SISTEMA DE INFORMACIÓN PARA EL MANEJO DE DATOS MOLECULARES EN CAFÉ: II. DESARROLLO DE BASES DE DATOS

Por

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

Resumen

Rivera, L.F., C.E. Orozco, A. Chalarca & M.A. Cristancho Ardila: Sistema de información para el manejo de datos moleculares en café: II. desarrollo de bases de datos. Rev. Acad. Colomb. Cienc. **32**(124): 325-330, 2008. ISSN 0370-3908.

Cenicafé ha realizado el desarrollo de una plataforma de Bioinformática para el almacenamiento, análisis y fácil accesibilidad de los datos del proyecto de estudio del genoma del café, la broca y el hongo *Beauveria bassiana*. Los proyectos que involucran la construcción de librerías de cDNA y el desarrollo de secuencias de ESTs tienen como objetivo general obtener el catálogo de genes de una especie en particular.

En el presente trabajo se describen las bases de datos desarrolladas en Cenicafé y las estadísticas de las secuencias depositadas en las mismas hasta Mayo del 2007. Estas bases de datos incluyen las correspondientes al análisis de "Trancript Assemblies" a partir de secuencias de ESTs, marcadores moleculares microsatélites (SSRs) y BES (BAC-End sequences). El sistema desarrollado tiene implementadas para su acceso interfaces Web, de fácil acceso y utilización por parte de los investigadores del proyecto de estudio de genoma desde cualquier computador de Cenicafé a través de un sistema de autenticación que permite mantener la seguridad de los datos. El sistema desarrollado permite un amplio crecimiento y el acceso a la información actualizada de cada especie estudiada en forma rápida y eficiente.

Palabras clave: Expressed Sequenced Tags, ESTs, bases de datos relacionales, análisis de secuencia, ensamblaje de transcritos.

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

Abstract

Cenicafé has carried out the development of a Bioinformatics platform for the storage, analysis and easy accessibility of the data of the project that study the genomes of several coffee species, the coffee berry borer and the fungus *Beauveria bassiana*. The projects that involve the construction of cDNA libraries and the development of sequences of ESTs have as a common goal to catalogue the genes of a particular species.

In the present study we describe the databases developed at Cenicafé and the statistics of the sequences stored until May 2007. These databases include the corresponding analysis of "Trancript Assemblies" from sequences of ESTs, microsatellites molecular markers (SSRs) and BES (BAC-End sequences). The system developed has implemented Web interfaces for the access, easy accessibility and utilization for the scientists in genome projects since the system can be accessed from any computer at Cenicafé through a system of authentication that permits to maintain the data secure. The system developed permits an ample growth and the access to the information brought up to date of each species studied in an efficient and fast way.

Key words: Expressed Sequenced Tags, ESTs, relational databases, sequence analysis, transcript assemblies.

Introducción

A pesar de ser uno de los cultivos de mayor importancia a nivel mundial, la investigación genómica en café es muy reducida y existen muy pocas publicaciones sobre el tema en la literatura científica y muy pocas secuencias de su genoma depositadas en bases de datos públicas. En abril del 2007 existían menos de 10.000 secuencias de nucleotidos y secuencias de ESTs de la especie *Coffea arabica* y únicamente alrededor de 50.000 secuencias de la especie *C. canephora* depositadas en GenBank; para tener una comparación, de algunos cultivos como arroz, papa o tomate existen más de 1 millón de secuencias depositadas a la misma fecha. Las secuencias de ESTs son ensambladas en unigenes a través de sistemas especializados; el Instituto TIGR es uno de los sitios pioneros en el ensamblaje y análisis de secuencias de ESTs en secuencias únicas transcritas, sistema denominado "Transcript Assemblies" (http://plantta.tigr.org/). El único depósito relativamente grande de secuencias de la especie *C. canephora* lo realizó Nestlé en conjunto con investigadores de la Universidad de Cornell (**Lin *et al.,*** 2005). Dentro de los avances recientes en secuenciación de la especie *C. arabica*, investigadores de la Universidad de Florida completaron la secuencia del cloroplasto de la especie (**Samson *et al.,*** 2007); el análisis de la secuencia mostró una alta homología del cloroplasto de café con el de especies Solanaceas en cuanto a tamaño, número y orden de genes en su genoma.

Cenicafé ha realizado el desarrollo de una plataforma de Bioinformática para el almacenamiento, análisis y fácil accesibilidad de los datos del proyecto de estudio del genoma del café, la broca y el hongo *Beauveria bassiana*. Los proyectos que involucran la construcción de librerías de cDNA y el desarrollo de secuencias de ESTs tienen como objetivo general obtener el catálogo de genes de una especie en particular (**Teufel *et al.,*** 2006).

En el presente trabajo se describen las bases de datos desarrolladas en Cenicafé y las estadísticas de las secuencias depositadas en las mismas hasta mayo del 2007. En un artículo adjunto se describen las herramientas, sistemas de bases de datos, software y hardware utilizados en el desarrollo del sistema.

Materiales y métodos

El Hardware y Software utilizados para el desarrollo del sistema Web y de las bases de datos se describen en un artículo acompañante.

Desarrollo del Sistema de información y análisis de ESTs, BACs y BES

A continuación se hará una descripción de cada una de las fases en el proceso de análisis de los productos de secuenciación provenientes de los ESTs (Expressed Sequenced Tags), secuencias de librerías BACs y secuencias BES (BAC End Sequences).

Fase 1. Descarga de Cromatogramas ("Trace Files")

Se obtienen los accesos a los respectivos sitios Web donde se encuentran depositados los cromatogramas y se inician las descargas haciendo uso del servidor central (denominado Quimbaya) el cual cuenta con un canal dedicado de 1Mbps.

Fase 2. Asignación de Nombres de Librería genómica

Se procede a renombrar, según sea necesario, cada una de las librerías y cada uno de los cromatogramas, tomando como referencia ciertos patrones de nomenclatura de acuerdo al tipo de librería, los tratamientos de los tejidos usados para su construcción y la especie.

Fase 3. Generación de Matrices de Calidad

Se hace uso del software Phred (**Ewing *et al.,*** 1998) para realizar la asignación de calidad a cada una de las bases que componen los cromatogramas. Como resultado se obtienen dos archivos por librería, uno que contiene las secuencias en formato Fasta y otro que contiene las matices de calidad (rango 0-60).

Fase 4. Limpieza de Secuencias y resumen estadístico de calidad

Haciendo uso del Software Lucy (disponible en http://www.tigr.org/software/sequencing.shtml), se hace la limpieza de vectores y adaptadores usados en el proceso de clonación, además de la generación de códigos de descarte por secuencia (E= Secuencia Vacía, Q= Baja calidad, S= Inserto Corto, P= Inserto corto por PolyA, V= Vector). Las secuencias que no posean código de descarte pasan a la fase 5.

Fase 5. Limpieza de contaminantes

Haciendo uso del software Seqclean, se efectúa una búsqueda y eliminación general de secuencias contaminantes de *Escherichia coli* y las contenidas en la base de datos mundial de vectores Univec (http://www. ncbi. nlm.nih.gov/VecScreen/UniVec.html). Como resultado tenemos las secuencias limpias con sus respectivas matrices de calidad, además de un archivo de referencia para identificar los contaminantes encontrados.

Fase 6. Ensamblaje de secuencias (solo para secuencias de ESTs y BACs)

Haciendo uso del Software Tgi_cl (disponible en http://www.tigr.org/software/other.shtml), se realiza el ensamblaje de los clones resultantes de la fase 5. Como salidas del proceso se cuenta con la información detallada de conformación de Contigs (secuencias ensambladas) y Singletons (secuencias no ensambladas), la cual será posteriormente almacenada en la base de datos.

Fase 7. Enmascaramiento de secuencias (solo para secuencias BACs y BES)

Debido a que las secuencias producto de DNA genómico presentan un alto grado de repeticiones, se hace necesario utilizar la herramienta RepeatMasker (**Smit *et al.*** 1996-2004), con la cual además de marcar las secuencias repetidas, se identifica el tipo y clase de repetición.

Fase 8. Identificación de Marcadores SSRs

Haciendo uso de la herramienta MISA-MicroSatellite identification tool (**Thiel *et al.,*** 2003), se realiza la búsqueda de secuencias SSR o microsatélites (Simple Sequence Repeats), las cuales son utilizadas como marcadores moleculares para la construcción del mapa molecular de *C. arabica*.

Fase 9. Predicción de Genes (solo para secuencias BACs y BES)

Haciendo uso de la herramienta GenScan (http://genes.mit.edu/GENSCANinfo.html) se puede hacer una aproximación en la localización de las regiones que representan genes y su respectiva representación a nivel de proteínas.

Fase 10. Creación del esquema de base de datos

Tomando como base los resultados en cada una de las fases anteriormente señaladas, se hizo la construcción de las bases de datos que internamente se denominan "Coffee TA" y "Coffee BACs" con el fin de almacenar y organizar sistemáticamente la información para análisis posteriores.

Fase 11. Carga de datos en el sistema

Haciendo uso de scripts desarrollados en el lenguaje de programación Perl y usando módulos de BioPerl (www.bioperl.org), se realiza el proceso de carga de información en las respectivas bases de datos.

Fase 12. Generación de estadísticas por librería

Haciendo uso del lenguaje SQL (Standar Query Languaje), se realizan las consultas necesarias sobre las bases de datos para la generación de la información requerida.

Fase 13. Proceso de Anotación

Tomando como referencia los Contigs generados en cada uno de los sistemas se procede a efectuar los correspondientes análisis de homología de secuencia con la herramienta MPI-BLASTX , con el fin de asociar a cada Contig una posible función.

Fase 14. Análisis Comparativo

Usando las herramientas WU-Blast (http://blast.wustl.edu/) y Blat (http://genome.ucsc.edu/goldenPath/help/blatSpec.html), se ejecutan los alineamientos necesarios para la identificación de secuencias homólogas con

secuencias de otros organismos en las secuencias de librerías BACs y secuencias BES.

Resultados y discusión

Ensamblaje de secuencias de ESTs

La construcción de ensamblajes de secuencias de ESTs se basó en el sistema "Transcript Assemblies - TAs" desarrollado en TIGR (http://plantta.tigr.org/). Las secuencias que se usan para construir los TAs son transcritos expresados colectados de librerías de cDNA (secuencias EST s). A diferencia de los ensamblajes de TIGR, en los análisis de Cenicafé no incluimos secuencias que no tengan cromatogramas de los cuales podamos deducir datos de calidad de secuencia; los ensamblajes de TIGR se realizan a partir de secuencias de la división dbEST de la base de datos de GenBank.

Los TAs fueron ensamblados por medio de la herramienta TGICL (**Pertea *et al.***, 2003), el programa especial de BLAST, Megablast (**Zhang *et al.***, 2000) y el programa para ensamblaje de secuencias CAP3 (**Huang y Madan**, 1999). TGICL es un programa que invoca en su programación Megablast y CAP3. Las secuencias son inicialmente agrupadas basadas en comparaciones de todas contra todas, con Megablast; los agrupamientos son ensamblados en "clusters" para la generación de secuencias consenso usando CAP3. Los criterios de ensamblaje incluyeron una región mínima de traslape de 50bp, 95% de identidad mínima en esta región y máximo 20bp que no traslaparan en regiones de homología.

Cualquier secuencia EST/cDNA que no se agrupa en TAs es incluida como "singleton". Los "singletons" corresponderán a los números de acceso de GenBank si son depositados allí. Las identificaciones de los TA de CENICAFE son de la forma Número_de_ TA/taxonID, donde el número de TA es una identificación consecutiva y taxonID representa la identificación de la base de datos taxon del NCBI.

Para proporcionar anotación para los TAs, cada TA/singleton fue alineado a una versión enmascarada de la base de datos UniProt/Uniref100. Las alineaciones requirieron tener por lo menos 20% de identidad y 20% de cubrimiento de secuencia. La anotación para la proteína con la mejor alineación a cada TA o singleton se utilizó como la anotación para esa secuencia. Adicionalmente, la orientación relativa de cada TA/singleton a la mejor secuencia de proteína se utilizó para determinar la orientación de cada TA/singleton. Algunas secuencias no alinearon con ninguna secuencia de las bases de datos de proteínas con los criterios de calidad utilizados por lo que esas secuencias no tienen ni anotación funcional ni orientación. Cada vez que se reciben datos de secuencia de nuevas librerías se procede a realizar un nuevo ensamblaje de esa especie en particular, aunque las versiones pasadas de ensamblajes siguen siendo almacenadas en la base de datos.

Las secuencias ensambladas de *C. arabica* fueron utilizadas para análisis BLAST de homología con secuencias de especies de la familia Solanaceae, cercana filogenéticamente a la familia del café (Tabla 3). El mayor porcentaje de secuencias homólogas se obtuvo con la especie *Solanum lycopersicum* (55%) y se obtuvieron porcentajes de similaridad muy variables entre todas las especies. En esta tabla se puede observar que los porcentajes de homología entre conjuntos de secuencias de diferentes especies son muy dependientes del número de secuencias analizadas; cuando se analiza un número pequeño de secuencias los porcentajes son bajos y cuando se analiza un número de secuencias alto este número aumenta. Es aconsejable analizar un número de secuencias similar entre dos conjuntos de secuencias de dos especies diferentes para poder tener una idea real sobre su relación. Un análisis más completo de homología, aunque con un set más pequeño de secuencias de *C. arabica* se publicó recientemente (**Montoya *et al.*** 2006).

Análisis de secuencias BES (BAC-end Sequences)

La construcción y secuenciación de la librería BAC de *C. arabica* var. Caturra se realizó en la Universidad de Arizona por parte del grupo liderado por el Dr. Rod Wing. Se secuenciaron un total de 86.782 BES de los cuales se seleccionaron por calidad (phred>20) un total de 81.444; el porcentaje de secuencias de buena calidad fue por tanto de 94%. En este grupo se identificaron adicionalmente un total de 488 secuencias de organelos (mitocondria y cloroplasto) por lo que la contaminación de la librería con este tipo de secuencias es despreciable (<1%).

El número final de BES después de estos filtros de limpieza fue de 80.056 secuencias con un tamaño promedio de 619bp. De estas secuencias un total de 38.199 corresponden a clones que tienen la secuencia de los 2 extremos BES y 3660 corresponden a clones que solo poseen un extremo BES secuenciado. Se identificaron un total de 2169 secuencias BES que poseen microsatélites de 2-5 unidades repetitivas, predominando los dinucleotidos con un 57% del total de microsatélites identificados. Estos microsatélites están siendo utilizados en la construcción del mapa molecular de *C. arabica*.

Tabla 1. Estadísticas de los ensamblajes de secuencias de ESTs depositadas en las bases de datos de bioinformática a mayo del 2007.

	N° de Contigs	N° de Singletons	N° de Unigenes	Promedio Longitud de Contigs	Máxima Longitud de Contigs	Mínima Longitud de Contigs	Promedio Longitud de Singletons	Máxima Longitud de Singletons	Mínima Longitud de Singletons	ESTs/Contig
Coffea arabica	8709	11802	20511	950 bp	3943 bp	102 bp	600 bp	1021 bp	101 bp	5
Coffea kapakata	496	1551	2047	741 bp	1786 bp	144 bp	579 bp	883 bp	101 bp	5
Coffea liberica	3231	7551	10782	1015 bp	3478 bp	102 bp	730 bp	1007 bp	101 bp	3
Beauveria bassiana[1]	2297	4119	6416	733 bp	2549 bp	110 bp	437 bp	779 bp	101 bp	5
Hypothenemus hampei	216	698	914	591 bp	1614 bp	156 bp	411 bp	806 bp	101 bp	7

1. Parte de las secuencias de B. bassiana fue obtenida en la Universidad de Florida y parte en Cenicafé.

Tabla 2. Número de secuencias depositadas en la base de datos de marcadores moleculares, marcadores que han sido polimórficos en la especie *C. arabica* y marcadores que han sido utilizados en la construcción del mapa de la especie *C. arabica*.

Organismo	Marcadores Polimórficos	Marcadores Mapeados	Número de secuencias depositadas
Coffea arabica	158	86	3950
Coffea canephora	100	36	986
Coffea kapakata	2	2	64
Hedyotis spp.	12	-	326
Total	272	124	5478

Tabla 3. Análisis de homología de secuencias de ESTs de *C. arabica* con secuencias de especies de la familia Solanaceae. El análisis se realizó con el programa BLAST con un valor $E=1e^{-}$

	N° de secuencias	Hits	%
Capsicum annuum	13176	3888	36%
Hedyotis centranthoides	4311	2097	19%
Solanum lycopersicum	45585	5942	55%
Solanum pennellii	3370	1327	12%
Nicotiana sylvestris	6773	2549	24%
Nicotiana tabacum	38612	5494	51%
Solanum chacoense	5450	2796	26%
Solanum habrochaites	3553	2214	21%

Bases de Datos

En las Tablas 1 y 2 se describen detalles del número de secuencias de EST´s y microsatélites respectivamente, depositadas en las bases de datos que se han desarrollado en Cenicafé a mayo del 2007. En este proyecto se implementaron bajo entorno Web bases de datos con la información generada en el proyecto de estudio del genoma del café y se espera que en el futuro el sistema creado permita implementar otras bases de datos. Las bases de datos que se han implementado hasta la fecha son:

Marcadores moleculares: esta base de datos contiene la información generada en el proyecto de identificación y utilización de marcadores microsatélites y otros tipos de marcadores para la construcción de un mapa molecular de café. En la actualidad esta base de datos contiene un total de 5.478 registros.

ESTs: esta base de datos contiene la información de los ESTs (Expressed Sequenced Tags) que son secuencias parciales de genes que se han generado en café, broca y *B. bassiana*. Dentro de las tablas que esta base de datos

contiene se encuentran los resultados de los ensamblajes de ESTs, información detallada del proceso de limpieza de las secuencias y la información de las librerías de las que provienen los ESTs. La base de datos está compuesta por las tablas Library, Raw_Clones, Transcripts, Contigs_Link, Contigs, Annotation, Analysis_Annotation, Contamination, Releases, Contig_History.

Las estadísticas de las secuencias que actualmente están depositadas en las bases de datos de bioinformática se detallan en la Tabla 1. La especie con el mayor número de secuencias depositadas en las bases de datos de EST´s es *C. arabica*, aunque al hacer un cálculo del porcentaje del genoma (1400 Mbp) de esta especie que ha sido secuenciado solamente está entre el 5-6%; sin embargo, aunque la especie *B. bassiana* tiene un número mucho menor de secuencias en la base de datos el porcentaje del genoma que ha sido secuenciado es de alrededor del 25%, debido a su pequeño tamaño de genoma (aprox. 12-20 Mbp). Es de anotar que la calidad de las librerías construidas ha mejorado notablemente en los últimos años así como la calidad en la secuenciación; estas mejoras tecnológicas han redundado notoriamente en el aumento de secuencias de buena calidad y de mayor longitud depositadas en las bases de datos (datos no mostrados). Algunas de las librerías de cDNA recientemente secuenciadas fueron construidas por la compañía Evrogen (Moscú, Rusia) a través de una novedosa técnica de normalización, la cual permitió obtener librerías con menor redundancia y por tanto mayores tasas de descubrimiento de genes.

Conclusiones

Fue posible implementar un sistema de bases de datos bajo entorno Web para el almacenamiento de datos de secuencia de varias especies de café, la broca y el hongo *B. bassiana*, basado principalmente en herramientas GNU-GPL.

El sistema es de fácil acceso y utilización por parte de los investigadores del proyecto de estudio de genoma desde cualquier computador de Cenicafé a través de un sistema de autenticación que permite mantener la seguridad de los datos.

El sistema desarrollado permite un amplio crecimiento y el acceso a la información actualizada de cada especie estudiada en forma rápida y eficiente.

Agradecimientos

A la Dra. Robin Buell y su equipo de bioinformática en TIGR por su colaboración en todas las etapas del trabajo y al Ministerio de Agricultura y Desarrollo Rural por la cofinanciación de este estudio.

Referencias

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

Huang X.; Madan A. 1999. CAP3: A DNA Sequence Assembly Program. Genome Research 9: 868-877.

Lin C, Mueller LA, Mc Carthy J, Crouzillat D, Pétiard V, Tanksley SD. 2005. Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. Theor Appl Genet. 112: 114-130.

Montoya, G.; Cristancho, M.A.; Moncada M.P. 2006. Análisis de secuencias de genes de C*offea arabica* var. Caturra. Cenicafe 57: 79-87.

Pertea G, Xiaoqiu Huang, Feng Liang, Valentin Antonescu, Razvan Sultana, Svetlana Karamycheva, Yuandan Lee, Joseph White, Foo Cheung, Babak Parvizi, Jennifer Tsai, John Quackenbush. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19: 651-652.

Samson, N., Michael G. Bausher, Seung-Bum Lee, Robert K. Jansen, Henry Daniell 2007. The complete nucleotide sequence of the coffee (*Coffea arabica* L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms Plant Biotechnology Journal 5 (2): 339-353.

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

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.

Zhang Z., Schwartz S., Wagner L., & Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol 2000; 7(1-2): 203-214.

Recibido: octubre 30 de 2007.

Aceptado para su publicación: noviembre 21 de 2008.