The use of the *PHPH* tool to assembly the gene sequences that are candidate to the biotic and abiotic stress in *Musa acuminata*

Roberto C. Togawa[1], Marcelo M. Brigido[2], Candice M. R. Santos[1], Manoel T. S. Júnior[1]

[1]Laboratório de Bioinformática - Embrapa Recursos Genéticos e Biotecnologia.
Parque Estação Biológica final W5 Norte Caixa Postal: 02372 70770-900, Brasília, DF - Brasil.

[2] Laboratório de Biologia Molecular - Departamento de Biologia Celular, IB - Universidade de Brasília.
Campus Universitário, Asa Norte 70910-900, Brasília, DF - Brasil.

UnB
Embrapa
Recursos Genéticos e Biotecnologia
Ministério da Agricultura, Pecuária e Abastecimento
GOVERNO FEDERAL

Introduction

Banana (*Musa* spp.) is cultivated in numerous tropical countries throughout the world, and in many of these countries its cultivation and marketing play very important roles, both economically and socially. In 2004, Brazil produced 6,602,750 ton of bananas in an area covering 484,981 ha [1]. The majority of the banana farmers are small-scale producers with the crop grown predominantly as a supplementary source of income. Developments in molecular biology have provided tools to enable insights into changes in the transcriptome that arise when a plant is submitted to different kinds of stresses. For the assessment of gene expression, methodologies such as large-scale single pass sequencing of cDNA clones to generate expressed sequence tags (ESTs) can be utilized. ESTs provide a quantitative method to measure specific transcripts within a cDNA library and represent a powerful tool for gene discovery, gene expression, gene mapping and the generation of gene profiles [2]. Stress in plants may affect physiological and biochemical processes [3], which are transduced through a chain of signaling molecules that ultimately, affect regulatory elements of stress-inducible genes.

To help to identify genes related to biotic and abiotic stress in banana a web based tool called *PHPH* was used [4]. We describe a keyword-based search in the DATA_*Musa* database for genes known as related to biotic and abiotic stress, as well as the base calling, quality assignment and assembling of 20 candidate genes sequences using *PHPH* tool.

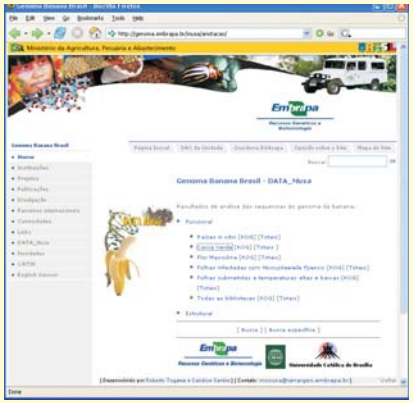

Figure 1- A screen shot of the DATA_*Musa* web page, hosted at: http://genoma.embrapa.br/musa

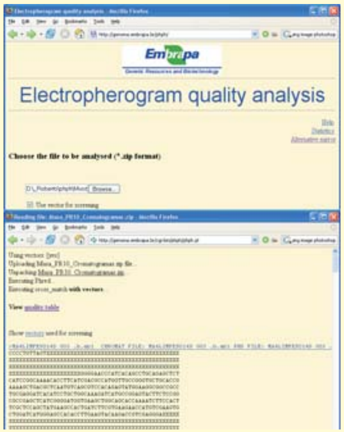
Embrapa
Electropherogram quality analysis
Choose the file to be analysed (*.zip format)
Browse...

Figure 2 -*PHPH* Initial screen and the result after submitting the sequence. From this screen the user can see the quality table, the vectors used for screening and the quality of individual sequence.

Materials and Methods

To identify genes related to the biotic and abiotic stress resistance in Musa acuminata a "virtual screening" was made in the transcriptome part of the DATA_*Musa* database (http://genoma.embrapa.br/musa/index.html/DATA_musa.html) [5]. The transcriptome part of this database consists of 5,317 *Musa acuminata* Assembled EST Sequences (MaAES). DATA_*Musa* was a result of a collaborative project sponsored by CNPq, and developed by Embrapa Genetic Resources and Biotechnology (Embrapa Cenargen), Brasilia Catholic University (UCB) and the Agricultural Research for Developing Countries (CIRAD) in France. These three institutions are also part of the Global Musa Genomics consortium (GMGC).

From the selected sequences retrieved from the DATA_*Musa* database, their correspondent electropherograms were analyzed using the *PHPH* tool. The sequences were submitted (zip format) using as an interface a web-browser. All the file manipulations and the calls for the analysis programs were developed using a PERL programming language [6] and a CGI interface. For the quality analysis a PHRED [7, 8] package was used. To mask out the vector parts that might be present within each sequence the CROSSMATCH [9] program was used. Optionally the user can run a CAP3 [10] program for the assembly, checking the sequences of interest (figure 3). A color code showing the sequence quality was used as shown in figure 4-A. A freely available chromatogram viewer (applet) [11] was used in other to show the trace (figure 4-C). This Applet can read SCF files, generated by PHRED (version 2 or 3) and ABI sample files. Also the user can save the coloured sequence in RTF format.

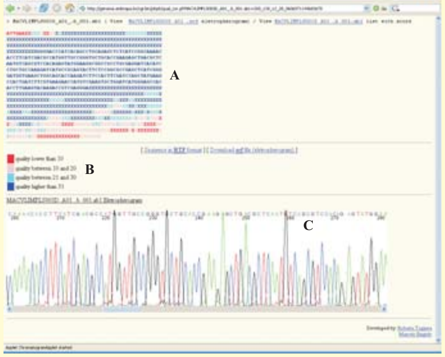

Figure 4 - The sequence quality screen. The sequence (A) are shown using the color code (B) depending on its quality. The eletropherogram is show using a JAVA applet (C).

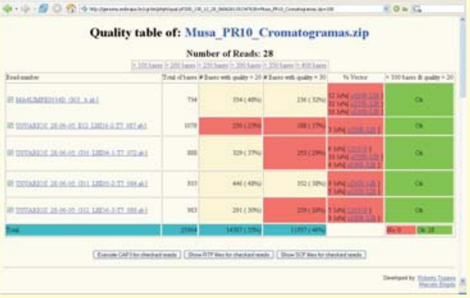
Quality table of: Musa_PR10_Cromatogramas.zip
Number of Reads: 28
Read number
Total of bases
# Bases with quality > 20
# Bases with quality > 30
% Vector
> 100 bases & quality > 20
Ok
Total
Execute CAP3 for checked reads
Show RTF files for checked reads
Show SCF files for checked reads
Developed by Roberto Togawa Marcelo Brigido

Figure 3 - Created quality table. From this screen the user can see the individual sequence quality and select the sequences to execute the CAP3 program

Results and Discussion

Using *PHPH* was possible to check the sequence quality automatically using the PHRED program. The user parameterized the PHRED quality and their sequences were grouped by CAP3 program resulting in contigs and singlets. Using the generated consensus sequence (figure 5), a BLAST [12] search was made against SwissProt database [13]. So far, using the *PHPH* tool was identified 20 genes that are related to biotic and abiotic stress in *Musa acuminata*, such as chitinase, pathogenesis-related protein (PR-10), germin-like protein, ascorbate peroxidase, glutathione peroxidase, selenium binding protein, heat shock proteins, polygalacturonase, peroxiredoxin, superoxide dismutase, salt tolerance protein, lectin, 14-3-3 protein among others.

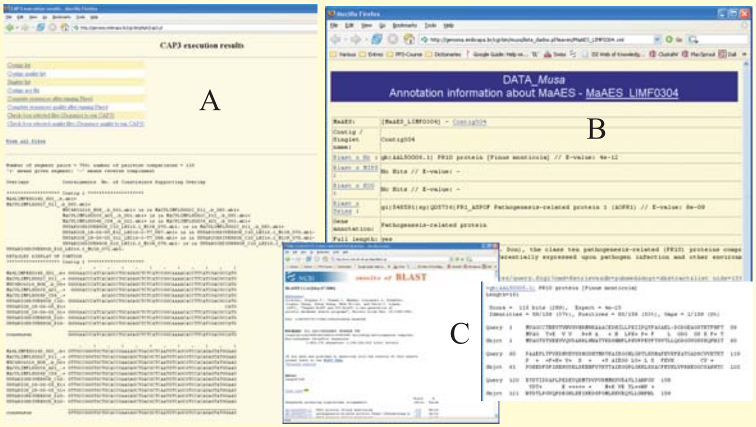

Figure 5 - (A) show the result after running the CAP3 program; (B) Search result from the DATA_*Musa* database using the keyword 'PR10'; (C) NCBI Blast result from the assembled contig. This contig was analysed by 'ORF finder' tool and the best frame was chosen to be used.

Conclusion

The data presented in this study provide a first general overview of the genes related to biotic and abiotic stresses present in DATA_*Musa* database and the possibility to use *PHPH* tool as assembler for small EST projects, with the usual pipeline from the electropherogram analysis to sequence assembly, all built-in in a single run. The *PHPH* tool can also be used for rapid quality analysis of the sequences generated by the automatic sequencer. Thanks to BIOFOCO (http://www.biofoco.org) a group of researchers engaged in the bioinformatics multidisciplinary work, *PHPH* can be accessed in the different addresses:

http://adenina.biomol.unb.br/phph (since August 2001 at Brasilia University)
http://genoma.embrapa.br/phph (at Embrapa Genetic Resources and Biotechnology)
http://bioinformatica.ucb.br/phph (at Brasilia Catholic University)

BIOFOCO is a group of researchers engaged in the bioinformatics multidisciplinary work. The main field is the development of new tools for genomics using state of the art in information technology, and gather four institutions: Embrapa - Recursos Genéticos e Biotecnologia, UCB (Universidade Católica de Brasília), UnB (Universidade de Brasília) and UFMS (Universidade Federal de Mato Grosso do Sul).

centro-oeste bioinformatics network
BIOFOCO

Bibliography

[1] FAO (http://www.fao.org)
[2] Rudd S. (2003) Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci,* **8**:321–329.
[3] Grover A., Agarwal M., Katiyar-Agarwal S., Sahi C., Agarwal S. (2000) Production of high temperature tolerant transgenic plants through manipulation of membrane lipids. *Curr Sci,* **79**:557–559.
[4] Togawa R.C. and Brigido M.M. (2003). *PHPH*: Web based tool for simple electropherogram quality analysis. 1st International Conference on Bioinformatics and Computational Biology - *IcoBiCoBi* 14th to 16th May 2003. Ribeirão Preto.
[5] DATA_*Musa* - (http://genoma.embrapa.br/musa/pt/DATA_musa.html)
[6] PERL - Practical Extraction and Report Language. (http://www.perl.com/)
[7] B. Ewing and P. Green. (1998) Base-calling of automated sequencer traces using phred. II. error probabilities. *GenomeResearch*, **8**:186-194.
[8] B. Ewing, P. Green, L. Hillier, and M. C. Wendl. (1998). Base-calling of automated sequencer traces using phred. I. accuracy assessment. *Genome Research*, **8**:175-185.
[9] P. Green. Crossmatch website documentation. (http://www.phrap.org/phredphrap/general.html)
[10] Huang X. and Madan A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, **9**:868-877.
[11] Chromatogram Applet, Release 1, 6/30/96. by Eugen Buehler (http://www.nematode.net/EST/Programs/TRACE_VIEWER/Chrom_Applet/TaggedRecord.java)
[12] 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.
[13] Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**:45-48.

This poster was presented on the 35th Annual Meeting of the Brazilian Society of Biochemistry and Molecular Biology (SBBq)
July 1 - 4, 2006 - Águas de Lindóia, São Paulo, Brazil,.

This work was supported by CNPq, Embrapa and BIOFOCO (Centro-Oeste Bioinformatics Network)