# PHPH: Web based tool for simple electropherogram quality analysis

# Roberto C. Togawa<sup>1</sup>, Marcelo Macedo Brigido<sup>2</sup>

<sup>1</sup> Laboratório de Bioinformática - Embrapa Recursos Genéticos e Biotecnologia. Parque Estação Biológica final W5 Norte Caixa Postal: 02372 70770-900, Brasília, DF - Brasil. <sup>2</sup>Laboratório de Biologia Molecular - Departamento de Biologia Celular, IB - Universidade de Brasília. Campus Universitário, Asa Norte 70910-900, Brasília, DF - Brasil.



**Recursos Genéticos e Biotecnologia** 



#### Introduction

any genome projects are undertaken worldwide. One important issue in the sequencing process is the quality analysis for the generated sequence. In many cases **I V I** 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.

## **Results and Discussion**

he 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, *I* 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 (<u>http://www.ebi.ac.uk</u>) (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.

🚈 Electropherogram quality analysis - Microsoft Internet Explorer	
<u>A</u> rquivo E <u>d</u> itar E <u>x</u> ibir <u>F</u> avoritos F <u>e</u> rramentas Aj <u>u</u> da	2
🚱 - 🕑 - 🛃 😰 🏠 🔎 🤧 🧐 🥙 🍰 - 🥃 🖉 - 🛄 🥥	
Endereço 😸 http://condor.genoma.cenargen.embrapa.br/phph/	🔻 芛 Ir 🛛 Links »
Electropherogram quality	analysis
	Help
Choose the file to be analysed (*.zip format)	
Procurar Submit Clear	
Reading file: Z:_brigido_poster_Seqlab030203.zip - Microsoft Internet Explorer	
<u>A</u> rquivo E <u>d</u> itar E <u>x</u> ibir <u>F</u> avoritos F <u>e</u> rramentas Aj <u>u</u> da	
🕒 • 🕞 - 🛃 🛃 🏠 🔎 🧙 💓 🤗 🍰 - 🤤 🖓 🥥	
Endereço 🧃 http://condor.genoma.cenargen.embrapa.br/cgi-bin/phph/phph.pl	🗾 🔁 Ir 🛛 Links ≫
	<u> </u>
Uploading Z:_brigido_poster_Seqlab030203.zip IIIe	
Executing Phred	
Executing cross_match	
View <u>quality table</u>	
Show <u>vectors</u> used for screening	
>A01.esd CHROMAT FILE: A01.esd PHD FILE: A01.esd.phd.1 CHEM: unknown DYE: unknown TI)	ME: Fri May 9 10:15:29 2003
TCTTGATGGCTAAGGAGTTGACATACTTCGCTAAGGCCCTGGAAAACCCT	
GTCAGACCATTCTTGGCCATCCTTGGTGGCGCCTANNAGGTTTCTNGAACC	
A A GNANT FUNNNAAT FUAT FUAUAAT FIN FUUUGUAUUAAGU NUUUGUN ATTNNCCTATCCCNAAANNTNCCCNAATTNTTCGGGGGGGGTTNGGGGGGGG	
GGGCCAAANGGGCGCAACACTCCGCGGCCCCNTTTTTCTTCTTCCCCCAA	
ACCCTINNITICTCC	
>A02.esd CHROMAT FILE: A02.esd PHD FILE: A02.esd.phd.1 CHEM: unknown DYE: unknown TI	ME: Fri May 9 10:15:23 2003
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
GTGATTGGCCACGCGTCGACTAGTGGTAGTGGTAGTGGTAGTGGTAG	
TGGTAGTGGTAGTGCTGCCACAGCCGTATCGGTAAGTTCTGCACAGGCAA	
U.A.T. TUYUA TI TUYUTTUTTA UUFTA TUYUA A UUFTTTTU AA TUA UUFUTTUA TUUUTTUUFA UTUUFA UUF	

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.

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.

							- Chi 3 CACCULOT TESUICS - THE OSOIC Internet Explorer	
							<u>A</u> rquivo E <u>d</u> itar E <u>x</u> ibir <u>E</u> avoritos F <u>e</u> rramentas Aj <u>u</u> da	<b>R</b>
Quality table - Mic	Quality table - Microsoft Internet Explorer					l)	🌀 • 🕥 - 😰 💰 🔎 🤧 💓 🤣 🍰 - 🎍 🗩 - 🛄 🥥	
rquivo Editar Exibir Eavoritos Ferramentas Ajuda							Endereco 📳 http://condor.genoma.cenargen.embrana.br/cgi-bip/phph/can3.pl	Ir Links »
😌 • 🕑 • 📕	) • 🕑 • 💌 🖉 🏠 🔎 🛠 🏵 🔗 😓 • 🧫 🗭							
indereço 🧿 http://co	dereço 🔄 http://condor.genoma.cenargen.embrapa.br/cgi-bin/phph/qual.pl?200_17_10_13_0305101052606728+Z:_brigido_poster_Seqlab030203.zip+250 💽 💽 Ir 🛛 Links 🌺				🔽 🄁 Ir 🛛 Links 🌺		CAP3 execution results	-
Quality table of: Z:_brigido_poster_Seqlab030203.zip				_poster_Seqlab03020	- 3.zip			
			Number of <b>F</b>	Reads: 63	_		Contigs list	
		> 100 bases > 20	0 bases  > 250 bases  > 1	300 bases > 350 bases > 400 bases			Contigs quality list	
Read number To	otal of bases #Ba	ases with quality > 20 $\#$ Ba	uses with quality > 30	% Vector	> 250 bases & quality > 20		Singlets list	
A01.esd	264	93 ( 35%)	85 ( 32%)		No		Contigs ace file	
☑ <u>A02.esd</u>	294	257 ( 87%)	230 ( 78%) 24	% (Cgil537468lgblU14121.1lCVU14121)	Ok		Complete sequences after running Phred	
A03.esd	589	48 (8%)	22 (4%)		No		Complete sequences quality after running Phred	
A04.esd	2598	383 ( 15%)	150 ( 6%)		Ok		Check-box selected files (Sequence to run CAP3)	
A05.esd	2414	292 ( 12%)	128 ( 5%)		Ok		Check-box selected quality files (Sequence quality to run CAP3)	
A07.esd	2890	218 ( 8%)	71 (2%)		No			
A09.esd	339	122 ( 36%)	55 ( 16%)		No			
₩ <u>H02.esd</u>	1611	596 ( 37%)	514 ( 32%)		Ok		View all files	
M H03.esd	1934	424 ( 22%)	291 ( 15%) 2 9	% (Cgil416323 gb U03452.1 PRS425)	Ok		Number of segment pairs = 3906; number of pairwise comparisons = 67	
M H05.esd	2460	548 ( 22%)	398 ( 16%)		Ok		'+' means given segment; '-' means reverse complement	
M H06.esd	1272	0 ( 0%)	0 ( 0%)		No	1	Greelens Containments No. of Constraints Supporting Greelen	
M H07.esd	4183	1 ( 0%)	0 ( 0%)		No		overlaps concarmments No. of constraints supporting overlap	
<u>         H10.esd         </u>	977	11 ( 1%)	0 ( 0%)		No		******************** Contig 1 *******************	
₩ <u>H11.esd</u>	681	1 ( 0%)	0 ( 0%)		No		A11.esd+	
M H12.esd	463	0 ( 0%)	0 ( 0%)		No		BU2.esd-	
Total	112973	20515 ( 18%)	13842 ( 12%)		No: 23 Ok: 40		BO3.esd-	
Execute C	CAP3 for checked	reads					A02.esd+ is in B03.esd-	
							B04.esd+	
							BO6.esd-	
					Developed by: <u>Roberto Togawa</u> <u>Marcelo Brigido</u>		B05.esd+	
1					🔹 🔮 Internet		B11.esd-	
ioure 2 - 0	Treated au	ality table From	m this screen t	he user can see the individ	ual sequence quality	-	B10.esd+ is in B11.esd-	
nd select t	he sequent	ces to execute t	he CAP3 proc	fram	ual sequence quality		B12.esd+ ********************** Contig 5 ******************	-
								111

## **Materials and Methods**

he 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.

List of sequence by quality color - Microsoft Internet Explorer	_ 🗆 ×
<u>A</u> rquivo E <u>d</u> itar E <u>x</u> ibir <u>F</u> avoritos F <u>e</u> rramentas Aj <u>u</u> da	<b>.</b>
З - 🕥 - 📧 🛃 🏠 🔎 🧙 🧐 🧐 🖾 - 🌺 🔜 - 🛄 🗭	
ndereço 🙆 http://condor.genoma.cenargen.embrapa.br/cgi-bin/phph/qual_cor.pl?C07.esd+200_17_10_13_03050910524 🗾 📄	Ir Links »
> CO7.esd ( View <u>CO7.scf</u> eletropherogram) / View <u>CO7.esd</u> list with score	<u>^</u>
GATCCCACTTATTTGAATTCTATATGCAAAAAGTAGGTCATTAAAACCGT	
AAAATGCATATAGGTAATTAAAACAAGTAAGAAATTAAGGCCTGGAATTC	
АСТСТТСТТСТАТАСАААААСТАТТСААААСТСАССАТСТТССТТАССССТ ГССАТТТСТТТААААА ТТСТАССАААСТСАТСАТСАТААААСТ	
ACTGAAAAAATCAAAAAATAGCTAATTAAAGAAAAAATTTGACTTAGCGG	
ATATTTTTCTAAAACAAAGTGAATTTTTGTTTTCATTT <mark>ACTNG</mark> CCTCCTTT	
THAGT CTAAAAANGC TAAATANAAT TAGTAACTTTTTTTTTACAATTGGGC	
TACATTTCTAGGATTTAAGTTCCAATAGGAACTTATTATATCATAATTTT	
ATTACTCCAAAAATAAATTTTTTTTTTTATTA	
quality lower than 10	
quality lower than 10 quality between 10 and 20	
quality lower than 10 quality between 10 and 20 quality between 21 and 30	

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

#### Conclusion

or 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).



🚰 Vectors used for screen - Microsoft Internet Explorer		
<u>A</u> rquivo E <u>d</u> itar E <u>x</u> ibir <u>F</u> avoritos F <u>e</u> rramentas Aj <u>u</u> da		
🎯 • 🕥 - 💽 😰 🏠 🔎 🧙 😻 🎯 🖉 - 🦕 📼 • 🛄 🚳		
Endereço 🍘 http://condor.genoma.cenargen.embrapa.br/cgi-bin/phph/list_vector.pl	💌 🄁 Ir	Links
Vectors used for screening		_
<ol> <li>gb X66730 BBPLAS B.bronchiseptica plasmid pBBR1 genes for mobilization and replication</li> <li>gb J01749 SYNPBR322 Cloning vector pBR322, complete genome</li> <li>gb J02400 SV4CG Simian virus 40 complete genome</li> <li>gb W00604 INW13X Phage W13 genome</li> </ol>		
🚰 http://www.ebi.ac.uk/cgi-bin/dbfetch?X66730 - Microsoft Internet Explorer		-02
<u>A</u> rquivo E <u>d</u> itar E <u>x</u> ibir <u>F</u> avoritos F <u>e</u> rramentas Aj <u>u</u> da		
😋 • 🕑 - 📓 🛃 🔎 🧙 🔮 🥴 🖉 🍃 🔤 • 🛄 🗭		
Endereço 🙋 http://www.ebi.ac.uk/cgi-bin/dbfetch?X66730	💌 🄁 Ir	Links '
ID BBPLAS standard; circular DNA; PRO; 2687 BP. XX		



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).

A00/30; 540300;
X66730.1
11-AUG-1992 (Rel. 33, Created)
12-MAR-1999 (Rel. 59, Last updated, Version 6)
B.bronchiseptica plasmid pBBR1 genes for mobilization and replication
plasmid DNA.

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



[1] PERL - Practical Extraction and Report Language. <u>http://www.perl.com/</u> [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. <u>*Http://java.sun.com/*</u>

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)