PHPH: Web based tool for simple electropherogram quality analysis

Roberto C. Togawa[1], Marcelo Macedo Brigido[2]

[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.

Embrapa Recursos Genéticos e Biotecnologia

Universidade de Brasília

Introduction

Many genome projects are undertaken worldwide. One important issue in the sequencing process is the quality analysis for the generated sequence. In many cases the user needs to know if the obtained sequence has an acceptable quality to proceed with the sequencing process or simply to check the generated sequence in an uncomplicated interface. We have developed a web-based tool for simple electropherogram quality analysis called PHPH and it is available at http://adenina.biomol.unb.br/phph since August 2001.

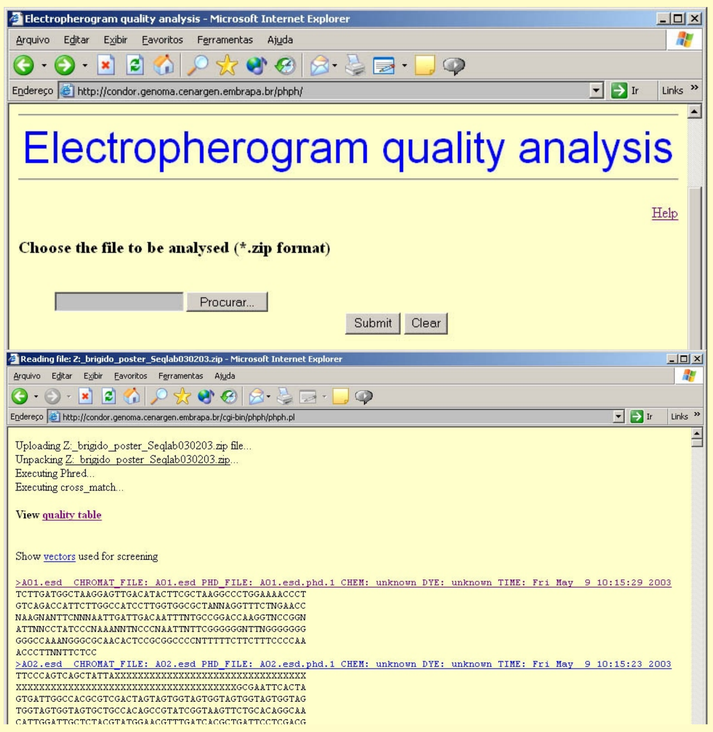

Figure 1 - Initial screen and the result after submiting the sequence. From this screen the user can see the quality table, the vectors used for screening and the quality of individual sequence.

Materials and Methods

The sequences are submitted (zip format) using a web-browser such as Mozilla, IE or Opera. All the file manipulations and the calls for the analysis programs were developed using a PERL programming language [1] and a CGI interface. For the quality analysis a PHRED [2, 3] package was used. To mask out the vector parts that might be present within each sequence a CROSSMATCH [4] program was used. Optionally the user can run a CAP3 [5] program for the assembly, checking the sequences of interest (figure2). A color code showing the sequence quality was used as shown in figure 3-A. A freely available chromatogram viewer [6] developed in JAVA programming language [7] was used in other to show the trace generated by the sequence. This Applet can read SCF files, generated by PHRED (version 2 or 3) and ABI sample files.

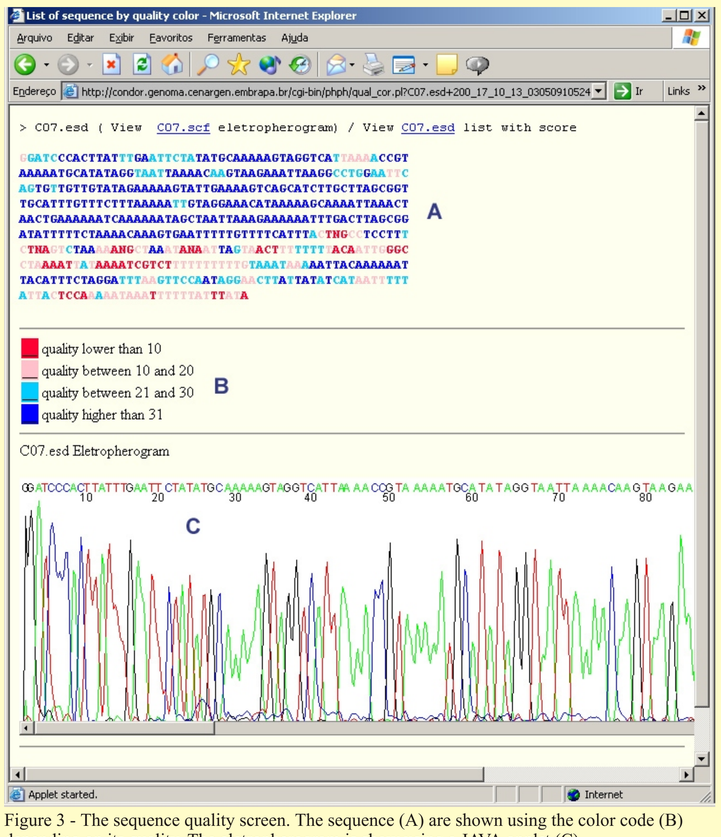

Figure 3 - The sequence quality screen. The sequence (A) are shown using the color code (B) depending on its quality. The eletropherogram is show using a JAVA applet (C).

Results and Discussion

The developed tool analyzes the sequences generated by the automatic sequencer and gives its quality using PHRED package via a web browser interface. The process of uploading the sequence trace data, call PHRED program, manipulation of the generated files and the call of CROSSMATCH, which masks out the vector parts that might be present within each sequence are automated by a script written in PERL programming language. The user can check a list of vectors used for this screening that is linked using the GenBank accession code to the EBI web site (http://www.ebi.ac.uk) (figure 5). After the processing a table containing a statistic with the acceptable sequences is shown (figure 2). Looking at this table it is possible to visualize at a glance the overall sequence status. This table contains the number of processed sequences, the number of acceptable sequences depending on the number of bases with determined quality score, the name and the percentage of the vector found in the sequence (if occur) and the total number of bases in the analysis.

Looking at the individual sequence analysis page, the user can visualize the sequence with respective quality scores using different colors for each base depending of its quality range (figure 3-A). Also is possible to visualize the eletropherogram graph using a JAVA applet (figure 3-C).
From the quality table page, the user can select the sequences for assembling using CAP3 program. The interface output presents the CAP3 assembly alignment and allows the access to the generated contigs, singlets and quality file.

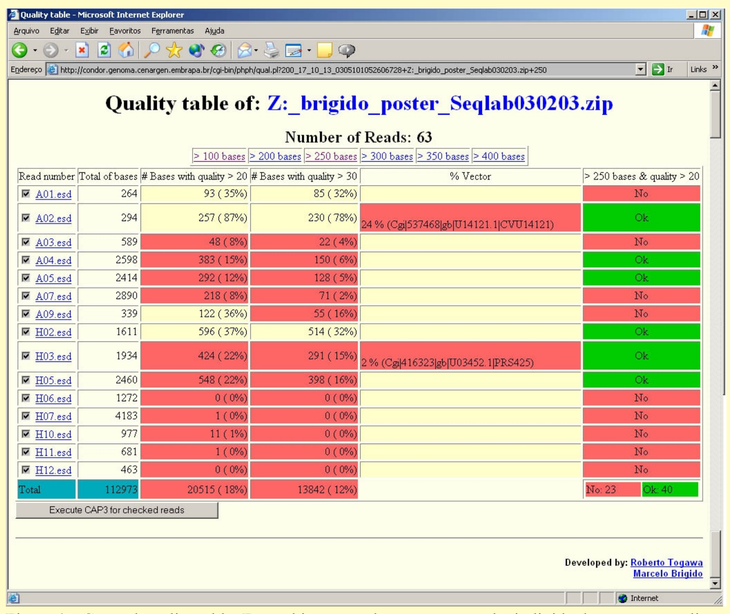

Figure 2 - Created quality table. From this screen the user can see the individual sequence quality and select the sequences to execute the CAP3 program

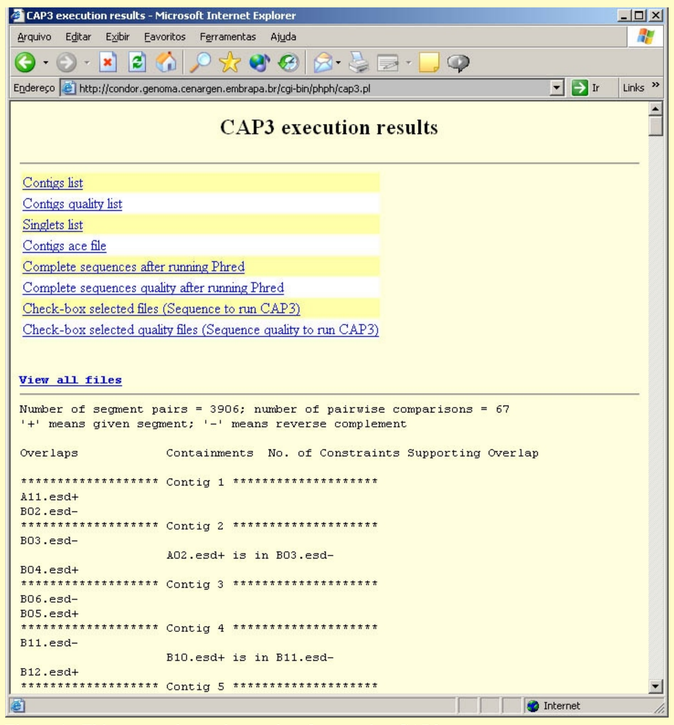

Figure 4 - This screen shows the results after running CAP3 program.

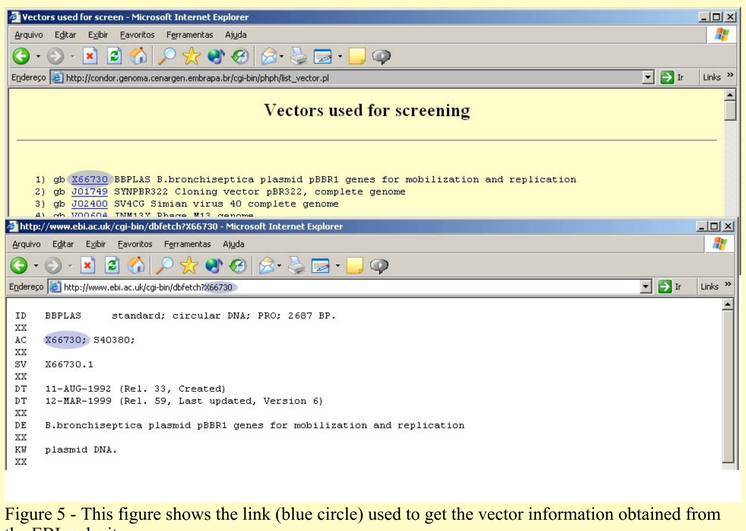

Figure 5 - This figure shows the link (blue circle) used to get the vector information obtained from the EBI web site.

Conclusion

For the scientific community working in many different genome projects, the developed tool is useful in terms of rapid quality analysis. Several times the user wants to check the sequence quality in order to adjust the experiment or simply to make a small amount of sequencing in a specific project. Using a web-browser environment it is possible to go all the way from the generated sequences until the contig assembly just "zipping" the sequences and submitting to the web server.

In terms of sequence quality visualization the user can have an idea how the sequence is at glance due to the use of different colors for different quality scores.

Thanks to BIOFOCO a group of researchers engaged in the bioinformatics multidisciplinary work, this service is mirrored at:

http://bioinformatica.ucb.br/electro.html
http://condor.genoma.cenargen.embrapa.br/phph

The BIOFOCO main field is the development of new tools for genomics using state of the art in information technology, and gather three institutions: UCB (Universidade Católica de Brasília), UnB (Universidade de Brasília) and EMBRAPA (Recursos Genéticos e Biotecnologia).

centro-oeste bioinformatics network BIOFOCO

Bibliography

[1] PERL - Practical Extraction and Report Language. *http://www.perl.com/*
[2] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. error probabilities. *GenomeResearch*, 8:186-194, 1998.
[3] B. Ewing, P. Green, L. Hillier, and M. C. Wendl.Base-calling of automated sequencer traces using phred. I. accuracy assessment. *Genome Research*, 8:175-185, 1998.
[4] P. Green. Crossmatch website documentation.
http://genome.uc.edu/genome/HelpPages/phred-phrap-polyphred/swat-crossmatch.html.
[5] X. Huang and A. Madan. CAP3: A DNA sequence assembly program. *Genome Research*, 9:868-877, 1999.
[6] Chromatogram Applet, Release 1, 6/30/96. by Eugen Buehler
(*Http://www.nematode.net/EST/Programs/TRACE_VIEWER/Chrom_Applet/TaggedRecord.java*)
[7] JAVA - The Java platform. *Http://java.sun.com/*

This poster was presented on 1st International Conference on Bioinformatics and Computational Biology - ICoBiCoBi
May 14th - 16th 2003 - Ribeirão Preto - SP - Brasil

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