo futuro se pretende automatizar pasos previos referidos al primer procesamiento de datos, entre ellos la adjudicación semiautomatizada del Entrezgene ID a la tabla de expresión diferencial generada por los programas bioestadísticos para tal fin, en este caso particular edgeR [13]. Mediante un escarbado web análogo al ejecutado para UniprotKB se podría obtener esta información de la base de datos Entrezgene, sin tener que hacer una búsqueda manual de aquella que corresponda a la especie de estudio. Además, sería muy interesante poder generar una API (Interfaz de Programación de Aplicaciones) para que sea más fácil el proceder para el usuario.

Disponibilidad de datos y materiales

La herramienta que se desarrolló y utilizó para la búsqueda de metadata asociada a los transcriptos diferencialmente expresados a las 26 semanas para colección biológica PRJNA417324, junto a los paquetes y versiones pertinentes se encuentran disponibles en el siguiente repositorio:

https://github.com/MACHARODRIGO/Buscador-de-metadata.git

Referencias

1. Hoheisel JD (2006) Microarray technology: Beyond transcript profiling and genotype analysis. Nat Rev Genet 7:200–210
2. Marguerat S, Bähler J (2010) RNA-seq: From technology to biology. Cell Mol Life Sci 67:569–579
3. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL (2016) RNA-seq experiments with HISAT , StringTie and Ballgown. Nat Protoc 11:1650–1667
4. Glez-Peña D, Lourenço A, López-Fernández H, Reboiro-Jato M, Fdez-Riverola F (2013) Web scraping technologies in an API world. Brief Bioinform 15:788–797
5. Richardson L (2019) Beautiful Soup Documentation Release 4.4.0. MediaReadthedocsOrg 1–72
6. Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. https://doi.org/10.1155/2008/619832
7. Koski LB, Gray MW, Lang BF, Burger G (2005) AutoFACT: An automatic functional annotation and classification tool. BMC Bioinformatics 6:1–11
8. Muñoz-Mérida A, Viguera E, Claros MG, Trelles O, Pérez-Pulido AJ (2014) Sma3s: A three-step modular annotator for large sequence datasets. DNA Res 21:341–353
9. Chin EL, Ramsey JS, Mishchuk DO, et al (2020) Longitudinal Transcriptomic, Proteomic, and Metabolomic Analyses of Citrus sinensis (L.) Osbeck Graft-Inoculated with " Candidatus Liberibacter asiaticus." J Proteome Res 19:719–732
10. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A (2009) BioMart - Biological queries made easy. BMC Genomics 10:1–12
11. Chapman BA, Chang JT (2000) Biopython: Python tools for computational biology. ACM SIGBIO Newsl 20:15–19
12. Cokelaer T, Pultz D, Harder LM, Serra-Musach J, Saez-Rodriguez J, Valencia A (2013) BioServices: A common Python package to access biological Web Services programmatically. Bioinformatics 29:3241–3242
13. Robinson M, McCarthy… D (2010) edgeR: differential expression analysis of digital gene expression data. BioconductorFhcrcOrg 1–76