Development of a Bioinformatics Platform at the Colombia National Coffee Research Center

M. CRISTANCHO, L. RIVERA, C. OROZCO, A. CHALARCA, L. MUELLER

1National Coffee Research Center, CENICAFE, Planalto, Chinchiná, Caldas, Colombia
2Plant Breeding Department, Cornell University, Ithaca, NY, 14850.

Summary
We have implemented a web-based Bioinformatics platform that functions as a genomics information resource for coffee and other organisms studied at the Colombia National Coffee Research Center - CENICAFE.

The Bioinformatics platform includes a Laboratory Integrated Management System (LIMS), the implementation of wEMBOSS, home-developed perl tools for data analysis, InterproScan for annotation of sequence domains, and the implementation of wBLAST and wNetBLAST among other tools available. The main backbone of the system is an adaptation of the SOL Genomics Network (SGN) databases developed at Cornell University for ESTs, molecular markers and BAC sequences storage and analysis (http://sgn.cornell.edu). The system is based on the postgresQL relational database, the use of perl scripts for the manipulation of data, the Apache Web server with the mod_perl integrated perl interpreter, and the servers run the Debian distribution of the GNU/Linux operating system.

Although SGN has mainly developed as a plant genomics oriented resource, the Cenicafe platform has implemented several new tools and databases for the analysis of other organisms sequence data such as fungi and insects.

The Cenicafe databases contain to date 32,000 coffee EST sequences from 22 libraries organized in 9,257 C. arabica and 1,239 C. liberica unigenes, 6,000 Beauveria bassiana EST sequences organized in 2,404 unigenes, and 4,000 Hypothenemus hampei (coffee berry borer) EST sequences organized in 885 unigenes, besides the more than 100,000 Solanaceae unigene sequences annotated at SGN. The sequences are annotated based on Solanaceae, Arabidopsis, Swissprot and Genbank sequence comparisons using BLAST homology searches, aminoacids are predicted using ESTScan, the domains are annotated using InterproScan and Gene families are annotated using a perl script developed at SGN.

The system will implement in the near future a database of coffee genetics resources developed at Cenicafe, a proteomics platform, and a Microarray database. We will also be incorporating other components to the platform specially for the visualization of genetic maps from the Gmod project (Gbrowse), the SGN system, TIGR, and other open source projects.
Introduction

Coffee is one of the most important agricultural commodities in the world, providing large resources for the economies of many developing countries. Despite its global importance, very little information has been gathered from this plant at the genetic level. As of July 2006, roughly 3,000 DNA sequences from the species *C. arabica* had been deposited in the GeneBank database. Only recently, a large Expressed Sequenced Tag data set from the species *C. canephora* developed jointly by Nestlé and Cornell University scientists were deposited in public databases (Lin et al. 2005).

Genomic research is a field that continuously faces the problem of storing, indexing and retrieving large amounts of data; fortunately for bioinformaticians there is a trend in the field to rely to a greater extent on standard methods for the analysis of this data. It is possible nowadays to share Bioinformatics resources between different research groups and the integrity of the data is not jeopardized in anyway (Teufel et al. 2006).

ESTs are being produced for a number of plants as a rapid method for gene discovery. For instance rice has more than 1 million EST sequences and there are 12 plant species, most of them grasses, with at least 200,000 EST sequences in dbEST (release 082506, August 2006). The ultimate aim in most projects is to catalogue all the expressed genes in a particular genome.

The genus Coffea includes two cultivated species of economic importance, *C. arabica* L. and *C. canephora* Pierre. *C. arabica* (2n = 4X = 44) is an amphidiploid formed by a recent event of hybridization between the diploid species *C. eugenioides* and *C. canephora* (Lashermes et al. 1999); all other Coffea species are diploid (2n = 2X = 22). ESTs and microsatellite markers have not been extensively developed in coffee as in other crops. Only eleven microsatellite markers were obtained by Combes et al., (2000) and they have been used for the study of allele number and heterozygosity level in several diploid and tetraploid coffee species.

The aim of the present work was the development of a Bioinformatics platform for the storage, comprehensive analysis and easy retrieval of molecular data generated in Colombia from the Coffee Genome Initiative taking place in Cenicafe.

Development of the Platform

The coffee genomics project started in 2003 in Colombia with the financing of the Ministry of Agriculture and the National Coffee Growers Federation. The main outcomes expected from this research include the development of molecular tools and markers for coffee, construction of a *C. arabica* genetic map, identification of agronomic important genes, and the development of a Bioinformatics platform to store and analyse the data generated in the project. A major part of the molecular tools been developed involve; the generation of a large set of ESTs; construction, fingerprinting and sequencing of a *C. arabica* BAC library; and the detection of microsatellite, COS and SNP markers. We have also incorporated as part of our genomics research the development of tools to study the genomes of *Hypothenemus hampei* and *Beauveria basiana*.

From the start, the Bioinformatics platform has been a major component of the genomics research at Cenicafe. Our efforts have concentrated in the development of
relational databases, tools for data analysis, and web-based user-friendly interfaces to access data, based on open source technology. We use commercial software just in very special scenarios of our analysis. We have engaged in close collaborations with research groups that adopt this kind of approach in their Bioinformatics developments and consequently our main partners in this area are the Solanaceae Genomics Network based in the Department of Plant Breeding at Cornell University and The Institute for Genomic Research – TIGR in Rockville, Maryland.

**Computer resources:** We have a cluster that consists of a master server (IBM x346. Disk space RAID5: 1.2 TB, RAM: 5 GB, CPU: 2 x 3.6GHz Intel Xeon) and 7 server nodes for data processing (Opteron e-325, 64 bits). For the functioning of the cluster, we use NFS (Network File System) and samba to share directories.

**Software resources:** The master server runs the Apache webserver, most machines run Debian GNU/Linux as operating system, and generally we develop programs in perl for processing data, and perl-cgi and PHP for Web development. Our first Database Management System was constructed using the MySQL relational database and we are currently migrating to the PostgreSQL database system. We implemented the applications MPI-BLAST and ClustalW-mpi to run on the cluster. By running these processes on the cluster we have calculated up to 6 times reduction in processing of data.

The Bioinformatics group efforts are concentrated in two major areas, service and production (development). The service routines include the analysis of sequence and other types of data produced by Cenicafe genomics scientists. The development activities include the setting up and administration of Bioinformatics servers, construction of structured databases, development of web-based interfaces for the display of data and the writing of scripts in perl and other languages for the manipulation of data.

**Implementation of Software**

The Bioinformatics system is accessible through a web-based interface from which all databases and tools are available (Figure 1). The system is built-in a Laboratory Integrated Management System that incorporates a project administration resource, coffee and other organisms databases (SGN Schema), the wEMBOSS suite of tools (Rice et al 2000) and local implementations of BLAST (Altschul et al., 1990) to run particular types of analysis (wBLAST, BLASTXtract and wNetBLAST among others).

**Data Analysis**

While the Sol Genomics Network has mainly developed as a plant genomics oriented resource, the Cenicafe platform has implemented several new tools and databases for the analysis of other organisms molecular data such as fungi and insects. The core of the system is to date the SGN database schema, but our platform has incorporated several additional modules to annotate fungi and insects, given that as mentioned above, genomics data is also been produced from *H. hampei* (the coffee berry borer) and *B. bassiana* (biological control agent).
ESTs are analyzed based on an adaptation of the SGN pipeline (Mueller et al. 2005). In synthesis, chromatograms are called with phred (Ewing et al., 1998), assemblies are performed with CAP3 (Huang and Madan 1999), full length EST sequences are computed by TargetIdentifier (Min et al. 2005), aminoacid prediction is accomplished by ESTScan (Iseli et al. 1999), and functional annotation of sequences is performed with several databases among them GenBank, and more specialized databases like Solanaceae and Arabidopsis for plants, Sacharomyces and Magnaporthe for fungi, and Drosophila and Tribolium for insects. Additional functional annotation of Gene Ontology terms (Ashburner et al., 2000) is performed in house with InterProScan (Mulder et al., 2003; Zdobnov and Apweiler 2001). Several steps of the process include scripts written in perl at SGN and Cenicafe.

Other Bioinformatics analysis include the discovery of SSR markers, development of specific PCR primers, prediction of SNPs and homology comparisons between large sets of sequences. SNP prediction is accomplished through the comparison of several homologous sequences and their visualization is performed with the software CodonCode Aligner (CodonCode Corporation, Dedham, MA).

Table 1 illustrates the number of ESTs analyzed sequences deposited in Cenicafe databases. These numbers are continuously increasing and there will shortly be an update of the databases to incorporate a number of new coffee ESTs and BAC-end sequences. It is possible to retrieve the data from the databases in very specific ways according to the scientists needs; an example of a tissue-specific expression analysis of *C. arabica* transcripts is shown in Table 2.

Cenicafe databases also include more than 1000 coffee microsatellite sequences used for the construction of coffee genetic maps and diversity studies and data gathered from COS and SNP markers studies.

**Final remarks and Future Prospects**

The system will implement in the near future a database of coffee genetics resources developed at Cenicafe, a proteomics platform, and a Microarray database. We will also be incorporating other components to the platform specially for the visualization of genetic maps from the Gmod project (Gbrowse “Generic Genomic Browser”, Stein et al 2002), the SGN system, TIGR, and other open source projects. A BAC relational database is in the process of construction and it will include over 60,000 BAC-end sequences been generated at Arizona University.

We are in the process of mirroring the Solanaceae Genomics Network site (http://sgn.cornell.edu/) that will be accessible from (http://sgn.cenicafe.org/). The projected web interface of the mirror site can be viewed in Figure 2. We will have to develop ways of defining complex interactions, functional annotation and integration of proteomics, microarray and other data that will emerge from the project. Several new ways of integrating these data are emerging (Rhee et al. 2006).
Figure 1. Web interface of the Cenicafe LIMS system which integrates access to databases and analysis tools.

Figure 2. Web interface of the SGN mirror at Cenicafe. The service of this mirror must be available in the next few months.
### Table 1. Number of EST sequences deposited in Cenicafe databases

<table>
<thead>
<tr>
<th>ORGANISM</th>
<th>LIBRARIES</th>
<th>CHROMATOGRAMS</th>
<th>UNIGENES</th>
</tr>
</thead>
<tbody>
<tr>
<td>COFFEE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Coffea arabica</em></td>
<td>17</td>
<td>35.000</td>
<td>9.257</td>
</tr>
<tr>
<td><em>Coffea liberica</em></td>
<td>4</td>
<td>3.613</td>
<td>1.239</td>
</tr>
<tr>
<td><em>Coffea canephora</em></td>
<td>5</td>
<td>47.000</td>
<td>13.750</td>
</tr>
<tr>
<td><em>Coffea spp.</em></td>
<td>1</td>
<td>497</td>
<td>210</td>
</tr>
<tr>
<td>FUNGI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Beauveria bassiana</em></td>
<td>8</td>
<td>5.300</td>
<td>2.404</td>
</tr>
<tr>
<td>INSECTS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Hypothenemus hampei</em></td>
<td>2</td>
<td>3.563</td>
<td>885</td>
</tr>
</tbody>
</table>

### Table 2. BLAST homology searches of *C. arabica* Unigene sequences, number of ESTs that compose each unigene and representation in 3 tissue-specific libraries.

<table>
<thead>
<tr>
<th>Unigene No.</th>
<th>No. Members</th>
<th>FR</th>
<th>L</th>
<th>FL</th>
<th>BLAST</th>
<th>ORGANISM</th>
</tr>
</thead>
<tbody>
<tr>
<td>sgn\U269499</td>
<td>244</td>
<td>133</td>
<td>110</td>
<td>1</td>
<td>No hits</td>
<td></td>
</tr>
<tr>
<td>sgn\U269496</td>
<td>201</td>
<td>101</td>
<td>76</td>
<td>24</td>
<td>Metallothionein</td>
<td><em>C. arabica</em></td>
</tr>
<tr>
<td>sgn\U269332</td>
<td>192</td>
<td>25</td>
<td>113</td>
<td>54</td>
<td>acidic endochitinase</td>
<td><em>At</em></td>
</tr>
<tr>
<td>SGN-U269331</td>
<td>168</td>
<td>21</td>
<td>108</td>
<td>39</td>
<td>acidic endochitinase</td>
<td><em>At</em></td>
</tr>
<tr>
<td>SGN-U268736</td>
<td>143</td>
<td>26</td>
<td>100</td>
<td>17</td>
<td>lipid transfer protein</td>
<td><em>At</em></td>
</tr>
<tr>
<td>SGN-U268469</td>
<td>138</td>
<td>40</td>
<td>29</td>
<td>69</td>
<td>DNA replication</td>
<td>SV40</td>
</tr>
<tr>
<td>SGN-U269498</td>
<td>122</td>
<td>86</td>
<td>36</td>
<td>0</td>
<td>No hits</td>
<td></td>
</tr>
<tr>
<td>SGN-U268200</td>
<td>98</td>
<td>72</td>
<td>19</td>
<td>7</td>
<td>metallothionein</td>
<td><em>Citrus unshiu</em></td>
</tr>
<tr>
<td>SGN-U269495</td>
<td>80</td>
<td>40</td>
<td>30</td>
<td>10</td>
<td>Metallothionein</td>
<td><em>C. arabica</em></td>
</tr>
<tr>
<td>SGN-U268647</td>
<td>68</td>
<td>16</td>
<td>52</td>
<td>0</td>
<td>chlorophyll a/b binding protein</td>
<td><em>At</em></td>
</tr>
<tr>
<td>SGN-U269497</td>
<td>64</td>
<td>48</td>
<td>16</td>
<td>0</td>
<td>No hits</td>
<td></td>
</tr>
</tbody>
</table>
Acknowledgements

The authors wish to thank the National Federation of Coffee Growers of Colombia and the Ministry of Agriculture of Colombia for their financial support.

References


