Uso y perspectivas en la SfGD: acceso eficiente e integridad de la información

Laura Angelone [1,2], Paula Fernández[3], Alfonso Pons[1], Bernardo Clavijo[3], Claudia Reynares[1], Pilar Bulacio[1,2], Norma Paniego[3], Elizabeth Tapia[1,2]

[1]*Facultad de Cs. Exactas e Ingeniería, Av. Pellegrini 250, Rosario, Argentina*
[2]*CIFASIS-Conicet, Bv. 27 de Febrero 210 Bis, Rosario, Argentina*
[3]*Instituto de Biotecnología, CICVyA, INTA Castelar, Las Cabañas y Los Reseros, (B1712WAA) Castelar, Provincia de Buenos Aires, Argentina*

Resumen
La base de datos SfGD (Sunflower Genomic Database) fue diseñada para almacenar, gestionar y consultar datos e información generados dentro del marco de un proyecto de genómica funcional de girasol llevado a cabo en INTA Castelar. La SfGD facilita el almacenamiento de ESTs y unigenes de girasol, la comparación contra bases de datos de otras especies más profundamente caracterizadas y la recuperación de subclases funcionales basadas en ontologías genéticas. La base se encuentra alojada en el servidor del INTA, bajo el dominio http://bioinformatica.inta.gov.ar/sunflower/, actualmente bajo acceso con usuario y contraseña. En este trabajo presentamos las mejoras que se han realizado a partir de la SfGD versión 2010 referidas a su interfaz web y a la administración de los datos, con el propósito final de la puesta en público. En el aspecto de la interfaz, se agregaron funciones que permiten guardar las consultas, y además se ha incorporado filosofía AJAX para agilizar la visualización de las búsquedas. El objetivo es perfeccionar la visualización y acceso a los datos. En lo referido a la administración de los datos, con la firme convicción de preservar la integridad de los mismos, se realizó un módulo independiente de ABM. La base de datos resultará más segura y los datos ómicos residentes en ella serán fiables. De esta forma consideramos que es inminente la implantación de la base como recurso de acceso público para la comunidad científico/académica interesada en consultar información molecular asociada a esta especie de interés agronómico.

Abstract
The SfGD(Sunflower Genomic Database) was designed to store, manage and consult data and information generated within the framework of a project of sunflower ESTs genomic informationcarried out at INTA Castelar. SfGD facilitates the storage of sunflower ESTs and unigenes, comparison with ESTs from other species vast studied and the recovery of functional subclasses based on genetics ontology. The database is hosted on the server of INTA, in http://bioinformatica.inta.gov.ar/sunflower/, currently accessible with username and password. We present the improvements that have been made from the 2010 version SfGD regarding its web interface and data management, with the ultimate aim of launching it to public. In terms of interface, functions were added to save queries, and also incorporated philosophy AJAX to speed the display of the search. The aim is to improve the visualization and data access. In regard to data management with the firm conviction of preserving the integrity of data, we wrote a separate module of CRUD. The database will be safer and omics data residing in it will be reliable. Thus we consider that the database is ready for its imminent introduction as a public access resource for the scientific/academic community interested in consulting molecular information related to this species of agronomical interest.

Palabras Clave
Base de datos, AJAX, girasol

Introducción

SfGD fue diseñada para alojar datos e información derivada del procesamiento de secuencias ADNc de girasol (*Helianthus* annuus L.) [1]. El conjunto de datos incluye, secuencias de ESTs, unigenes derivados del ensamblado de los mismos, su anotación funcional [2], catálogo de etiquetas de reconocimiento inequívoco para cada unigen y

datos de estudios de expresión ante desafíos frente a estrés abiótico (frío y salinidad) [3].
Los datos en SfGD han sido compilados con el fin de encontrar funcionalidades genéticas ligadas al conjunto de genes representados por los mencionados ESTs. Brevemente, las secuencias ESTs han sido procesadas, ensambladas y agrupadas para definir un conjunto no redundantes de genes, conformando el índice de genes disponible al momento para girasol el cual constituye la base para estudios funcionales basados en un microarreglo de oligonucleótidos de tecnología Agilent en el marco del proyecto en red ANPCyT PAE. Estos genes han sido luego comparados con secuencias de proteínas descriptas y disponibles en bases de datos públicas y de ontologías con el fin de asignar funciones moleculares putativas y anotaciones con vocabulario controlado. Siguiendo estas consideraciones, SfGD permite, i) el análisis de secuencias de ESTs, los contigs y supercontigs obtenidos de los procesos de ensamblado y agrupamiento, ii) la recuperación de anotaciones BLAST[4], y iii) la búsqueda de secuencias a partir de términos de funciones moleculares definidas según la ontología de genes GO (Gene Ontology)[5] iv) la recuperación de unigenes que cuentan con información experimental adicional (ensayos de expresión), v) la recuperación de unigenes representados por etiquetas únicas representadas en el chip de oligonucleótidos

La necesidad de almacenar digitalmente el enorme volumen de datos biológicos que resultan de esta investigación, hizo que se requiriera la creación de una base de datos, denominada SfGD *(Sunflower Genomic Database)*[6], tal que permitiese la carga, organización, actualización y visualización de los datos generados en el proyecto EST de girasol mediante una interfaz web amigable[7].

Cuando se inició el diseño de la SfGD se plantearon dos desafíos: 1) crear una base de datos con un modelo sencillo y eficiente, 2) diseñar un nivel de visualización sencillo y análogo al lenguaje coloquial y/o visual de comunicación. Puntos que fueron robustecidos en reiteradas ocasiones hasta llegar a la versión presentada en este trabajo, tal vez no la última.

La elección de las herramientas de software a utilizar para su implementación se basó en las siguientes premisas: software libre, portabilidad, independencia del sistema operativo, escalabilidad, arquitectura cliente servidor, humanizar la interfaz, conocimiento previo de las herramientas, y aprovechamiento de los recursos de Internet. Inicialmente, y como resultado del análisis de estas premisas, se decidió implementar una base de datos relacional utilizando el paquete de software LAMP (Linux, Apache, MySQL y Programación (PHP + SQL))[9]. MySQL para la creación de la SfGD, PHP para la programación de los clásicos formularios de ABM (Alta, Baja y Modificación), Apache como servidor web y Linux como sistema operativo soporte de los anteriores softwares.

Actualmente, la base se encuentra alojada en el servidor del INTA, bajo el dominio http://bioinformatica.inta.gov.ar/sunflower/, en este momento bajo acceso con usuario y contraseña. La información biológica fue curada y cargada completamente, y está siendo usada por los investigadores del Instituto de Biotecnología de INTA y miembros del PAE37100 en la extracción de información útil para el proyecto genómico.

Avances en la implementación de la SfGD

1) Re-diseño del modelo de la base de datos

El modelo de la base SfGD en su versión original del 2009[6], basado en un modelo teórico, ha sido modificado en varias ocasiones hasta llegar al actual presentado en la Fig.1. Este re-diseño surge como consecuencia de dos hechos relevantes: el primero que

las tablas no reflejaban la dinámica de los datos, y el segundo, atendiendo al enriquecimiento de la misma con nuevos datos producto de la investigación biológica. Se considera que este esquema evidencia una mejor comprensión e interrelación de los datos biológicos disponibles.

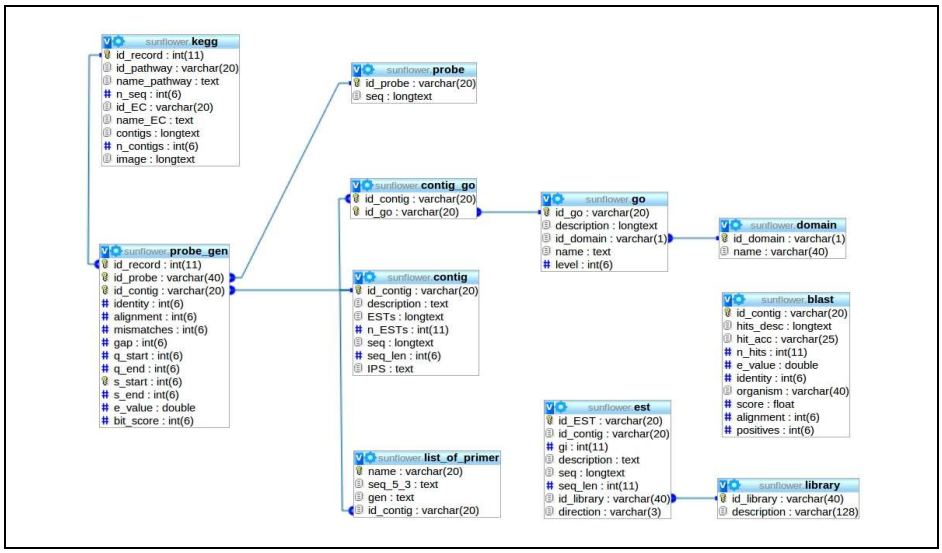

Fig. 1 Modelo relacional de SfGD

2) Independencia del sitio ABM

Para preservar la integridad de los datos, se decidió separar la sección ABM (Alta, Baja y Modificación) y alojarla en un sitio web seguro donde sólo el grupo INTA tiene injerencia. Esta sección estaba presente como opción *Submit Data* en el mismo sitio de usuario final, Fig.2. Esta decisión se tomó pensando en la confiabilidad y seguridad de los datos ómicos para la comunidad científica interesada en consultar información molecular asociada a esta especie de interés agronómico. De esta manera, la base crecerá y se actualizará en función de las nuevas investigaciones, independientemente de la web de usuario final.

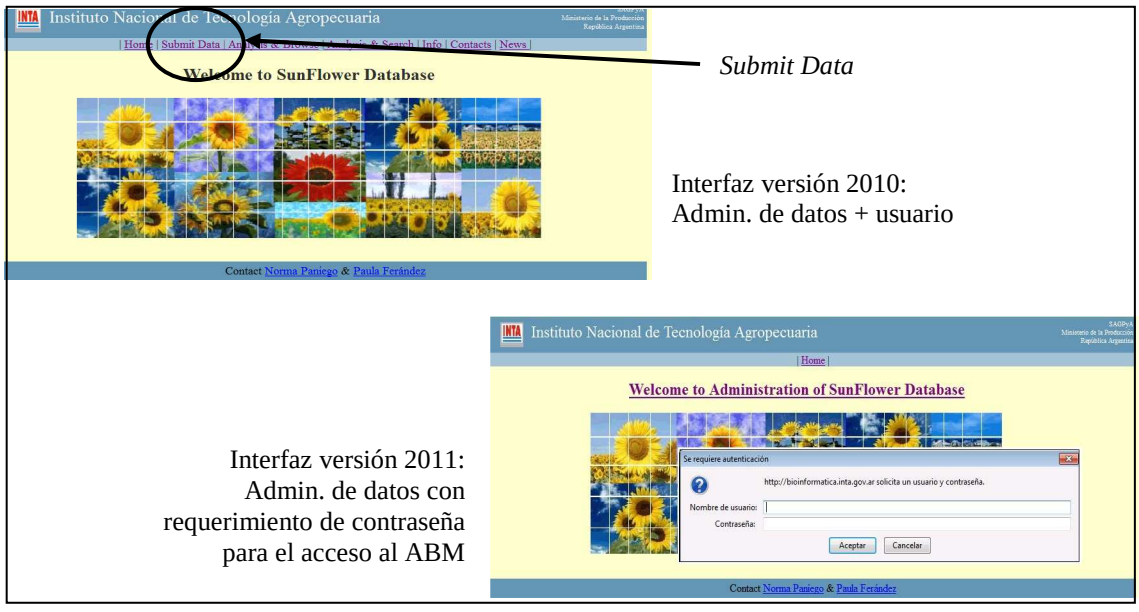

Fig.2 Sitio ABM seguro

3) Avances en la interfaz de consulta

Es indudable que Internet es un medio masivo y rápido para buscar, enviar y recibir información de diversos tipos. En este contexto, nuestro desafío fue construir una interfaz web que funcione como un puente de comunicación informático-biológico, con la premisa de lograr un nivel de visualización sencillo y análogo al lenguaje coloquial y/o visual de comunicación, que permita ser utilizada por un usuario final sin demasiadas explicaciones.

La interfaz web de la SfGD[7] es interactiva, de manera tal que el usuario puede construir una amplia gama de consultas, basándose en que tipo de información necesita obtener de la base de datos, sin necesidad de involucrarse en los aspectos técnicos relacionados con la administración de la misma. Lograr este objetivo requirió de permanentes consultas a las fuentes biológicas para comprender y corregir los programas en pos de mejoras que atiendan a su usabilidad.

En la Fig. 3 puede verse la interfaz de consulta. En la parte superior se selecciona uno a uno los elementos biológicos (*Select from*), el atributo relacionado (*the attribute*), y si se desea se puede realizar una condición de tipo relacional sobre el mismo (*with condition*). Pulsando OK, se van posicionando cada línea de consulta seleccionada y se unen a través de conectores AND u OR. Una vez compuesta la consulta, se pueden ver los resultados (*Submit query*) y/o guardarlos en un archivo de tipo csv (*whole data in csv file*) para luego procesarlos. El resultado de la consulta puede mostrase ordenado por alguno de los campos seleccionados (*sort*).

El botón *Saved queries* permite ver todas las consultas SQL guardadas, para poder reutilizarlas.

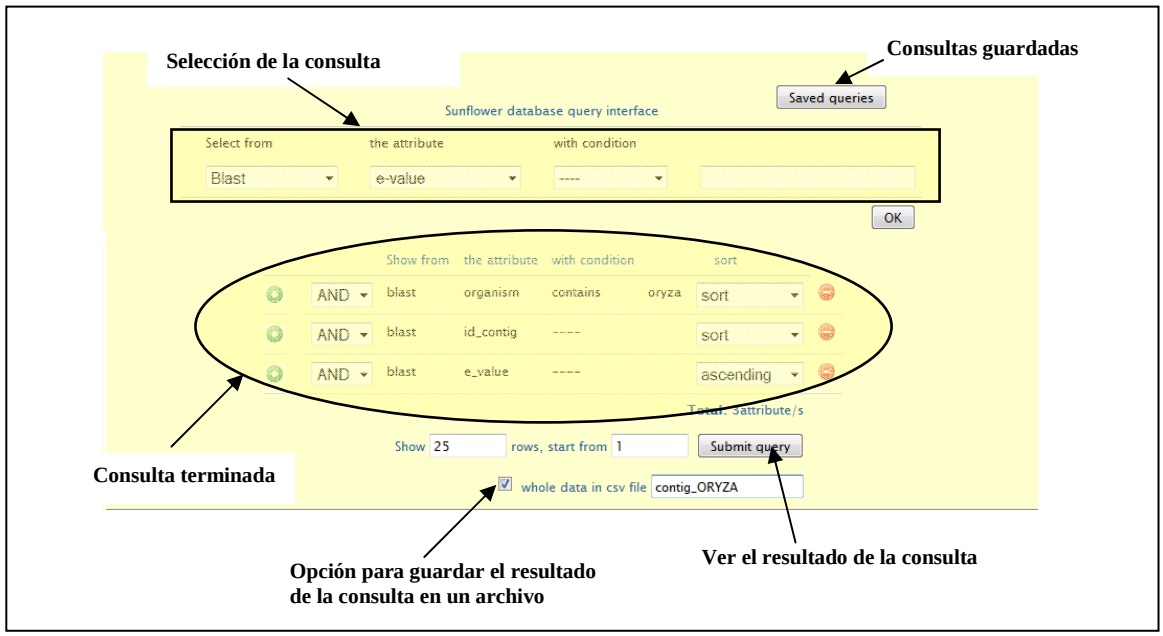

Fig. 3 Interfaz de consulta de SfGD

En la Fig.4 se muestra un ejemplo del uso de esta interfaz de consulta. Se observa una consulta de todos los ESTs públicos identificados con su número de acceso que conforman cada contig. Esta consulta es de fundamental relevancia para conocer e interpretar en qué sentido y con cuales ESTs están conformados los contigs descriptos y almacenados *in silico* por CAP3[8], que fue el software de ensamble utilizado en la construcción del índice de genes para girasol. A esta misma consulta se le agregó un

nuevo patrón de búsqueda e incorporar la anotación GO y un listado de oligonucleótidos descripto para esa secuencia en girasol, el resultado se observa en la Fig. 5.

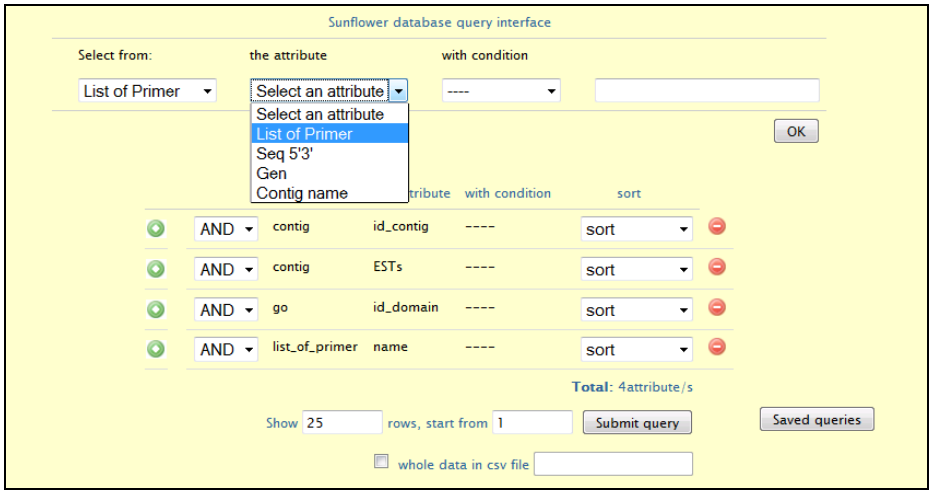

Fig. 4 Consulta de ESTs públicos con anotación GO

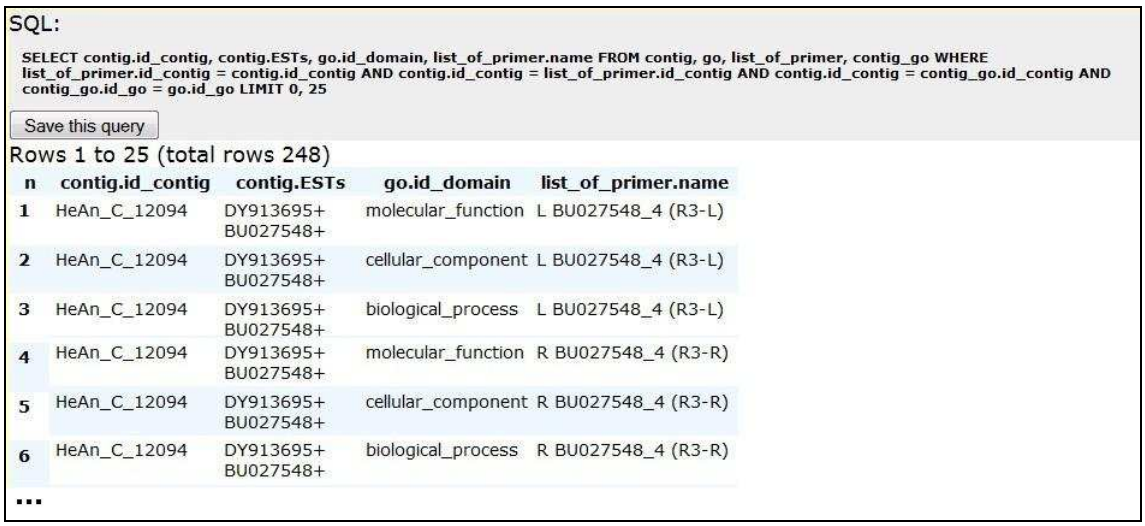

Fig. 5 Resultado de la consulta de Fig.4

En la pantalla de salida de resultados está presente un botón denominado *Save this query* que permite guardar la consulta SQL a fin de utilizarla en otra oportunidad sin necesidad de reconstruirla, ver en Fig.5.

Sin embargo, estas mejoras no fueron suficientes. Se evidenciaba, en algunas consultas complejas, que el tiempo que tomaba la visualización de los resultados era considerable, con la justificada impaciencia de los investigadores. Ante la necesidad de acelerar este proceso se recurrió a la filosofía AJAX, la cual permite actualizaciones asincrónicas y parciales de una página web, dando como resultado aplicaciones web más rápidas.

4) Ventajas introducidas con AJAX.

AJAX es el acrónimo de *Asynchronous JavaScript and XML*, y fue inventado en el 2005 por Jesse James Garrett de la empresa Adaptive Path [11] para designar a la tecnología que sustenta a las aplicaciones Web 2.0 que ofrecen interfaces de usuario más ricas y complejas. Hasta ese momento, no existía un término normalizado que hiciera

referencia a un nuevo tipo de aplicación web que estaba apareciendo. No es una tecnología en si misma, sino una técnica de desarrollo web para crear aplicaciones interactivas, que utiliza otras tecnologías ya existentes: HMTL, JavaScript y XML. Su éxito actual se debe a que los navegadores más importantes han estandarizado un objeto JavaScript, denominado XMLHttpRequest, que permite hacer peticiones al servidor desde la página actual sin recargarla. Google ha sido pionero en el uso de AJAX. Podemos ver su uso en Google Maps o Gmail.

En las clásicas aplicaciones web, todo el procesamiento recae en el servidor. Así, para cada petición de un usuario le supone al servidor un tiempo de procesamiento y el envío de una página web completa al cliente. Cualquier actualización del contenido de la página requiere una recarga completa de una parte de la misma y el usuario debe esperar a que esa recarga se complete. Debido a que el número de páginas que el servidor tiene que procesar y enviar a los clientes crece cuanto mayor sea la interactividad (usuario-aplicación) permitida por la aplicación y según aumenta el número de clientes simultáneos, este tipo de arquitecturas son poco escalables ya que se hacen insostenibles en aplicaciones con mucho intercambio de datos y/o que requieran atender a muchos usuarios simultáneos.

Con AJAX la comunicación con el servidor se implementa mediante el intercambio de mensajes cortos formateados con XML. Esos mensajes son interpretados en el servidor que formatea y envía una respuesta, también en XML, en vez de enviar una página web completa. El mensaje XML recibido por el cliente es interpretado mediante código JavaScript y utilizado para hacer las modificaciones oportunas en la página web actual. Este modo de funcionamiento es mucho más ágil para el usuario. Utilizando filosofía AJAX no es necesario recargar la página sino que sólo se altera su contenido, y el usuario no tiene que esperar a que este cambio se produzca.

Además, el uso de AJAX facilita el trabajo de los servidores de aplicaciones web que tienen que atender cada vez a más usuarios potenciales. Por otra parte, el hecho de reducir la cantidad de información que se intercambia entre el cliente y el servidor permite liberar el ancho de banda y por consiguiente otorga a las aplicaciones web mucha más interactividad que las tradicionales ya que la capacidad de procesamiento del servidor se puede utilizar para atender un mayor número de peticiones simultáneas.

5) Funcionamiento para distintos clientes web

Con el compromiso de hacer público el acceso a la SfGD, además de desarrollar una interfaz que cumpla con todas las premisas antes nombradas (software libre, interfaz amigable y acceso eficiente), exigió probar su funcionamiento en distintos clientes web (Mozilla Firefox, Chrome, Opera, Safari e Internet Explorer).

En una primer versión el funcionamiento fue aceptable en casi todos, salvo en Internet Explorer®[1]. Este comportamiento condicionaba el uso de la plataforma en sistemas Windows®[1], debiendo instalar otro navegador. Los errores de la interfaz provenían en su mayoría de código JavaScript no compatible con Internet Explorer. Para incluir este último navegador se decide utilizar un framework JavaScript, el jQuery[12], que si bien obligó a recodificar, también proporcionó nuevas herramientas para mejorar ergonómicamente la interfaz.

De este modo, la SfGD puede ser accedida con Firefox 2.0+, Internet Explorer 6+, Safari 3+, Opera 10.6+, Chrome 8+.

[1] ® marca registrada de Microsoft Corporation

Desde el punto de vista del servidor no hubo inconvenientes pues al utilizar Apache como servidor web y MySQL como servidor de base de datos, pueden ser instalados en servidores Linux o Windows debido a que existen versiones para ambos sistemas.

Conclusión

Las nuevas capacidades que se han incorporado en el diseño de la SfGD: interfaz de consulta sencilla y robusta, acceso eficiente a los datos, mayor dinamismo en la recuperación de la información con filosofía AJAX y la administración de los datos en forma exclusiva por el grupo de investigación INTA Castelar, acelera la inminente posibilidad de implantación de la base de datos del girasol cultivado como una herramienta de acceso público a la comunidad científico/académica interesada en consultar datos moleculares asociados a esta especie de interés agronómico.

Como trabajo futuro se prevé la incorporación de toda la información correspondiente al microarreglo de girasol, que ya cuenta con datos de hibridaciones de los respectivos laboratorios, con sus correspondientes categorías funcionales y/o genes con expresión diferencial.

Referencias

[1] Fernandez P, Paniego N, Lew S, Hopp HE, Heinz R (2003). Differential representation of sunflower ESTs in enriched organ-specific cDNA libraries in a small scale sequencing project. BMC Genomics 2003, 4:40.

[2] Lew, S., Fernández, P. y Paniego, N, (2008) eBiopipeline: una plataforma abierta para el procesamiento de datos bioinformáticas, Segundas Jornadas Argentinas de Agroinformática (JAIIO), Santa Fe, Argentina.

[3] Fernández P, Di Rienzo J, Fernandez L, Hopp HE, Paniego N, Heinz RA. (2008) Transcriptomic identification of candidate genes involved in sunflower responses to chilling and salt stresses based on cDNA microarray analysis. BMC Plant Biology; 8(1):11.

[4] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., October 1990. Basic local alignment search tool. J Mol Biol 215 (3), 403:410.

[5] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G., May 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nature Genetics 25 (1), 25:29.

[6] Fernández P, Angelone L, Bulacio P, Reynares C., Tapia E, Paniego N. (2009). SfGD: Base de Datos Genómicos de Girasol. 1er. Congreso Argentino de Agroinformática (CAI) JAIIO 2009, ISSN 1852-4850, 126:129.

[7] A Pons, C Reynares, L Angelone, P Fernández, P Bulacio, N Paniego, E Tapia (2010), Evolución de la Interfaz de Consulta de la SfGD. Un Puente de Entendimiento Informático-Biológico, 2do. Congreso Argentino de Agroinformática (JAIIO/CAI). ISSN 1852-4850, 726:733.

[8] Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. Genome Res., 9, 868-877.

[9] Lee, James; Brent Ware (December 2002). Open Source Web Development with LAMP: Using Linux, Apache, MySQL, Perl, and PHP. Addison Wesley. ISBN 0-201-77061-X

[10] Conrad Bessant, Ian Shadforth and Darren Oakley.(2009) Building Bioinformatics Solutions with Perl, R and MySQL. Oxford University Press. ISBN13: 9780199230235

[11] Garrett, J.J. (2005) Ajax: A New Approach to Web Applications. Adaptive Path.

[12] http://docs.jquery.com/Main_Page

Datos de Contacto

Laura Angelone. CIFASIS-Conicet, 27 de Febrero 210 bis, Rosario, 2000, Argentina. angelone@cifasis-conicet.gov.ar.