INForum 2013

Atas do 5º Simpósio de Informática

5 e 6 de Setembro de 2013
Universidade de Évora
Portugal

Editores
João Cachopo, Instituto Superior Técnico
Beatriz Sousa Santos, Universidade de Aveiro

Edição: Escola de Ciências e Tecnologia da Universidade de Évora
Editores: João Cachopo e Beatriz Sousa Santos
Ano: 2013
ISBN: 978-989-97060-8-8

Genome Inspector: A Web Tool for Exploring Bacterial Genomes

Tiago Inocêncio[*], Joana Vital[**], Jorge Vítor[**] and Andre O. Falcão[*]

[*]Department of Informatics & LaSIGE, Faculty of Sciences; [**]Department of Biochemistry and Human Biology & iMed.UL, Faculty of Pharmacy, University of Lisbon

tinocencio@lasige.di.fc.ul.pt, joanavital@ff.ul.pt, jvitor@ff.ul.pt, afalcao@di.fc.ul.pt

Abstract. Genome Inspector (GIN) is a new interactive web-based application for exploring bacterial genomes designed to provide users with a wide choice of options to facilitate the process of designing DNA primers for isolating genes from similar bacterial species and strains. GIN employs a project-based approach in which users can select any given set of available genomes and perform a comprehensive bioinformatics analysis. Currently, it encompasses the full set of annotated bacterial genomes available on GenBank, a total of 4300 files corresponding to over 740,000 annotated genes. GIN allows new data to be added directly from GenBank as it is being generated, as well as new genome uploading for personal analysis. The application interfaces were designed with potential users in mind to assist with their most common research goals. Users can visually explore full circular genomes, search for similar regions in other species and strains, and visualize amplified genomic regions for several strains simultaneously. This interactive tool allows for dynamic graphical exploration and refinement of search and exploration criteria. Furthermore, it includes a multiple alignment algorithm (MUSCLE) to help researchers in the process of designing primers. GIN is free and publicly available at http://gin.ul.pt/GIN2/index.php.

Keywords: Bioinformatics, Web Application, Genome Analysis, Data Visualization, Restriction endonucleases, *Campylobacter jejuni*, *Helicobacter pylori*.

1 Introduction

The ability to cleave DNA at specific sequences is the fundamental technology responsible for the development of genetic engineering. Ubiquitous among prokaryotes, restriction endonucleases, commonly referred as restriction enzymes (RE), are the workhorses of the genetic engineer, and today almost every investigator working in

the molecular biology field uses these enzymes. They are defined as nucleases that recognize specific double-strand DNA sequences and cleave both strands at a defined point within or close to that sequence [1].

Beginning in 1973, a concerted effort was undertaken to assemble a collection of the known restriction endonucleases and to screen extensively for new ones. A larger number of recognition sequences would obviously increase the likelihood of obtaining the appropriate piece of DNA to clone. By 1976 almost 50 type II endonucleases had been discovered, and according to REBASE there are currently several companies selling 235 different specificities out of 628 endonucleases. Over the years many thousands of bacterial strains have been examined for the presence of these endonucleases. The classical screening process was performed in two steps: the microbiological step, consisting in bacteria isolation and biomass production, and the biochemical step, which involved enzymatic crude extract preparation and DNA hydrolysis with different substrates (plasmid, phage or genomic DNA). The screened bacteria were isolated from several sources (soil, water, etc.), or were part of private or public collections [2].

The major limitation of this classical screening methodology was the extremely limited number of bacteria that could be grown in artificial media. With the complete sequencing of a bacterial genome in 1995 and the more recent development of less expensive and faster DNA sequencing technologies, the screening strategy for new enzymes has changed [3,4]. Nowadays, screening consists in the *in silico* analysis of complete genomes followed by the popular PCR screening method and cloning of the genes of interest (restriction endonucleases or others); the DNA can easily be obtained from the environment without isolation and cultivation of microorganisms. The last step involves biochemical characterization of the enzyme, as before.

There are publically available on NCBI databases 2567 completely sequenced prokaryote genomes, *Bacteria* 2410 and *Archae* 157, and more than 18000 in progress (July 2013); 15,586 gene sequences from restriction and modification systems were deposited on REBASE only last year [5]. Our primary goal was to screen by PCR our microbial collections for type IIG restriction endonucleases [6]: *Campylobacter jejuni*, *C. coli*, and *Helicobacter pylori*. We could not find a single application to help us design primers for PCR screening of endonuclease genes, based on completely sequenced bacterial genomes. We developed a web-based application, Genome Inspector (GIN), to perform multiple DNA alignments of the sequences or genes of interest based on completely sequenced bacterial genomes publically available on GenBank [7]. The *in silico* analysis is done by the genomic annotation process, usually automatically, and in the specific case of endonucleases that annotation is refined at REBASE.

To use GIN, the user first needs to choose an annotated genome that will be the reference of the analysis, the kernel genome. A database with all the completely sequenced genomes will then be built from a list of all genomes deposited in GenBank. Using the kernel genome as reference, it is possible to select a gene or a genomic region and perform a BLAST search [8,9] to see if there are similar sequences in the other genomes; it is also possible to use an unpublished sequence or genome in the search. GIN will produce circular and linear graphics of the hits, which will assist the

user with the BLAST analysis. The final multiple alignments will be executed by MUSCLE [10,11], based on the BLAST hit table edited by the user. These multiple alignments will give the user the sequences needed to design the primers for PCR screening of the bacterial collections. Here we describe GIN in detail and present the results of primer design for PCR screening of type IIG restriction endonuclease genes in the genus *Campylobacter*.

2 Application Requirements

The system is built based on several requirements, the first being that it should be designed as a web application. In this way, a single data repository can be maintained, the interfaces can be designed in a standardized and centralized way, and the tool can be used on most platforms and operating systems in a consistent manner without requiring any software installation.

Several features were identified as functional requirements:

a) The application should enable the user to search for specific genes, genomic regions [a predetermined section of an individual genome] and sequences. The user should be able to input specific genetic sequences for search;

b) Procedures for selecting and including different genomes in each analysis should be provided to end users. These selection procedures should be accomplished both before and after the alignments;

c) The application should use BLAST for searches, and the alignment results should be provided for all the genomes selected. The blastn tool of BLAST was used;

d) After each search, a global view of the circular genomes and the positions where matches are found should be given to assist in the identification of gene translocations and transpositions among different species and strains. It is important that the information displayed is clear and easy to understand owing to the large amount of extant information;

e) A localized linear view of selected regions of the genomes under analysis should be provided, allowing users to interactively obtain detailed information for each gene in the neighborhood of the selected regions including the distinct reading frames in which each gene is located;

f) It should be possible to perform multiple alignments with the selected genomes and respective simple BLAST alignments so as to facilitate primer design procedures.

2.1 A project-centered approach

A strategy was devised to help users select the adequate set of genomes for analysis, which involved adapting the database to different possible analyses for each user, identified as projects. A user can therefore have several projects, and each project is characterized by the genomes considered for the analysis. For instance, an user may only explore the genomes of *E. coli* and *C. jejuni* in one project and be interested solely in *H. pylori* in another project; different analyses can be performed for each set.

However, each project has a predetermined validity. This is important because a separate FASTA format [12] database is built on the server for each project, consuming a significant amount of computational resources. As soon as the project expires, the FASTA database is deleted.

3 Data management and system design

3.1 Data model

The GIN database is an integrated database of complete sequenced bacterial genomes structured in a relational model (Figure 1). The Genes table is at the center of the data model as it encompasses all the relevant information pertaining to each gene in each genome, including the nucleotide sequence and its translated amino acid sequence, and the Genomes table stores information specific for each genome. The Projects table captures relevant analysis information, establishing multiple associations between genomes and projects, which mean that a genome can be involved in many projects, even from different users.

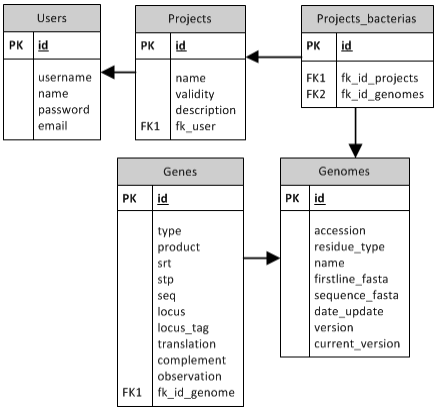


Fig. 1. Relational Model of the GIN database

3.2 Data

GenBank, one of the foremost repositories of genetic sequence data, is an annotated collection of all publicly available DNA sequences and is updated every two months. An exhaustive search of all bacterial genomes available was performed and downloaded on GIN. A Python script was programmed in order to retrieve the required information from each individual GenBank format file and store it in the central database defined for GIN use. This script is used by the back-office component of the application and can be accessed by administrators to update the database.

The system currently encompasses about 2567 different genomes and over 740,000 genes, thus requiring about 4 GB of available disk space.

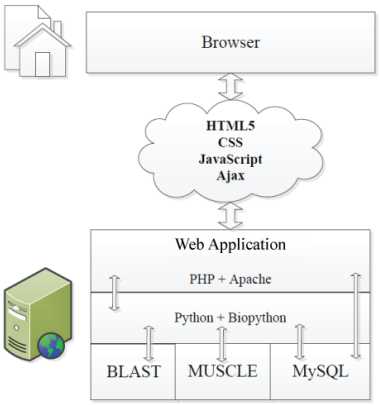
3.3 System architecture

Like a typical web application, GIN was designed as a three-tiered application. Figure 2 represents the proposed system architecture and makes explicit the interconnection between different modules.

The server module can be divided into three layers. The first layer is the web server and serves as the core of the front end. Web pages are created dynamically in PHP according to the user's requests. This layer communicates with the main database and with a second layer that supports most bioinformatics processing. This layer was fully developed in Python since BioPython [13] is an obvious choice for establishing an interface with the search and alignment tools (BLAST and MUSCLE).

Fig. 2. Two-component system architecture. The client side can access the application from the browser, and the server component has three distinct inter-connected layers.

3.4 MUSCLE



The Multiple Sequence Comparison by Log Expectation (MUSCLE) enables the gen-eration of multiple alignments of protein or nucleic acid sequences and provides a number of options, including the optimization of speed and accuracy of the nucleic acid alignment. MUSCLE is integrated into GIN and is accessed via Py-thon/BioPython, which is able to process the input data and retrieve the multiple alignment results. The output can be provided in several formats including HTML, ClustalW standard multiple alignments output, among others [9, 10].

4 Results

4.1 System development

According to the requirements described above, the system was developed using an agile software development methodology. An initial framework was developed and a limited amount of genomes were collected. The aim was to have a working prototype as soon as possible so that users could promptly start interacting with the system, always allowing greater flexibility in the implementation of the proposed solutions. Meetings were held periodically to define strategies to overcome the difficulties faced.

The system entered a public beta phase in September 2012, with the core functionalities implemented over a limited number of genomes. This version has been contin-

uously upgraded since then, and new functionalities have been introduced to provide a better user experience and facilitate the analysis process. The core features implemented were:

a) Separate user logins that enable a project-centered approach; registration is not mandatory, but access will be limited only to the most important bacterial families. A registered user is able to create individual projects with a selection of genomes for analysis. Projects exist physically in secondary storage within the server. This is a necessity since BLAST requires separate databases for analysis. Therefore, in order not to clutter the hard-drive with too many projects, each has a finite duration that can be extended at will by the user.

b) Several performance optimizations were implemented, which resulted in faster database access (even with modest server hardware), faster generation of web pages and improved mechanisms for server module computation. This enabled inclusion of GenBank's full set of bacterial genomes within the current system.

c) Improved graphics for both circular and linear plots, which were greatly extended in terms of the amount of information provided and navigational capabilities.

4.2 Core functionalities

All the functional requirements for the application were met. Among the most noteworthy are: the integration of a set of bioinformatics tools and data within a single web application; browsing of search results in circular graphics and inspection of gene neighborhoods with linear graphics; and visualization of multiple alignments for primer development.

4.3 Tool and data integration

One of the strengths of the present system is that it is a fully integrated environment for performing specialized analytical work on bacterial genomes without requiring any external tool. GenBank data have been transferred to the existing MySQL repository, and tools for genome selection, search and alignment, global and local analysis, as well as tools for robust multiple alignment, are all available within GIN.

Back-office tools were further developed to allow for easy database customization and update as more data become available.

Full genome exploration with circular graphics

For purposes of exploration, analysis and comparison of full genomes, circular graphics are very powerful representations that allow visualization of the relative distances of each gene or genomic sequence. These plots help the user explore the full bacterial genomes after a BLAST search and alignment process. Users are able to include several genetic markers (e.g., RNA coding regions), if available in the database. The output enables the identification of the direction of the selected gene read-

ing frame and its positional location within the full circular genome. When searching for a specific gene, it is possible to interactively obtain all the relevant characteristics for all the genes matched against the genomes analyzed (Figure 3).

Detailed view of local neighborhood with linear graphics

A typical second step for genome exploration is the identification of the search region in several species or strains (Figure 4).

Multiple alignment visualization

Multiple alignments are produced by running MUSCLE with the selected regions defined by BLAST (Figures 5, 6 and 7). It is possible to refine the regions selected by expanding or reducing the selected regions manually and individually for each genome being analyzed.

4.4 A test case

Our primary goal was to PCR screen two collections of *C. jejuni* and *H. pylori* for type IIG RE. Here we present the primers design methodology using GIN and also the PCR screening results for part of our *C. jejuni* collection.

The kernel genome, *C. jejuni* RM1221, was chosen from REBASE because it has the largest number of type IIG RE, with four putative genes (Table 1).

Table 1. Type IIG restriction endonucleases in *Campylobacter jejuni* RM1221

NC_003912 Gene no.	REBASE notation for putative type IIG RE	Start	Stop	Molecular weight kd (amino acid residues)	Coding strand
CJE0031	Cje1221ORF31P	44699	48472	148311 (1257 aa)	Direct
CJE0309	Cje1221ORF309P	272329	275679	130467 (1116 aa)	Direct
CJE0789	Cje1221ORF789P	727118	723366	147543 (1250 aa)	Reverse
CJE1195	Cje1221ORF1195P	1115282	1111251	156836 (1343 aa)	Reverse

A project was initiated with all *Campylobacter* complete genome sequences: *C. concisus* 13826, *C. curvus* 525.92, *C. fetus* 82-40, *C. hominis* ATCC BAA-381, *C. jejuni* (RM1221, 269.97, 81-176, 81116, IA3902, ICDCCJ07001, M1, NCTC 11168, NCTC 11168-BN148, PT14 and S3) and *C. lari* RM2100, totaling sixteen genomes.

Using *C. jejuni* RM1221 as the kernel genome and gene CJE0031 as the seed, the BLAST search identified hits on twelve genomes: eleven in *C. jejuni* and one in *C. lari* RM2100 (Table 2). The circular genomic maps show that all *C. jejuni* genomes have a hit in a similar genomic position, raising the hypothesis that there is at least one specific type IIG locus in *C. jejuni* genomes. *C. lari* also has a CJE0031 identical gene but in a different genomic position (Figure 3).

Table 2. BLAST hit table result on sixteen *Campylobacter* complete sequenced genomes: seed gene CJE0031, kernel genome *C. jejuni* RM1221

Genome Number	GenBank accession number, Species, Strain	Score	Max. ident.	Align length	Start	End
5	NC_003912 C. jejuni RM1221	3774	3774	3774	44699	48472
6	NC_009707 C. jejuni 269.97	1626	1938	2084	49738	51800
7	NC_008787 C. jejuni 81-176	2293	2607	2754	55010	57742
8	NC_009839 C. jejuni 81116	1676	1939	2061	53386	55426
9	NC_017279 C. jejuni IA3902	2356	2422	2455	46231	48685
10	NC_014802 C. jejuni ICDCCJ07001	1578	1718	1787	53532	55316
11	NC_017280 C. jejuni M1	1646	1930	2062	53409	55449
12	NC_002163 C. jejuni NCTC 11168	2907	3072	3152	46424	49571
13	NC_018521 C. jejuni NCTC 11168-BN148	2907	3072	3152	46424	49571
14	NC_018709 C. jejuni PT14	2323	2411	2455	44658	47112
15	NC_017281 C. jejuni S3	2979	3096	3152	44697	47843
16	NC_012039 C. lari RM2100	613	1949	2558	1403450	1401009

The BLAST hit table is ordered alphabetically and displays items chosen to help the user decide if the hit found is relevant or not: bit score, maximum identity, length of the hit alignment. The hit chromosomal positions, "Start" and "End", are also shown. In the present case, the bit score is high in all *C. jejuni* genomes indicating that these genes are identical (the higher the score, the better the alignment). *C. lari* has the smallest bit score, but the alignment length is greater than that of other genomes with higher scores, so we decided to maintain it in the first round of multiple alignments to verify if all the genes had similar initial and final sequences.

The first round of multiple alignments was done with CJE0031. Considering the first thousand nucleotides after the start codon, it was clear that there were at least two groups of similar genes, and the hit in *C.lari* was different from the *C. jejuni* genes (Figure 5). The multiple alignments of the carboxyl-terminus, also spanning 1000 nucleotides before the stop codon, confirm the results of the amino-terminus multiple alignment (data not shown). These results imply that more than one pair of primers will be needed when doing PCR screening for type IIG proteins in this locus. A second round of multiple alignments was done excluding *C. lari*, using the genes upstream and downstream of CJE0031. However, neither CJE0030 nor CJE0032 are conserved in the 11 genomes. CJE0029 and CJE0033 are conserved, but the resulting PCR fragment is too large, more than 14000 base pairs, when 3774 is the nucleotide size of target gene CJE0031.

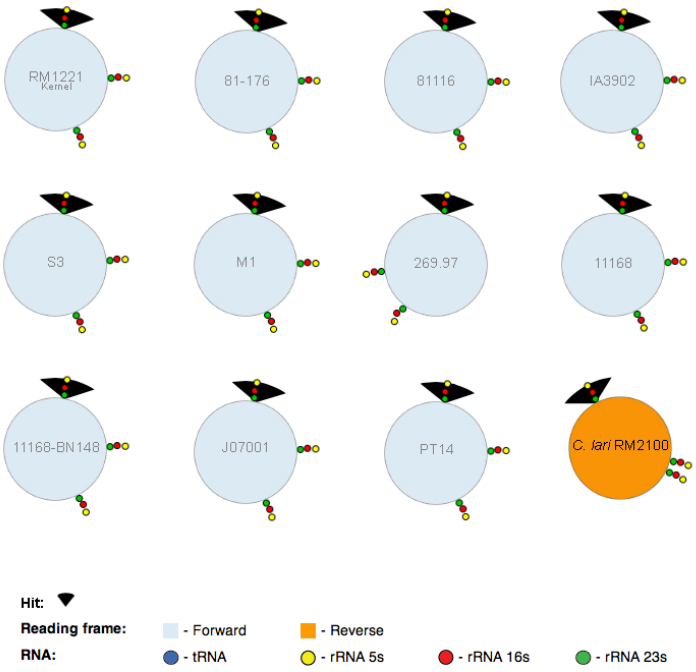


Fig. 3. Genomic maps showing the BLAST hit positions (black pies). *Campylobacter jejuni* RM1221 was the kernel genome, and the seed gene was CJE0031 (44699-48472). All *C. jejuni* genomes have an identical type IIG restriction endonuclease gene in an equivalent position. In strain 269.97, there is clearly a major genomic rearrangement: two of the three rRNA loci are on a different quadrant. *C. lari* has a gene similar to CJE0031, but in a different genomic position and coded on the reverse frame.

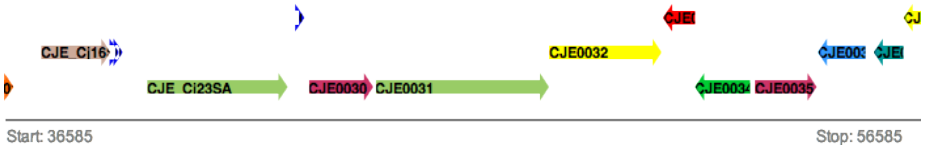


Fig. 4. Linear graphic showing the genes up-stream and down-stream CJE0031, 20000 nucleotide window, from *Campylobacter jejuni* RM1221.

There is an RNA locus between CJE0029 and CJE0030, and a sequence-based search approach was used for RNA23S and RNA5S. A small sequence, conserved among all genomes, was found and it is ideal for designing the forward primer (Figures 6 and 7). The resulting PCR fragment is still large, with 8000 base pairs, but we decided to use it and had 40% of PCR products with expected size (Table 4).

Fig. 5. MUSCLE HTML output of multiple alignments of BLAST results from table 1, showing the first 120 nucleotides. Genome 1 is the kernel, strain RM1221, and genome number 12 is *Campylobacter lari*. It is possible to use this result to design a primer including the start codon that will be efficient in 54% of the *C. jejuni* strains because the first 98 nucleotides are common to 6 of the 11 *C. jejuni* strains.

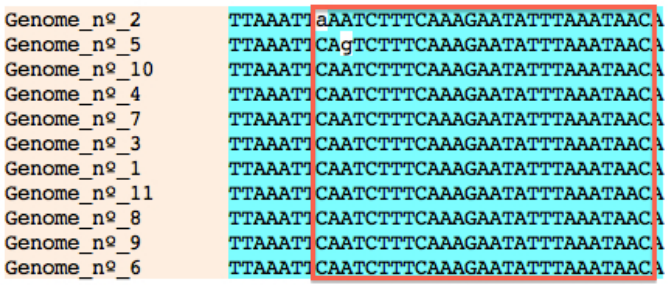


Fig. 6. MUSCLE HTML output of the multiple alignments based on region 42772-43017 of *C. jejuni* RM1221, genome 1, showing the last 35 nucleotides. The box shows the conserved region used to design the forward primer for Locus 1, CampL1f in table 3. The last "A" of genome 1 is nucleotide number 43017.

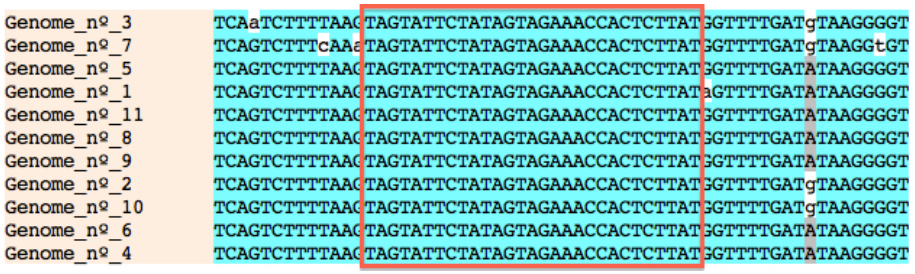


Fig. 7. MUSCLE HTML output of the multiple alignments based on gene CJE0033 of *C. jejuni* RM1221, showing the first 60 nucleotides. Genome 1 is the kernel strain RM1221. The box shows the conserved region used to design the reverse primer for Locus 1, CampL1r in table 3. The first "T" of genome 1 is nucleotide number 50958.

The gene search strategy was used on the other three putative type IIG RE genes, CJE0309, CJE0789 and CJE1195, and the results were similar to the ones already

described: no common sequences were found within the genes including the start and stop codons. Identical DNA sequences were found upstream and downstream of the seed genes, allowing the design of a pair of primers for each gene to screen the *C. jejuni* bacterial collection with a minimal set of primers (Table 3).

Table 3. Primers used in PCR screening of *Campylobacter jejuni* and *C. coli* type IIG RE

Primer	Sequence 5'-3'*	Tm – Phusion™ polymerase	Tm PCR	Amplicon size (nt)
CampL1f**	MARTCTTTCAAAGAATATTTAAATAAC	61	58	8000
CampL1r	ATAAGAGTGGTTTCTACTATAGAATACTA	58		
CampL2f***	ACTAAAGARTGTTGYGGTTGTGC	64	60	4485
CampL2r	CAGAAAGTATTAAAGAACAAACTGCA	61		
CampL3f	AAGGCTAAATTCTTTAAATTTTTTCAT	60	60	3964
CampL3r	CTACAGATGAAGAATTAGAAATCGC	61		
CampL4f	AAACTCATAAAAGTGTCAAGGGC	62	62	4188
CampL4r	TCCGCTTAATATCATAATGTTTTTCAT	63		

* Based on *C. jejuni* RM1221 genome; ** M: A or C; *** Y: T or G

Using these 4 pairs of primers, 169 *C. jejuni* strains were screened by PCR using Phusion™ polymerase (Table 4; details will be published elsewhere). Four type IIG endonucleases specific loci were found in *C. jejuni*, and the PCR screening of 168 strains resulted in 340 PCR products with lengths compatible with those of type IIG restriction endonucleases. Once the extremities are sequenced, it is possible to design new primers to clone the PCR products into *E. coli* expression vectors. This strategy enables a safe and efficient screening of pathogenic bacterial strains for genes of potential biotechnological interest, and GIN plays a key role in this process.

Table 4. Results of *Campylobacter* PCR screening for type IIG RE

169 Campylobacter strains screened	**CL1**	**%**	**CL2**	**%**	**CL3**	**%**	**CL4**	**%**
One PCR band	68	40.2	145	85.80	114	67.45	139	82.25
PCR product with expected size	68	40.2	36	21.30	114	67.45	122	72.19
No amplification	89	52.6	23	13.61	41	24.26	23	13.61
Unspecific amplification	21	12.43	1	0.59	14	8.28	7	4.14

5 Conclusion

This article describes the analytical system GIN and how it was structured and implemented to solve the issues initially identified. With this system it is possible to make simple alignments through BLAST, and the results produced are presented in text or graphic format (linear and circular graphs).

Linear graphics enable viewing of matched genes within each selected genome as well as their neighborhoods, with the user having direct access to the full annotated information in one single and comprehensive interface. Additionally, with circular

graphics it is possible to view the entire genome and to identify where the gene is located. This system also enables multiple alignments using MUSCLE and shows these results in text format only.

Finally, we can conclude by the user tests conducted that the system is easy to interact with. The tests produced comprehensive, diverse information that allowed users to choose which analysis they wanted.

Acknowledgements

We would like to thank Patrícia Fonseca for editorial support and New England Biolabs, Inc. (USA) for funding Jorge Vítor's research. We would also like to thank Fundação para a Ciência e Tecnologia for supporting LaSIGE and iMed.UL through the pluriannual funding program, which awarded Tiago Inocêncio's research grant.

References

1. Roberts, R. J. & Halford, S. E. (1993). "Type II restriction enzymes", in Nucleases, ed. S.M. Linn, R.S. Lloyd & R.J. Roberts (Cold Spring Harbor: Cold Spring Harbor Laboratory Press), 35-88.
2. Shildkraut, I. (1984). Screening for and characterizing restriction endonucleases. In Genetic Engineering, Principles and Methods. Vol 6. Setlow, J. and Hollaender, A. (eds). (New York, Plenum Press), 117-140.
3. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, *269*(5223), 496–512.
4. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138.
5. Roberts, R.J., Vincze, T., Posfai, J., and Macelis, D. (2010). REBASE—a database for DNA restriction and modification enzymes, genes and genomes. *Nucl. Acids Res*. 38: D234-D236.
6. Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickle, T. A., Bitinaite, J., et al. (2003). A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic acids research*, *31*(7), 1805–1812.
7. Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler (2005). GenBank. *Nucleic Acids Res*. 2005 January 1; 33(Database issue): D34–D38.
8. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J Mol Biol*. 215(3): 403-10.
9. McGinnis, S. & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*. 32 (Web Server issue): W20–W25.
10. Robert, E. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004, 5:113.
11. Robert, E. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res*. 32 (5): 1792-1797.
12. Lipman, DJ; Pearson, WR (1985). "Rapid and sensitive protein similarity searches". *Science* 227 (4693): 1435–41.
13. Chapman BA and Chang JT. (2000) Biopython: Python tools for computational biology. *ACM SIGBIO Newsletter* 20, 15-19.