SISTEMA DE INFORMACIÓN PARA EL MANEJO DE DATOS MOLECULARES EN CAFÉ: I. DESARROLLO Y USO DE HERRAMIENTAS

Por

Luis Fernando Rivera, Carlos Eduardo Orozco, Andrés Chalarca, Álvaro León Gaitán Bustamante & Marco Aurelio Cristancho Ardila[1]

Resumen

Rivera, L.F., C.E. Orozco, A. Chalarca, A.L. Gaitán Bustamante & M.A. Cristancho Ardila: Sistema de información para el manejo de datos moleculares en café: 1. Desarrollo y uso de herramientas. Rev. Acad. Colomb. Cienc. **32**(124): 317-324, 2008. ISSN 0370-3908.

Con la gran cantidad de información que en la actualidad se obtiene en los proyectos de estudio de genomas, es necesario crear una estrategia para que los datos sean almacenados y estructurados de forma que sean fácilmente accesibles.

Para tal efecto se desarrolló en Cenicafé un sistema de información de datos genómicos LIMS el cual, basado en su mayoría en herramientas libres, permitió construir un sistema a bajo costo y de alta calidad, con los aplicativos necesarios para asistir las necesidades en estudios de genómica. Este sistema está especializado para el manejo de información relacionada con secuencias de EST's, BACs y Microsatélites de varias especies de café, la broca *Hypothenemus hampei* y el hongo *Beauveria bassiana*. Para el análisis de la información se crearon "pipelines" específicos para proceder con el agrupamiento, análisis y anotación de las secuencias. Con esta información se generó un modelo relacional de bases de datos para su almacenamiento, se diseñaron interfaces Web con motores de búsqueda especializados y se incorporaron herramientas para despliegue gráfico de ensamblajes de genes, anotaciones, datos estadísticos y otra información relacionada.

El presente sistema es accesible desde la Intranet de Cenicafé mediante un mecanismo de autenticación de usuarios, permitiendo acceder a los datos de forma rápida y eficiente. Este sistema se encuentra en constante cambio debido a las continuas actualizaciones de los datos obtenidos en los proyectos de genoma de Cenicafé y de los datos de secuencias contenidos en los bancos de datos mundiales.

Palabras clave: bases de datos biológicas, bioinformática, software libre, café, LIMS, Sistema Integrador de Manejo de Laboratorio.

1 Centro Nacional de Investigaciones de Café, CENICAFÉ, Plan Alto, Chinchiná, Caldas. Correo electrónico: marco.cristancho@cafedecolombia.com

Abstract

Given the large amount of information that is currently obtained in genome study projects, is necessary to create a strategy to store and organize the data so that it is easily accessible to scientists.

For such effect Cenicafé developed a genomic oriented Laboratory Integrated Management System - LIMS which, based on GPL tools, permitted to build a low cost and high-quality system, with the essential applications needed to fulfill genomic studies. The system is specialized for the management of information related to EST sequences, BACs and Microsatellites of several species of coffee, the coffee berry borer *Hypothenemus hampei* and the fungus *Beauveria bassiana*. For the analysis of the information we wrote specific pipelines to proceed with the grouping, curation and annotation of the sequences. A database relational model was created for storage of the information generated, Web interfaces with specialized search engines were designed and tools for displaying gene assemblies graphics, annotations, statistical data and other related information were incorporated in the system.

The current system is accessible from the Cenicafé Intranet by means of a user authentication mechanism, permitting an efficient and fast way of accessing the data. The system is in constant change due to the continuous updating of the data obtained in the genome projects at Cenicafé and of sequence data stored in world databases.

Key words: biological data base, bioinformatics, open source software, coffee, LIMS, Laboratory Integrated Management System.

Introducción

Con los grandes volúmenes de información actualmente en el orden de Terabytes que se obtienen en los proyectos de estudio de genomas, es necesario crear una manera para que los datos sean almacenados y catalogados de forma que sean fácilmente accesibles. Este almacenamiento debe realizarse de una manera en que los investigadores puedan adquirir y comparar datos particulares almacenados en grandes volúmenes de información. Los datos deben ser también organizados de forma que las relaciones entre ellos sean simples de entender. De igual manera y en lo posible, los datos deben ser almacenados en un lenguaje unificado que evite confusión con datos similares de otros laboratorios para lo cual se optó por SQL para la manipulación de información en bases de datos y XML para la comunicación con sistemas externos. De esta forma es posible en la actualidad, compartir recursos de Bioinformática entre diferentes grupos de investigación sin que la integridad de los datos se vaya a ver en peligro (**Teufel *et al.*,** 2006).

Aunque existen muchos sistemas desarrollados para el manejo de datos moleculares como ESTIMA (**Charu *et al.*,** 2004) o el sistema del SGN en la Universidad de Cornell (**Mueller *et al.*,** 2005), es necesario en la mayoría de los casos integrar varios de estos sistemas en uno propio o desarrollar un sistema totalmente novedoso (**Rhee *et al.*,** 2006). Para el caso de Cenicafé decidimos desarrollar un sistema totalmente nuevo pero teniendo en cuenta los desarrollos de otros grupos, especialmente de la Universidad de Cornell y TIGR (The Institute for Genomic Research), centros que han sido nuestros cercanos colaboradores.

En el presente trabajo se describe el desarrollo de un sistema de información de datos genómicos, sistema accesible desde la Intranet de Cenicafé mediante un mecanismo de autenticación de usuarios, permitiendo acceder a los resultados de los análisis realizados de forma rápida y eficiente. Este sistema se encuentra en constante cambio debido a la producción de nuevos datos en los proyectos de genoma desarrollados en Cenicafé y a las continuas actualizaciones de los datos biológicos contenidos en bancos de datos mundiales.

Materiales y métodos

1. Hardware

Para ofrecer un eficiente servicio a los investigadores se requiere de una plataforma estable y de alto desempeño. Para ello el área de Bioinformática ha adquirido equipos de última tecnología. La plataforma que se utilizó para los desarrollos del sistema está compuesta por:

- 4 IBM e325 con 2 procesadores AMD opteron de 2.4GHz, 4GB de RAM, 2 discos duros SCSI de 74GB c/u.
- 1 IBM X346 con 2 procesadores Intel Xeon EMT64 de 3.6GHZ, 5GB de RAM, 6 discos duros UltraSCSI 320 de 300GB c/u.

- 1 IBM X345 con 2 procesadores Intel Xeon de 3.06 GHz, 4GB de RAM, 6 discos duros SCSI de 74.6 GB c/u.
- 1 SunFire v240.
- 1 Power Mac G5 con 2 procesadores G5 de 2.7GHz, 2 GB de RAM, 2 discos duros serial ATA de 250 c/u.
- 1 Apple xserve quad Xeon de 2GHz, 4GB de RAM, 1 disco duro serial ATA de 750GB.
- 2 IBM Intellistation aPro.
- Rack netbay 42.
- Cisco Catalyst 3750.

El poder de cómputo del sistema se expresa en FLOPS, operaciones de punto flotante por segundo, para cuantificar esto se utiliza la herramienta flops.c desarrollada en el lenguaje C que genera estadísticas de rendimiento del sistema mediante operaciones intensivas de punto flotante:

IBM x346: 1,7 GFLOPS

IBM e325: 3,1 GFLOPS

Power Mac G5: 2 GFLOPS

Apple xserve: 6.1 GFLOPS

IBM Intellistation A pro: 1,1 GFLOP

GFLOPS totales: 1,7+(3.1x4)+2+6.1 +(1,1x2)= 24,4 GFLOPS

En cuanto al almacenamiento de los datos se cuenta con un servidor IBM x346 con 1.2 TeraBytes en RAID 5 al 70 % de su capacidad máxima lo que muestra la gran cantidad de datos procesados por las diferentes herramientas computacionales.

El sistema de respaldo de los datos consta de una unidad de cinta de 800 Gigas de capacidad máxima, adicionalmente se hacen copias de respaldo de los datos más importantes en DVDs.

Esta estructura en hardware es utilizada para prestar servicios de análisis de datos genómicos, procesamiento masivo a través de Grid (varias máquinas computan un mismo proceso simultáneamente), interfaces para consulta de información y almacenamiento de datos. Para lograr la correcta sincronización del sistema, se debe contar con la debida configuración entre cada uno de los componentes del hardware.

2. Software

Para el desarrollo e implementación del sistema se ha utilizado en su mayoría software de distribución libre. El software se puede agrupar de acuerdo a sus características y responsabilidades en: sistema Operativo, Servidores, lenguajes de programación, herramientas de análisis, cluster, Grid y software de desarrollo.

2.1 Sistemas Operativos: como sistemas operativos se tienen instalados Mac OSX 10, Solaris10, Debian 3.1 Sarge y Ubuntu. Se optó por el uso de estos sistemas operativos ya que permiten la administración de los recursos del hardware de una manera eficiente y estable.

2.2 Sistemas manejadores de Bases de datos: MySQL versión 5.0.27 (www.mysql.com) es una de las bases de datos más populares desarrolladas bajo la filosofía de código abierto; PostgreSQL versión 8.1.4 (www.postgres.org): es un servidor de base de datos relacional libre, liberado bajo la licencia BSD (Berkeley Software Distribution).

2.3 Servicio de Directorios: Samba versión 3.0.14a (www.samba.org) permite al usuario final el acceso a archivos, impresoras u otros dispositivos compartidos en una red Intranet o en Internet. NFS (Network File System) es un sistema de archivos distribuido para un entorno de red de área local entre máquinas Unix/Linux. Posibilita que distintos sistemas conectados a una misma red accedan a archivos remotos como si se tratara de locales.

2.4 Servicios Web: Apache versión 1.3.34 y Apache2 versión 2.0.59 son servicios que trabajan a través del protocolo HTTP (Hypertext Transfer Protocol).

2.5 Conexión remota: Open SSH es una herramienta que permite la administración remota de servidores con la característica de que la información transmitida a través de este servicio viaja encriptada.

2.6 Lenguajes de programación: PERL: Practical Extraction and Report Language. Un lenguaje de script de propósito general desarrollado principalmente para la manipulación de texto pero está ampliamente difundido en muchas otras tareas como administración del sistema, desarrollo web, programación para redes entre otras; PHP5: acrónimo de "Hypertext Preprocessor", originado inicialmente del nombre PHP Tools, o Personal Home Page Tools es un lenguaje de programación interpretado utilizado para desarrollo web; XML: XML es el acrónimo del inglés Extensible Markup Language (lenguaje de marcado ampliable o extensible) desarrollado por el World Wide Web Consortium (W3C); JavaScript: es un lenguaje interpretado orientado a las páginas Web, con una sintaxis semejante a la del lenguaje Java; C: lenguaje de programación compilado.

2.7 Herramientas de análisis: se cuenta con una gran variedad de herramientas de análisis y se utilizan para realizar diferentes tipos de análisis, limpieza de vectores y contaminantes, asignación de valores de calidad a las secuencias, herramientas de anotación, alineación, predicción y comparación. Entre ellas podemos encontrar herramientas como Blast (**Altschul *et al.,*** 1990), MPI-Blast (**Darling,** 2008), MPI-Clustalw (**Li,** 2003), Codon Code Aligner (www. codoncode. com/aligner/), GenScan (**Burge y Karlin,** 1997), InterproScan (**Zdobnov y Apweiler,** 2001), Lucy (www.tigr.org), MEME (**Bailey *et al.,*** 2006), MISA (**Thiel *et al.,*** 2006), PHRED (**Ewing *et al.,*** 1998), PRIMER3 (**Rozen y Skaletsky,** 1998), RedBlast (perl script desarrollado en Cenicafé), RepeatMasker (**Smit *et al.,*** 2004), SeqClean (**Pertea *et al.,*** 2006), Tgicl (**Pertea *et al.,*** 2006) y WU-Blast (**Gish,** 2005).

De igual forma el sistema incorpora la herramienta de visualización Gbrowse (**Stein *et al.,*** 2002) la cual permite ver detalles de secuencias largas como seudomoléculas (BACs); en la actualidad esta herramienta despliega detalles de la secuencia completa del cloroplasto de *C. arabica*.

2.8 Cluster y Grid: LAM-MPI: librerías para desarrollo de aplicaciones de procesamiento paralelo, utilizadas en HPC (High Performance Computing) cuya ventaja radica en que divide un proceso en varios más pequeños, permitiendo ejecutarlos en un grupo de máquinas que integran el cluster; por otro lado SGE: Sun Grid Engine, es un sistema de computación paralela que a diferencia del cluster puede estar conformado por máquinas de diferentes plataformas y que además pueden encontrarse localizadas en diferentes sitios.

Actualmente en estos sistemas de computación paralela se ejecutan los procesos de:

InterproScan: usa el sistema grid SGE.

MPI-BLAST: usa el sistema cluster LAM-MPI.

2.9 Software de desarrollo: Macromedia DreamWeaver: es una herramienta IDE (Entorno de desarrollo integrado), que permite desarrollar elementos Web de una forma dinámica y eficiente; SubVersion: herramienta multipropósito para administración de versiones, principalmente utilizada en grupos de desarrollo de software.

Resultados

1. Desarrollo del Sistema

Contamos con una plataforma heterogénea de hardware que reúne la arquitectura Spark de Sun Mycrosystem, Mac de Apple, X86_64 de AMD e Intel, la cual está actualmente conectada a través de una red Gigabit (1Gb/s). El esquema de la plataforma se detalla en la Figura 1.

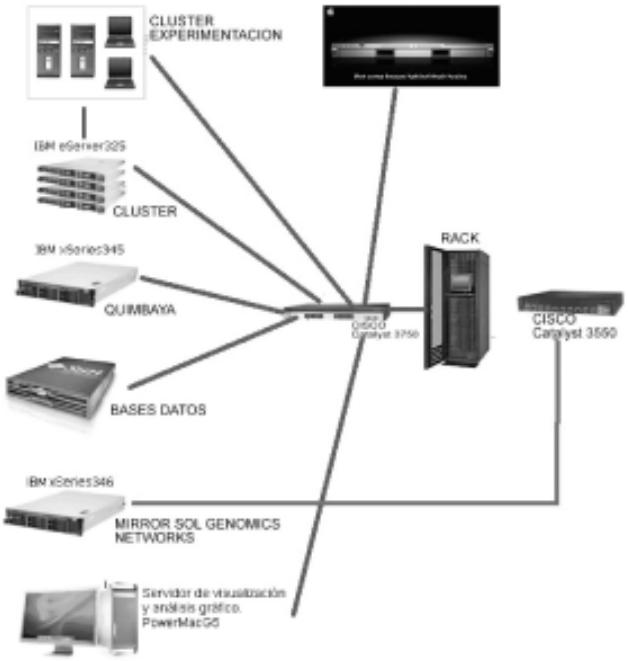


Figura 1. Organización física del Hardware.

La plataforma de software está distribuida teniendo en cuenta la arquitectura con el fin de evitar conflictos entre algunos procesos. El servidor Quimbaya ofrece los servicios Web, directorios compartidos, aplicaciones, autenticación de usuarios y sobre el cual se ejecutan las copias de seguridad. Sobre este además se encuentran instalados los scripts para análisis de secuencias tanto de DNA como de cDNA y el sistema de información como tal. Las bases de datos se encuentran ubicadas en el servidor SUN FIRE V240 denominado TAYRONA, sobre el cual se instalaron los motores de bases de datos MySql y PostgreSql. Para suplir las demandas de procesamiento se optó por la construcción de tres Clusters independientes donde el primero está conformado por 4 IBM Series 345 y se encarga principalmente de procesos como InterproScan y BlastX. El segundo cuenta con dos máquinas MAC Apple Xserver y G5 respectivamente donde además de ejecutar los procesos anteriores se usa para efectos de despliegue gráfico de información. Como complemento a la labor de adaptación de nuevas tecnologías de procesamiento masivo, se cuenta con un cluster de experimentación conformado por dos IBM Work Station. Finalmente se cuenta con un servidor IBM Xseries 246 cuyo único propósito es el de alojar una copia del sistema de información de SGN de Cornell el cual brinda información genómica de diferentes especies de plantas. El sistema para consultas Web está diseñado para manejar gran cantidad de usuarios simultáneos (Figura 2).

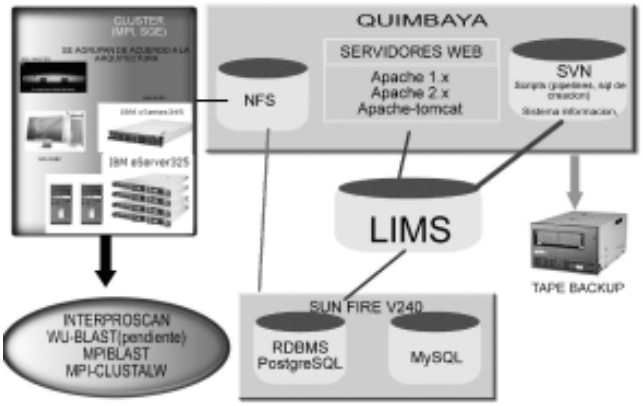


Figura 2. Organización de servicios y herramientas Bioinformáticas.

En la Figura 3 se detalla el sitio de acceso al servidor que contiene las bases de datos y las herramientas de Bioinformática que están disponibles para los investigadores del proyecto del genoma en Cenicafé. El acceso al sitio se realiza por autenticación de palabras claves que son asignadas a los líderes de experimentos de este proyecto.

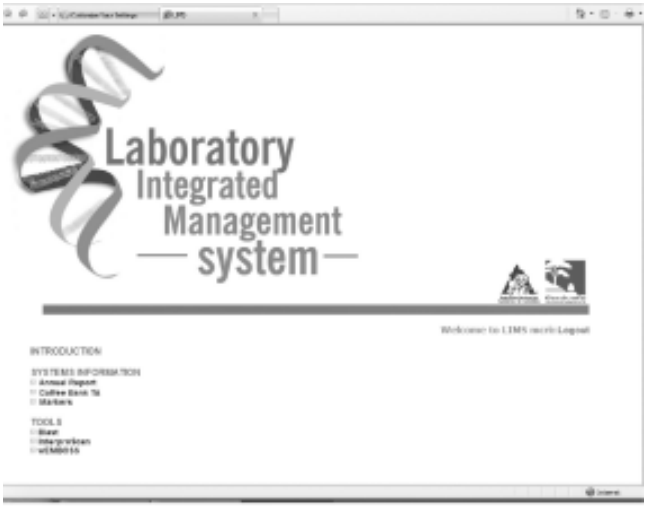


Figura 3. Sitio de acceso interno (quimbaya. cenicafe. org) a las bases de datos y herramientas de Bioinformática de Cenicafé.

Administración del sistema

El sistema actualmente se administra remotamente a través de OpenSSh (Security shell) con autenticación a través de password, a excepción del servidor *mirror*, el cual se detallará más adelante, que se autentica con llave privada DSA de 1024 bits. El DSA es un algoritmo de firma digital, estándar del Gobierno Federal de los Estados Unidos de América o FIPS para firmas digitales. Fue un Algoritmo propuesto por el Instituto Nacional de Normas y Tecnología de los Estados Unidos para su uso en su Estándar de Firma Digital (DSS), especificado en el FIPS 186. DSA se hizo público el 30 de agosto de 1991; este algoritmo, como su nombre lo indica, sirve para firmar y para cifrar información.

A futuro se proyecta crear interfaces de usuario para la administración de servicios como SAMBA, NFS, MPI y SGE y realizar e implementar las reglas establecidas para la realización de copias de seguridad a través de un tape Backup.

Descarga remota de archivos

Las descargas de bases de datos biológicas se realizan de forma automática mediante el uso de programas especializados, mientras que las actualizaciones al Sistema Operativo y software de Bioinformática se efectúan de manera manual según se requiera.

Visualización de la información

Se ha optado por el uso de un ambiente Web para la visualización de todos los análisis realizados en el sistema. Las interfaces para el manejo de secuencias de ESTs han sido dotadas de motores de búsqueda especializados así como herramientas de despliegue gráfico de ensamblajes, todo esto con el fin de adaptarse a las necesidades de los investigadores del centro.

2. Servicios Implementados

wEMBOSS

El wEMBOSS es una interfaz web para el Emboss que es una herramienta biológica para el análisis de secuencias. Esta herramienta es el resultado de un esfuerzo combinado entre el nodo EMBnet de Argentina y el nodo EMBnet de Bélgica (**Rice *et al.,*** 2000).

INTERPROSCAN

InterPro (**Mulder *et al.,*** 2003) es un sistema integrador de diversas bases de datos de familias de proteínas, dominios y regiones funcionales, las cuales identifican características encontradas en proteínas conocidas con el fin de ser utilizadas en proteínas desconocidas y la interfaz de búsqueda del sistema se conoce como InterProScan (**Zbodonov y Apweiler**, 2001). Las anotaciones de Interpro también permiten a los investigadores tener la información de "Gene Ontology" (**Ashburner *et al.,*** 2000) para cada una de sus secuencias.

En nuestro sistema contamos con esta herramienta para que los investigadores puedan hacer análisis localmente, ya que esta herramienta hace uso del cluster que se tiene implementado en el área de Bioinformática de Cenicafé. Actualmente contamos con las bases de datos de InterPro, IprMatches, PIR, Pfam, Pfam-A, Pfam-Clan, PRINTS, Smart, SUPERFAMILY y TIGR.

BLAST a través de la Web

Esta herramienta permite hacer búsquedas eficientes de secuencias de interés dentro de las secuencias generadas en los proyectos de Cenicafé por medio de la implementación Web del programa WU-BLAST, un derivado optimizado del programa original BLAST (**Altschul *et al.,*** 1990).

Servidores de servicio Web externo

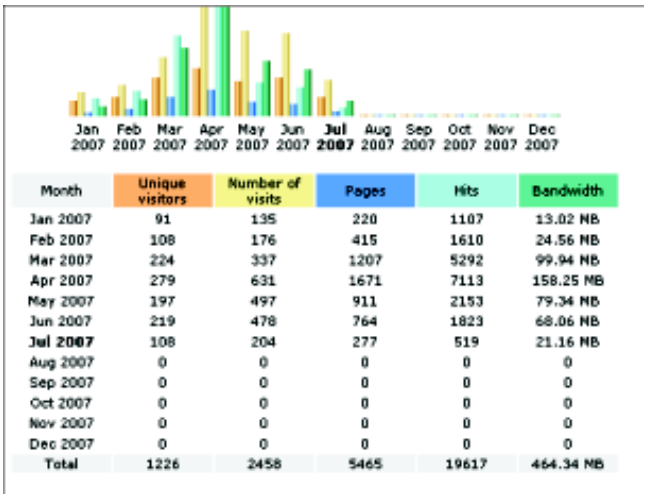
Servidor *mirror* (*sgn.cenicafe.org*)

El sistema puso al servicio externo de la comunidad científica un "mirror" (sgn.cenicafe.org) del sitio SGN de la Universidad de Cornell que estudia las plantas de la familia Solanaceae y otras familias relacionadas. El servidor "mirror" nos ha permitido conocer internamente en detalle la administración de un sitio reconocido de estudio de genomas de plantas e implementar muchos de sus desarrollos en nuestros sistemas de uso interno.

Servidor Web del proyecto del genoma (*bioinformatics.cenicafe.org*)

Este es el sitio de información general del proyecto de estudio del genoma del café. Su interfaz gráfica se puede observar en la Figura 4 y el tráfico que ha tenido en los cuatro primeros meses de uso en la Figura 5. Es de anotarse

Figura 4. Sitio de acceso externo (bioinformatics.cenicafe.org) a la información general del proyecto del genoma de Cenicafé.



Month	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Jan 2007	91	135	220	1107	13.02 MB
Feb 2007	108	176	415	1610	24.56 MB
Mar 2007	224	337	1207	5292	99.94 MB
Apr 2007	279	631	1671	7113	158.25 MB
May 2007	197	497	911	2153	79.34 MB
Jun 2007	219	478	764	1823	68.06 MB
Jul 2007	108	204	277	519	21.16 MB
Aug 2007	0	0	0	0	0
Sep 2007	0	0	0	0	0
Oct 2007	0	0	0	0	0
Nov 2007	0	0	0	0	0
Dec 2007	0	0	0	0	0
Total	1226	2458	5465	19617	464.34 MB

Figura 5. Estadísticas del sitio http://bioinformatics.cenicafe.org, durante el primer semestre del año 2007. Esta estadística de visitas se obtuvo con el programa awstat (http://awstats.sourceforge.net/) que es distribuido bajo licencia GNU-GPL.

que ya se cuenta con un tráfico elevado en este servidor Web, especialmente desde sitios en los Estados Unidos, lo cual aumenta la visibilidad del proyecto de estudio del genoma del café ante la comunidad científica internacional.

Discusión

Cenicafé ha desarrollado un sistema propio para el almacenamiento, análisis y despliegue en sistemas Web de la información de secuencias de ADN de café, la broca del café, el hongo *Beauveria bassiana* y otros organismos que se estudian en el Centro. El sistema ha sido desarrollado casi completamente con software de distribución libre lo que ha permitido rebajar los costos de desarrollo y al mismo tiempo, estar a la par con los centros de estudios de genoma del mundo ya que hemos utilizado un estilo de desarrollo similar gracias a las colaboraciones que hemos tenido con el grupo SGN de la Universidad de Cornell (sgn.cornell.edu) y el Instituto TIGR (en la actualidad el Craig Venter Institute).

El sistema descrito desarrollado en Cenicafé hace parte de las múltiples aplicaciones que en el ámbito mundial son desarrolladas en el área de la Bioinformática haciendo uso de herramientas Open Source (Código Abierto), las cuales en su mayoría bajo licenciamiento GPL (General Public License), permiten la libre distribución de software, reduciendo casi en un 100% los gastos de licenciamiento. Este sistema utiliza algunas herramientas desarrolladas en TIGR, el grupo SGN y otros grupos de investigación de Bioinformática en el mundo; sin embargo, el sistema nuestro

es único ya que está adaptado específicamente para analizar secuencias de diversos tipos de organismos estudiados en Cenicafé, utiliza los despliegues de información que solicitan los investigadores de genoma del centro, y tiene una plataforma Web de despliegue desarrollada completamente en Cenicafé.

En comparación con otros sistemas similares desarrollados, la plataforma de Cenicafé es muy novedosa ya que incorpora herramientas de análisis del wEMBOSS que casi ningún sistema incorpora, incluye el sistema de anotación InterproScan para uso en línea de los investigadores y tiene incorporados sistemas gráficos novedosos como el Gbrowse y otros sistemas gráficos para la visualización de ensamblajes de secuencias de ESTs. Para nuestro conocimiento, es el sistema más avanzado para el análisis y almacenamiento de secuencias de café en el mundo.

Actualmente los sistemas operativos de libre distribución como LINUX y otros licenciados como UNIX brindan el mejor desempeño y estabilidad para procesar y ejecutar volúmenes de información de diversas áreas de conocimiento, sin ser la Bioinformática la excepción. Además de tener en cuenta de que no solo el software licenciado brinda grandes prestaciones a los análisis y procesamiento de información, una muestra de ello son los grandes desarrollos que se encuentran en el mundo del software libre, donde compañías de gran prestigio como SUN, IBM y Novell colaboran en muchos de estos desarrollos.

Debido al gran crecimiento de Internet en los últimos años se ha hecho prácticamente una necesidad el montaje de sistemas basados en Web (**Tanenbaum,** 1987), que con solo el uso de un navegador permita a la comunidad científica desde cualquier punto del planeta acceder a su información, procesarla y redistribuirla si es del caso, con el fin de agilizar sus procesos de producción científica.

La seguridad es un factor decisivo a la hora de concebir sistemas basados en Web, debido a los múltiples ataques que día a día se presentan por parte de "Hackers", los cuales se enfocan en el bien más preciado de una compañía o institución, como lo es la información, la cual debe ser protegida ante cualquier perturbación o divulgación (en caso de ser información confidencial). Para tal efecto el sistema desarrollado cuenta con un sistema de autenticación basado en las normas actuales de privacidad de información, sistema que además de brindar seguridad al acceso de información vía web también cuenta con altos estándares de seguridad al nivel de la plataforma de hardware. Sobre este esquema planteado de seguridad se puede tener un alto control de las actividades que realizan los usuarios sobre el sistema de información.

También se espera mejorar la documentación y administración del sistema, haciendo uso de procesos de ingeniería de software para obtener un alto nivel de representación e interacción de los datos que contiene el sistema de información. De esta forma la meta es a futuro poder brindar un sistema funcional, escalable y seguro a toda la comunidad científica.

Para lograr las futuras metas es necesaria la generación de nueva información, tanto de librerías BAC, microarreglos y proteómica. Teniendo en cuenta que aunque se cuenta con un valioso grupo interdisciplinario, el grupo de Bioinformática necesita explorar más estas áreas de conocimiento para abarcar muchos aspectos que involucran escalar el sistema para administrar y representar este tipo de información. Además los procesos que este tipo de datos demanda en los análisis son demasiado complejos y consumen mucho tiempo de CPU, por ello se requiere un mejor núcleo de procesamiento para obtener resultados en un menor tiempo, sobre todo en los datos referentes a proteómica.

La libertad que ha brindado desarrollar nuestro propio sistema, nos ha posibilitado un mayor control a la hora de escalar el sistema a nuevos procesos y recursos. Donde muchos de estos cambios se efectúan basados en sistemas ya existentes que administran el mismo tipo de información (Básicamente el sistema de la Universidad de Cornell "SGN" y el sistema del TIGR "Transcript Assembly").

Conclusiones

Este trabajo es un buen ejemplo de las ventajas que trae la tecnología de la información a un centro de investigación, permitiendo que sus investigadores puedan compartir información de una forma rápida, flexible y segura, lo cual repercute en una mayor y veloz producción científica que trae grandes beneficios para Cenicafé, otros centros de investigación del país y otras dependencias de Fedecafé como el servicio de extensión. Para nuestro conocimiento, este es el primer sistema de información en Web, desarrollado usando herramientas de software libre, para manejo de datos moleculares en Colombia.

Agradecimientos

A la Dra. Robin Buell y su equipo de Bioinformática en TIGR por su colaboración en todas las etapas del trabajo, Sara Hunter y Martin Sarachu por sus recomendaciones para la instalación de las herramientas de Emboss e InterproScan, y al Dr. Lukas Mueller y el grupo de Bioinformática de la Universidad de Cornell por sus valiosas sugerencias y su ayuda para la implementación del

servidor *mirror*. Agradecemos muy especialmente al Ministerio de Agricultura y Desarrollo Rural por la co-financiación del presente trabajo.

Referencias

Altschul, S.F., Gish,W., Miller,W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol., 215, 403-410.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.

Bailey TL, Williams N, Misleh C, Li W. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research 34: W369-W373.

Burge, C., Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268: 78-94.

Charu G. K, Richard LeDuc, George Gong, Levan Roinishivili, Harris A. Lewin, Lei. Liu, W.M. Keck. 2004. ESTIMA, a tool for EST management in a multi-project environment. BMC Bioinformatics 5: 176-200.

Darling A. 2008. mpiBLAST open source project, http:// MPIBLAST.lanl.gov/

Ewing B, Hillier L, Wendl M, Green P. 1998. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8: 175-185.

Gish W. 2005. WU BLAST 2.0. [http://blast.wustl.edu/blast/ README.html]

Li K-B. 2003. ClustalW-MPI: ClustalW analysis using distributed and parallel computing. Bioinformatics 19: 1585-1586.

Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, Binns J, Lin C, Wright MH, Ahrens R, Wang Y, Herbst EV, Keyder ER, Menda N, Zamir D, Tanksley SD. 2005. The SOL Genomics Network. A Comparative Resource for Solanaceae Biology and Beyond. Plant Physiology 138: 1310-1317.

Mulder NJ, Apweiler R, Attwood RK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, et al. 2003. The Interpro database, 2003 brings increased coverage and new features. Nucleic Acids Res 31: 315-318.

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19: 651-652.

Rhee, S.Y., Julie Dickerson, Dong Xu. 2006. Bioinformatics and its Applications in Plant Biology. Annu. Rev. Plant. Biol. 57: 335-360.

Rice, P., Longden, I., Bleasby, A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16: 276-277.

Rozen S, Skaletsky HJ. 1998. Primer3. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

Smit, AFA, Hubley, R, Green, P. RepeatMasker Open-3.0. 1996-2004 <http://www.repeatmasker.org>.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, et al. 2002. The generic genome browser: a building block for a model organism system database. Genome Res 12: 1599-1610.

Teufel, A, Markus Krupp, Arndt Weinmann, Peter R. Galle. 2006. Current bioinformatics tools in genomic biomedical research (Review). International Journal of Molecular Medicine 17: 967-973.

Thiel T, Michalek W, Varshney RK, Graner A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor Appl Genet 106: 411-422.

Zdobnov EM, Apweiler R. 2001. InterProScan-an integration for the signature-recognition methods in InterPro. Bioinformatics 17: 847-848.

Recibido: octubre 30 de 2007.

Aceptado para su publicación: noviembre 21 de 2008.