Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

A new set of bioinformatics tools for genome projects

Luiz G.P. Almeida[1], Roger Paixão[1], Rangel C. Souza[1], Gisele C. da Costa[1], Darcy F. de Almeida[2] and Ana T.R. de Vasconcelos[1]

[1]Laboratório Nacional de Computação Científica (LNCC)/
Ministério de Ciência e Tecnologia, Petrópolis, RJ, Brasil
[2]Instituto de Biofísica Carlos Chagas Filho,
Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil
Corresponding author: A.T.R. de Vasconcelos
E-mail: atrv@lncc.br

Genet. Mol. Res. 3 (1): 26-52 (2004)
Received October 13, 2003
Accepted January 12, 2004
Published March 31, 2004

ABSTRACT. A new tool called System for Automated Bacterial Integrated Annotation - SABIA (SABIÁ being a very well-known bird in Brazil) was developed for the assembly and annotation of bacterial genomes. This system performs automatic tasks of assembly analysis, ORFs identification/analysis, and extragenic region analyses. Genome assembly and contig automatic annotation data are also available in the same working environment. The system integrates several public domains and newly developed software programs capable of dealing with several types of databases, and it is portable to other operational systems. These programs interact with most of the well-known biological database/softwares, such as Glimmer, Genemark, the BLAST family programs, InterPro, COG, Kegg, PSORT, GO, tRNAScan and RBSFinder, and can also be used to identify metabolic pathways.

Key words: Assembly, Automatic annotation, Software

INTRODUCTION

The SABIA (System for Automated Bacterial Integrated Annotation) software was developed to fulfill the computer needs of the Brazilian Genome Project for the management, assembly and annotation of the *Chromobacterium violaceum* genome. Its purpose was to integrate and automate the use of programs, and to facilitate access to public domain database, as well as those developed locally by the Bioinformatics Laboratory (LABINFO/ LNCC) team.

One of the particular features of this project was that the general coordination, the DNA laboratories, and the bioinformatics and sequencing laboratories were geographically distant from one another, being distributed throughout much of Brazil (for information on the origin and significance of the network, see Simpson, 2001). In order to deal with the drawbacks inherent to projects of this type, a series of follow-up and management reports were made available daily on the project's home page (www.brgene.lncc.br/cviolaceum). Tables containing information, such as the quality of the sequences submitted by each group, libraries and plates, among others, allowed decisions to be made and strategies to be established during the project's development.

The initial strategy to assemble the genome was large-scale sequencing of shotgun reads (Fleischmann et al., 1995) and cosmid ends. The contigs that were generated were ordered through a scaffold program (Setubal and Werneck, 2001). Following this phase the gap closure, or the finishing of the genome sequence generated by the shotgun sequences, was initiated.

Two basic gap types were identified: i) sequence or sequencing gaps, in which there is a DNA template (cosmid or shotgun read) with extremities in two adjacent contigs, and ii) physical gaps, for which there is no binding DNA template. The existence of the gaps could be explained by statistical or by functional and methodological reasons, as for instance, unstable regions or non-cloning toxic sequences, or a cloning bias associated with either the DNA fragmentation method or the cloning system used. Gaps are frequently associated with repetitive regions, such as the ribosomal operons, transposases and large genetic families. The sequencing gaps are easily closed after a careful selection of the shotgun clones for re-sequencing and subcloning. As for the physical gaps and the repetitive regions, specific closure methodologies were developed that are described elsewhere (Carraro et al., 2003).

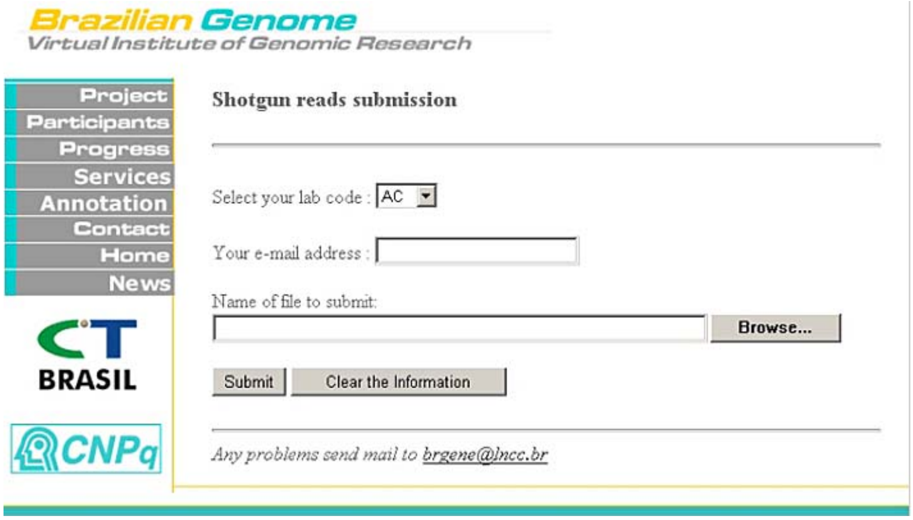
As the gaps were closed, the number of contigs decreased, the assembly was frozen, and annotation could be initiated. The SABIA method relies on the metabolic pathways of the organism; this is an approach distinct from those generally used by other genome projects, for it allows the premature identification of regions of particular interest. This system uses a group of well-known software and database, such as Glimmer (Delcher et al., 1999), GeneMark (Borodovsky and McIninch, 1993), tRNAscan (Lowe and Eddy, 1997), Blast (Altschul et al., 1990), InterPro (Mulder et al., 2003), KEGG (Kanehisa, 1996), and COG (Tatusov et al., 1997).

Software description

SABIA is made of two defined modules: assembly and annotation. Each module consists of a group of softwares written in the PERL programming language (version 5.6), executed in a command line fashion, or under the http Apache manager (version 1.3), and a relational database, implemented by means of MySQL software (version 1.3). The SABIA version used in this project was installed under the UNIX operating system. The annotation module requires the database nt, nr (www.ncbi.nlm.nih.gov), COG, KEGG, InterPro and GO (http://geneontology.org/) for proper functioning. The two modules are interconnected, thus allowing genomic sequences generated during the assembly phase to be used during annotation; likewise the information generated by the annotation can assist in the process of assembly analysis. The automatic assembly and annotation processes can be configured to be executed periodically.

Assembly

The large volume of data and tasks involved in the analysis and assembly of the *C. violaceum* genome motivated the construction of the SABIA assembly module. This module coordinates the automation, integration and organization of the results generated by the phred/phrap/consed programs (www.phrap.org). The package accomplishes tasks ranging from chromatogram analysis to assembly visualization, creating files that contain the assembly results to be used by SABIA. SABIA provides follow-up reports and supporting tools for the administration of the project, sequencing analyses and assembly of the genomes. The sequencing of the *C. violaceum* genome was divided into three phases: i) sequencing of the shotgun reads: approximately 80,000 reads with phred scores >20 were generated from both ends of plasmid clones ranging from 2.0 to 4.0 kb, providing a 13-fold genome coverage; ii) sequencing of the cosmid ends: both ends of 3,350 cosmid clones with an average insert size of 40 kb were also sequenced, thus providing a validation check of the final assembly, and iii) the finishing phase, where the quality of the assembled sequences was analyzed.



Brazilian Genome
Virtual Institute of Genomic Research

Shotgun reads submission

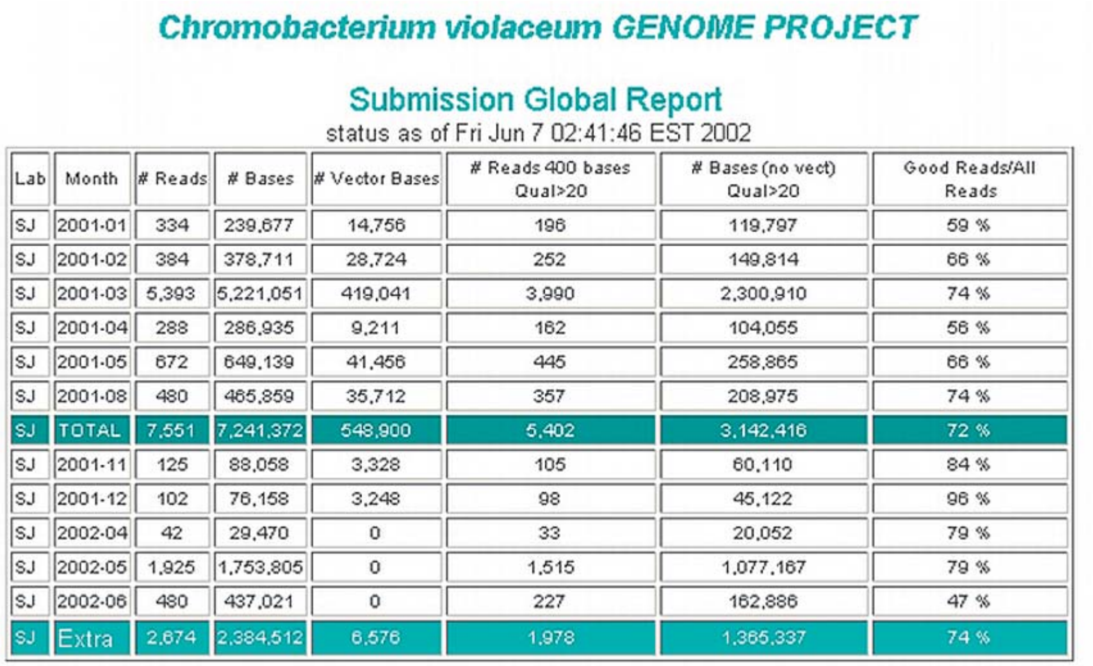
Select your lab code : AC

Your e-mail address :

Name of file to submit:

Any problems send mail to brgene@lncc.br

Figure 1. Sequence submission page.



Chromobacterium violaceum GENOME PROJECT

Submission Global Report

status as of Fri Jun 7 02:41:46 EST 2002

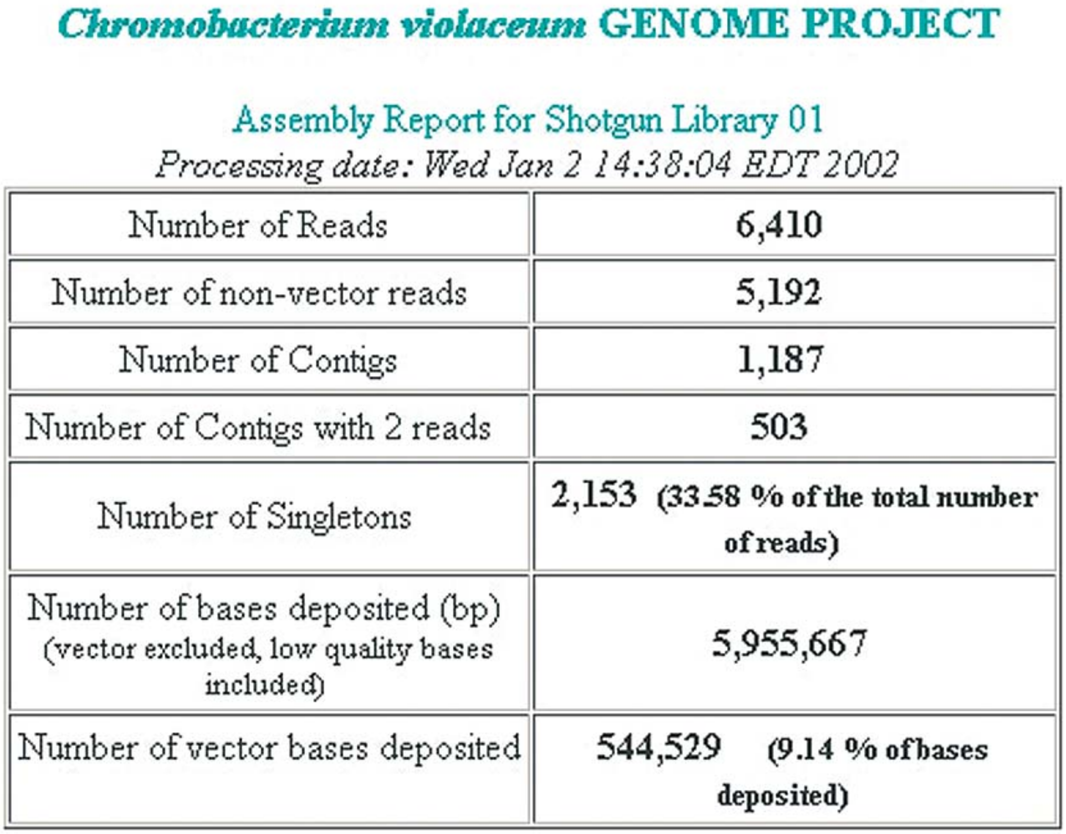
Lab	Month	# Reads	# Bases	# Vector Bases	# Reads 400 bases Qual>20	# Bases (no vect) Qual>20	Good Reads/All Reads
SJ	2001-01	334	239,677	14,756	196	119,797	59 %
SJ	2001-02	384	378,711	28,724	252	149,814	66 %
SJ	2001-03	5,393	5,221,051	419,041	3,990	2,300,910	74 %
SJ	2001-04	288	286,935	9,211	162	104,055	56 %
SJ	2001-05	672	649,139	41,456	445	258,865	66 %
SJ	2001-08	480	465,859	35,712	357	208,975	74 %
SJ	TOTAL	7,551	7,241,372	548,900	5,402	3,142,416	72 %
SJ	2001-11	125	88,058	3,328	105	60,110	84 %
SJ	2001-12	102	76,158	3,248	98	45,122	96 %
SJ	2002-04	42	29,470	0	33	20,052	79 %
SJ	2002-05	1,925	1,753,805	0	1,515	1,077,167	79 %
SJ	2002-06	480	437,021	0	227	162,886	47 %
SJ	Extra	2,674	2,384,512	6,576	1,978	1,365,337	74 %

Figure 2. Report of read production from a lab, showing the number of reads and bases, and the read qualities.

Submission of shotgun reads

SABIA manages the process of read submission and analyses by providing reports of read production (both quality and quantity), which help in the identification of the shotgun phase finalization. The submission process and the "nomination" of shotgun reads was standardized and established by a protocol that takes into account the name of the organism, the laboratory, the library, the plate and the orientation (the sequenced end in the forward direction is identified by the letter "b" and the sequenced end in the reverse direction by the letter "g"). For submission, the user informs a contact e-mail, the plate identification, the sequence orientation (b or g), and attaches the zipped file with the reads (Figure 1). After unzipping the file, the reads are nominated according to the previously determined pattern, and the information provided during submission. Whenever there are reads with the same name, or the name does not agree with the pattern, the read is rejected and the user notified. The phred (base calling) program is then executed; it checks for vector sequences that will be replaced by "X", in order to avoid their usage during assembly. SABIA then analyzes the file, calculating the size of each sequence, the number of bases with phred quality ≥20 and ≥30, the number of bases corresponding to vectors (total, ≥20 and ≥30). The result of this analysis is sent by e-mail to the project coordinators and to the laboratory submitting the file. These data are important to evaluate the quality of each file, the production of each laboratory and of the sequencing net, as well as the quality of the library that was used. The accounting data are stored at the assembly database, and updated reports are made available on the web.

Only reads containing 400 bases with phred quality ≥20 were considered for the sequencing of the *C. violaceum* genome. Two types of reports were created (Figure 2), one with



Chromobacterium violaceum GENOME PROJECT

Assembly Report for Shotgun Library 01

Processing date: Wed Jan 2 14:38:04 EDT 2002

Number of Reads	**6,410**
Number of non-vector reads	**5,192**
Number of Contigs	**1,187**
Number of Contigs with 2 reads	**503**
Number of Singletons	**2,153 (33.58 % of the total number of reads)**
Number of bases deposited (bp) (vector excluded, low quality bases included)	**5,955,667**
Number of vector bases deposited	**544,529 (9.14 % of bases deposited)**

Figure 3. Report of a library, showing the number of sequenced reads, the contigs and singlets formed.

the total and monthly production of each laboratory, the other with the total production and the production of each laboratory or library. A follow-up of the total production could also be made through the monthly graphic reports.

An important additional report is the “assembly report for shotgun library”, which provides a synthesis of the quality of the genome libraries built for the project. It includes some relevant information, such as the percentage of vector sequences and the average size of the clones, updated daily (Figure 3).

Assembly execution

To assemble the genome, SABIA automatically runs the phredphrap program and stores the results in the database for subsequent analysis. The execution of some tasks, such as creation of the repetition file, formatting the reads and contig sequence banks in the blast format, generation of the scaffold map, freezing of the assembly, and analysis of the repetitions are automatic, and may be executed when ordered by the administrator.

Assembly follow-up

To allow the monitoring of the genome assembly evolution, the phrap.out file is analyzed

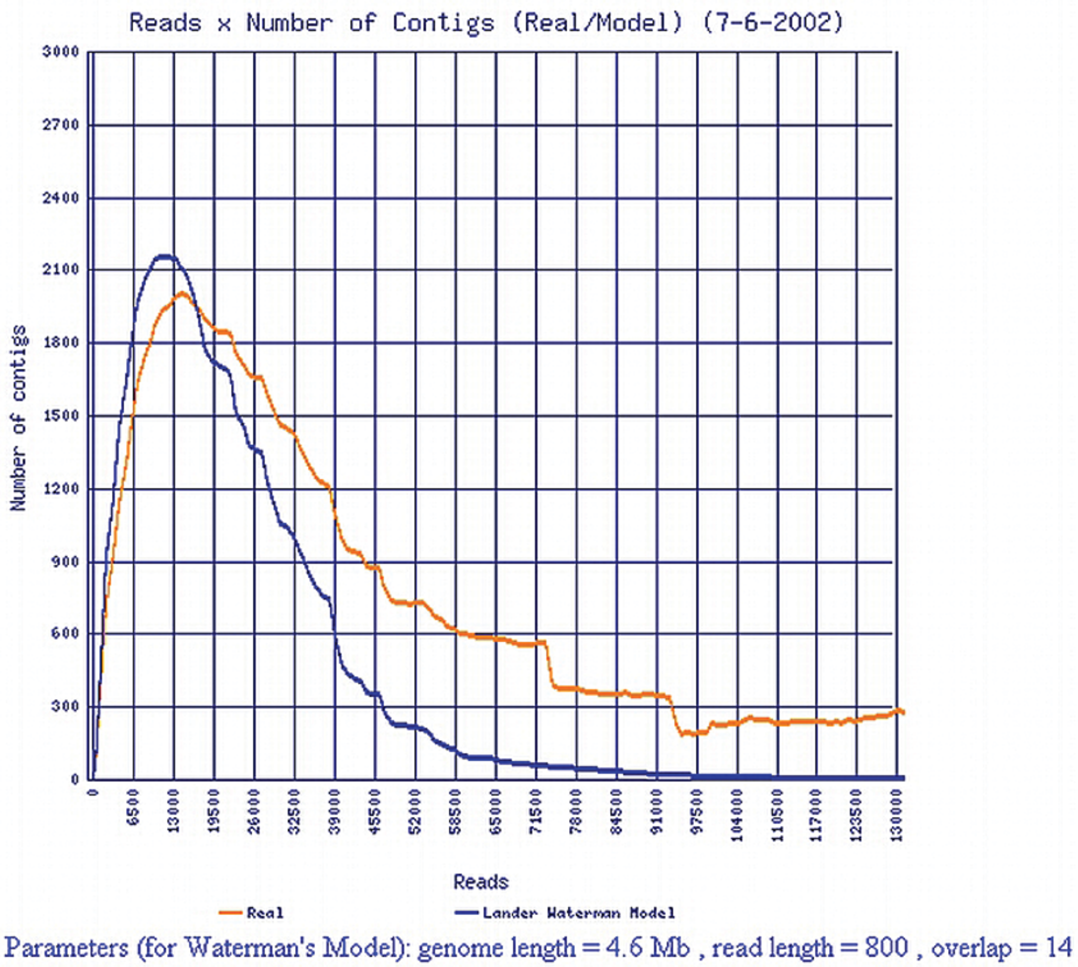
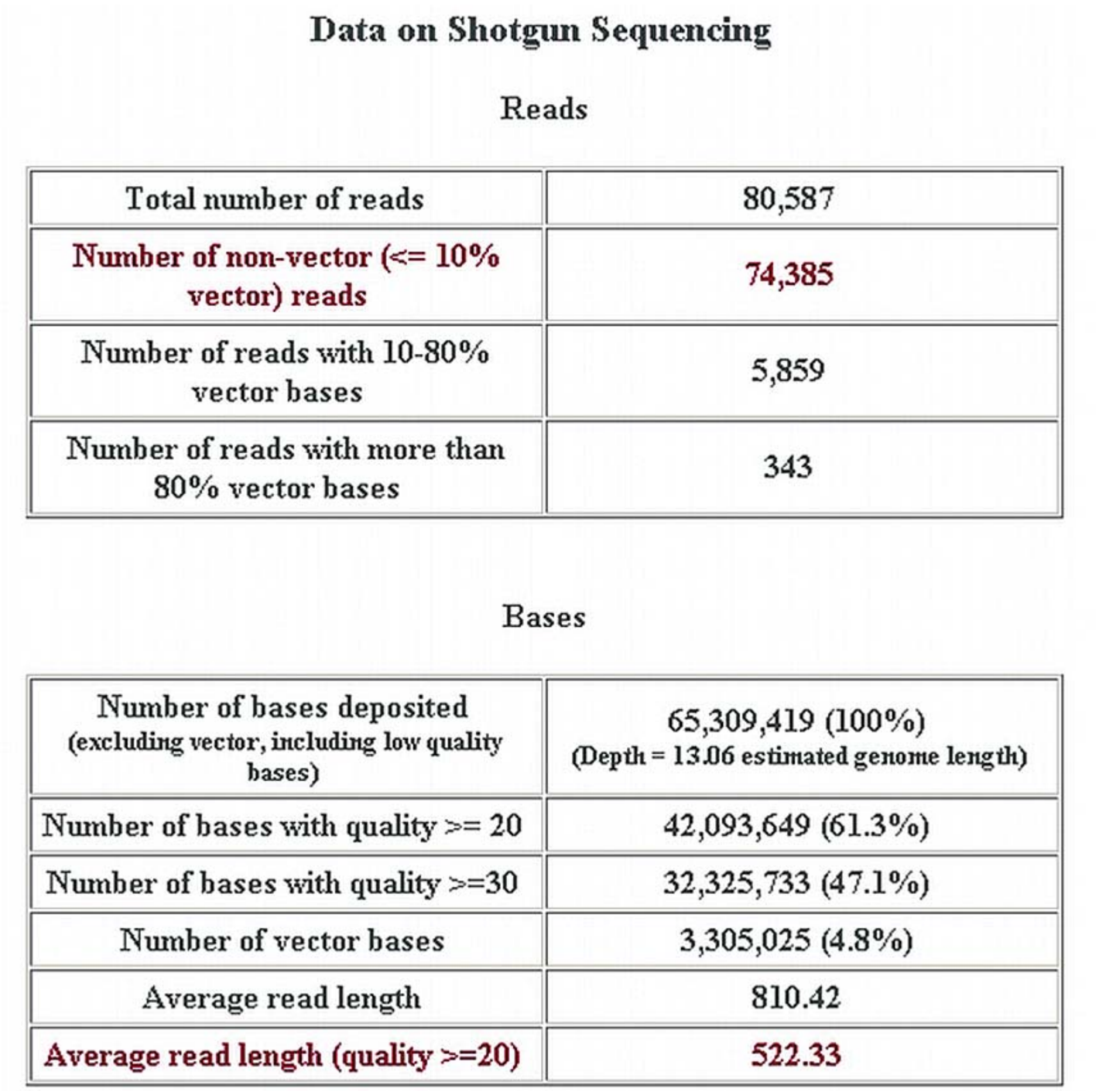


Figure 4. Evolution of contig number.

by SABIA and various types of data are transferred to the database. This follow-up can be made by means of graphic reports (Figure 4), or through text reports carrying information such as: total number of reads submitted with or without vectors, total number of bases with desired quality, number of singlets, singletons and contigs, number of reads used to assemble the contigs, distribution of the contigs according to the number of reads, among others (Figures 5, 6 and 7). These data, associated with the graphic reports, enable a vision of the assembly progress, indicating the end of large-scale sequencing and the beginning of genome finalization.

Repetitions

Repetitive regions in the genome can cause serious assembly problems, and therefore they should be filtered and analyzed separately. These regions can be identified by the occurrence of reads in a quantity far superior to the average of the rest of the genome, the existence of an elevated number of bases with HQD (high quality discrepancy), or the identification of repetitive regions, such as rRNA operons and transposases based on the search of sequence database (nr).



Data on Shotgun Sequencing

Reads

Total number of reads	80,587
Number of non-vector (<= 10% vector) reads	74,385
Number of reads with 10-80% vector bases	5,859
Number of reads with more than 80% vector bases	343

Bases

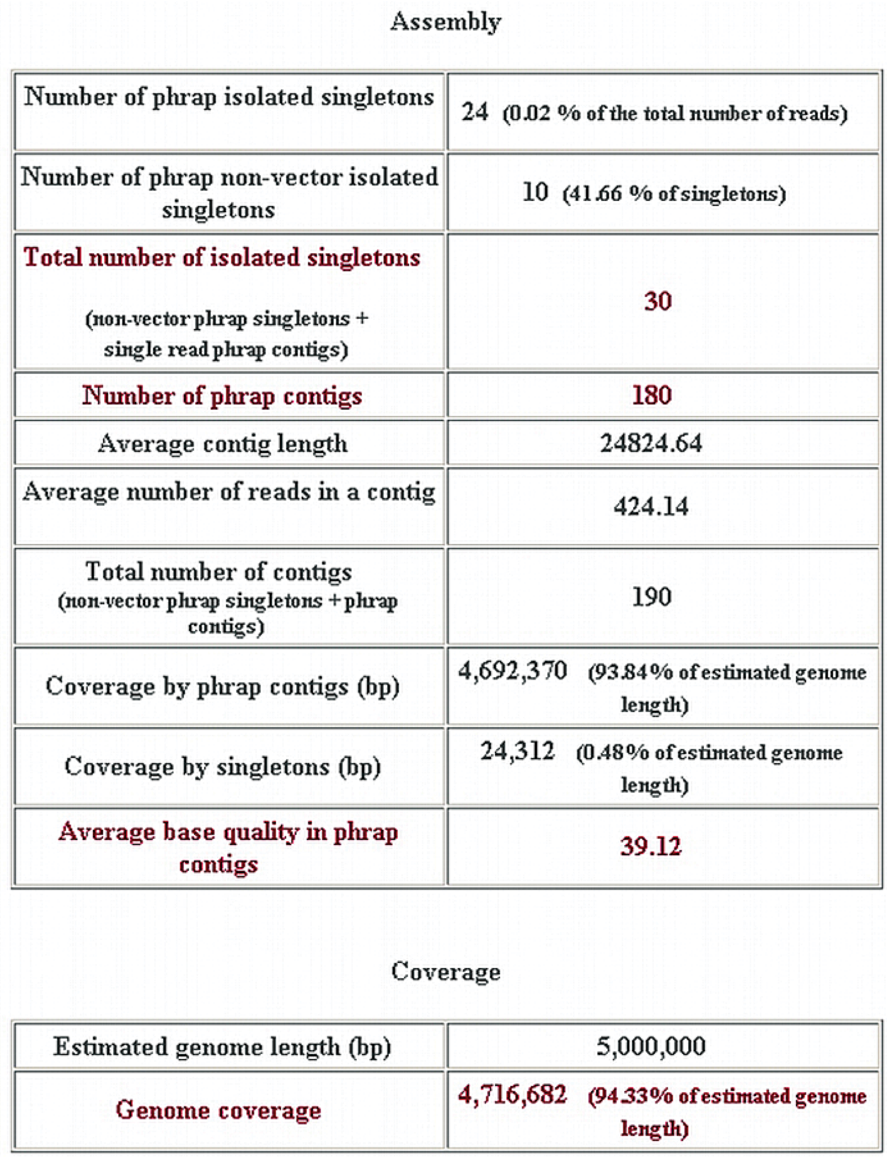
Number of bases deposited (excluding vector, including low quality bases)	65,309,419 (100%) (Depth = 13.06 estimated genome length)
Number of bases with quality >= 20	42,093,649 (61.3%)
Number of bases with quality >=30	32,325,733 (47.1%)
Number of vector bases	3,305,025 (4.8%)
Average read length	810.42
Average read length (quality >=20)	522.33

Figure 5. General assemblies report.

SABIA runs these tasks automatically, searching for regions where the density of the reads is greater than the average density in the rest of the genome and executing the alignment of the assembly contigs. The result of this alignment is stored in the database, and a report with the significant alignments is made available on the web, to be analyzed and eventually selected for screening.

Cosmids submission

Libraries of cosmids, with an average size of 40.000 bp, allowed the confirmation of the contig assemblies, as well as the identification of the connections between them. At first only the cosmid’s ends were sequenced, but as probable gap-closings were identified, they were completely sequenced. The nomenclature of the cosmid reads followed a particular pattern in



Assembly

Number of phrap isolated singletons	24 (0.02 % of the total number of reads)
Number of phrap non-vector isolated singletons	10 (41.66 % of singletons)
Total number of isolated singletons (non-vector phrap singletons + single read phrap contigs)	30
Number of phrap contigs	180
Average contig length	24824.64
Average number of reads in a contig	424.14
Total number of contigs (non-vector phrap singletons + phrap contigs)	190
Coverage by phrap contigs (bp)	4,692,370 (93.84% of estimated genome length)
Coverage by singletons (bp)	24,312 (0.48% of estimated genome length)
Average base quality in phrap contigs	39.12

Coverage

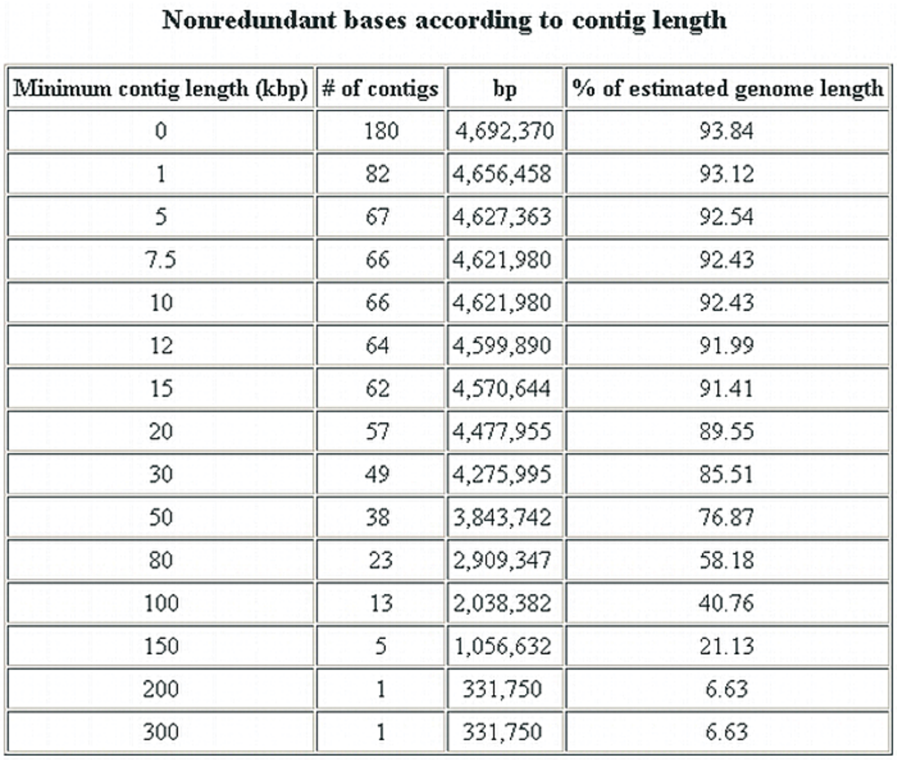
Estimated genome length (bp)	5,000,000
Genome coverage	4,716,682 (94.33% of estimated genome length)

Figure 6. General assemblies report.

order to distinguish them from other assembly reads. The sequencing laboratories could submit the cosmids in two ways: by means of an ace file generated by the phrap (assembly), or by means of the read list. Both the reads and the assembly of the cosmids, as well as the analysis of their quality, were stored in the database.

Scaffold analysis

The scaffold program was used as soon as the contig number began to decrease and the cosmid ends were submitted. This program generates a map from the phrap.out data, with



Nonredundant bases according to contig length

Minimum contig length (kbp)	# of contigs	bp	% of estimated genome length
0	180	4,692,370	93.84
1	82	4,656,458	93.12
5	67	4,627,363	92.54
7.5	66	4,621,980	92.43
10	66	4,621,980	92.43
12	64	4,599,890	91.99
15	62	4,570,644	91.41
20	57	4,477,955	89.55
30	49	4,275,995	85.51
50	38	3,843,742	76.87
80	23	2,909,347	58.18
100	13	2,038,382	40.76
150	5	1,056,632	21.13
200	1	331,750	6.63
300	1	331,750	6.63

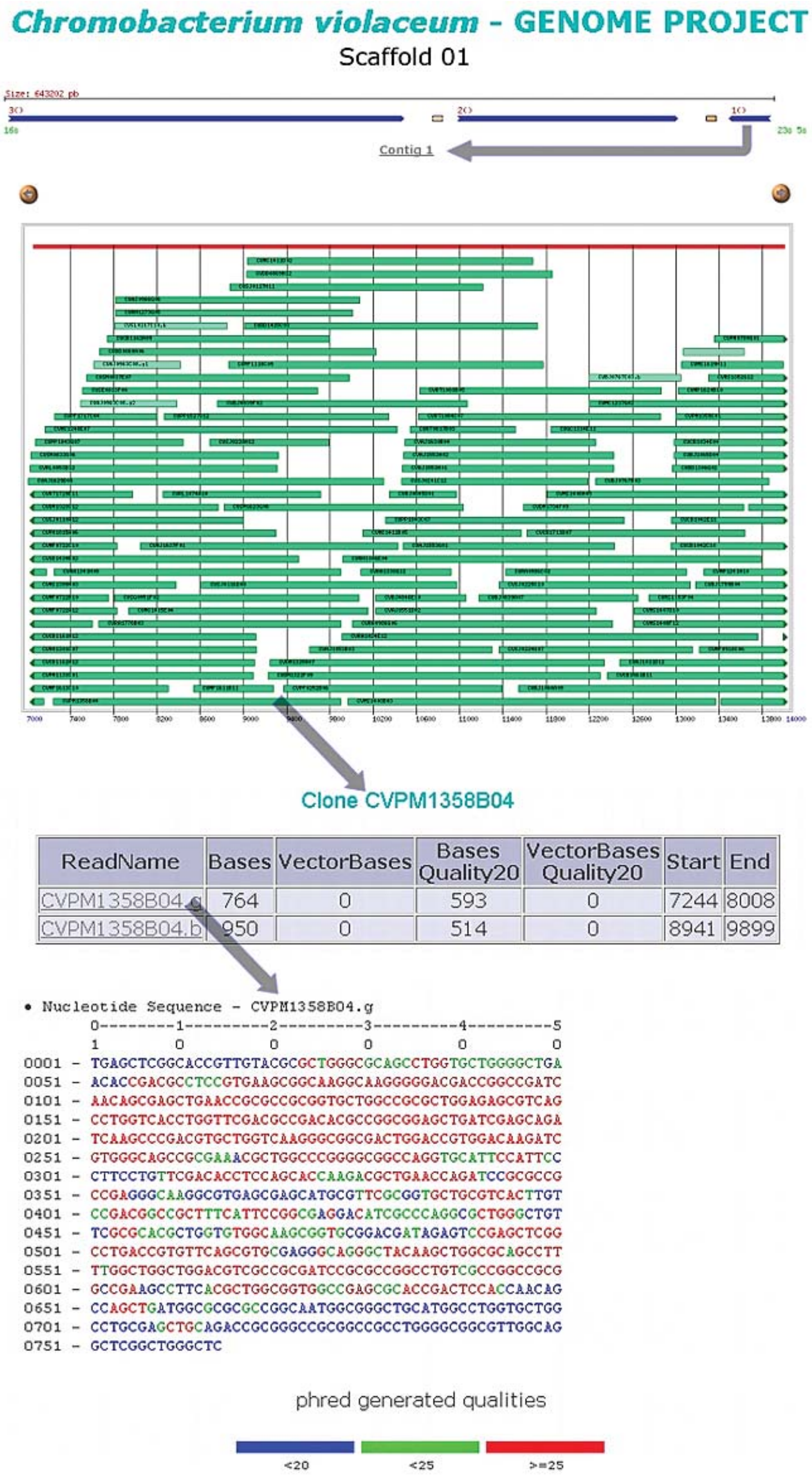
Figure 7. General assemblies report.

one or more contig chains, with corresponding ordering and orientation. This program also takes into account the phrap-estimated distance between the shotgun and cosmid read ends. This distance should be compatible with the estimated clone size of each library. In this way there is an indication of gaps between the contigs (virtual gaps) and those gaps that are not connected with other contigs (real gaps). As the output data of the scaffold program is loaded into the SABIA database, it becomes possible to access the list of clones covering the gap region (Figure 8). If a repetition filter has originated this gap, a list of the filtered read ends is shown. This information is useful for the genome closing process described below.

Genome closing

The closing phase includes two stages: first, evaluation of the contig quality; second, the identification of the solution for closing existing gaps.

The first stage consists of the identification of assembly problems, such as LCQ (low consensus quality): regions with phrap quality score below 25, and HQD (high quality discrepancy): high quality regions that differ from the consensus sequence and the NCBS (not confirmed both strands), since they do not show aligned reads in both orientations. The general assembly of the genome is then frozen (reference assembly) and the assembly manually executed for each contig. The related information is loaded into the database. After this stage, eventual problems are solved by the re-sequencing of shotgun read(s), by the specific primers drawings for the region, or by complete clone sequencing. This information is available to the



ReadName	Bases	VectorBases	Bases Quality20	VectorBases Quality20	Start	End
CVPM1358B04.g	764	0	593	0	7244	8008
CVPM1358B04.b	950	0	514	0	8941	9899

```
• Nucleotide Sequence - CVPM1358B04.g
       0---------1---------2---------3---------4---------5
       1         0         0         0         0         0
0001 - TGAGCTCGGCACCGTTGTACGCGCTGGGCGCAGCCTGGTGCTGGGGCTGA
0051 - ACACCGACGCCTCCGTGAAGCGGCAAGGCAAGGGGGACGACCGGCCGATC
0101 - AACAGCGAGCTGAACCGCGCCGCGGTGCTGGCCGCGCTGGAGAGCGTCAG
0151 - CCTGGTCACCTGGTTCGACGCCGACACGCCGGCGGAGCTGATCGAGCAGA
0201 - TCAAGCCCGACGTGCTGGTCAAGGGCGGCGACTGGACCGTGGACAAGATC
0251 - GTGGGCAGCCGCGAAACGCTGGCCCGGGGCGGCCAGGTGCATTCCATTCC
0301 - CTTCCTGTTCGACACCTCCAGCACCAAGACGCTGAACCAGATCCGCGCCG
0351 - CCGAGGGCAAGGCGTGAGCGAGCATGCGTTCGCGGTGCTGCGTCACTTGT
0401 - CCGACGGCCGCTTTCATTCCGGCGAGGACATCGCCCAGGCGCTGGGCTGT
0451 - TCGCGCACGCTGGTGTGGCAAGCGGTGCGGACGATAGAGTCCGAGCTCGG
0501 - CCTGACCGTGTTCAGCGTGCGAGGGCAGGGCTACAAGCTGGCGCAGCCTT
0551 - TTGGCTGGCTGGACGTCGCCGCGATCCGCGCCGGCCTGTCGCCGGCCGCG
0601 - GCCGAAGCCTTCACGCTGGCGGTGGCCGAGCGCACCGACTCCACCAACAG
0651 - CCAGCTGATGGCGCGCGCCGGCAATGGCGGGCTGCATGGCCTGGTGCTGG
0701 - CCTGCGAGCTGCAGACCGCGGGCCGCGGCCGCCTGGGGCGGCGTTGGCAG
0751 - GCTCGGCTGGGCTC
```

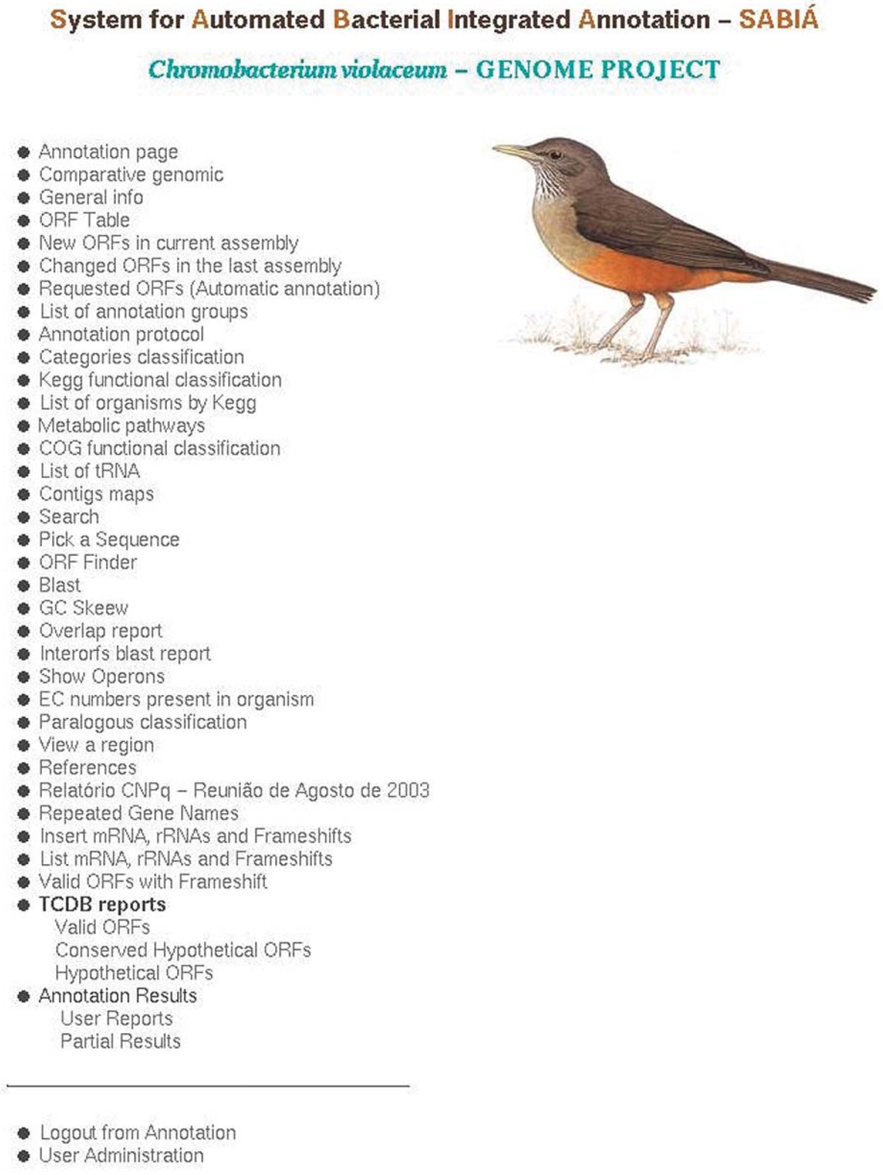
phred generated qualities

🟦 <20 🟩 <25 🟥 >=25

Figure 8. Scaffold's map showing the contigs with the respective clones.

sequencing groups in the project home page, which is updated in parallel with the reception of the respective submissions, provided the phrap quality is ≥25 for all bases.

At the second stage, two approaches are adopted: first, the automatic identification of the contig read ends that have not yet been submitted in both directions. This list is automatically generated and made available on the web. Second, the scaffold visualization tool is used to select the plasmid and cosmid clones that might close the gaps. For gaps generated by the repetition filters, the region is assembled by means of a read subgroup that is filtered until a sequence is found that can be anchored in both gap contigs, or else through the sequencing of cosmid or plasmid clones.



System for Automated Bacterial Integrated Annotation – SABIÁ

Chromobacterium violaceum **– GENOME PROJECT**

- Annotation page
- Comparative genomic
- General info
- ORF Table
- New ORFs in current assembly
- Changed ORFs in the last assembly
- Requested ORFs (Automatic annotation)
- List of annotation groups
- Annotation protocol
- Categories classification
- Kegg functional classification
- List of organisms by Kegg
- Metabolic pathways
- COG functional classification
- List of tRNA
- Contigs maps
- Search
- Pick a Sequence
- ORF Finder
- Blast
- GC Skeew
- Overlap report
- Interorfs blast report
- Show Operons
- EC numbers present in organism
- Paralogous classification
- View a region
- References
- Relatório CNPq – Reunião de Agosto de 2003
- Repeated Gene Names
- Insert mRNA, rRNAs and Frameshifts
- List mRNA, rRNAs and Frameshifts
- Valid ORFs with Frameshift
- **TCDB reports**
  Valid ORFs
  Conserved Hypothetical ORFs
  Hypothetical ORFs
- Annotation Results
  User Reports
  Partial Results

- Logout from Annotation
- User Administration

Figure 9. SABIA page showing all tools available.

Finally, with the help of consed, the frozen assembly with the ordered and closed gaps are converted to the FASTA format and transferred to the SABIA module.

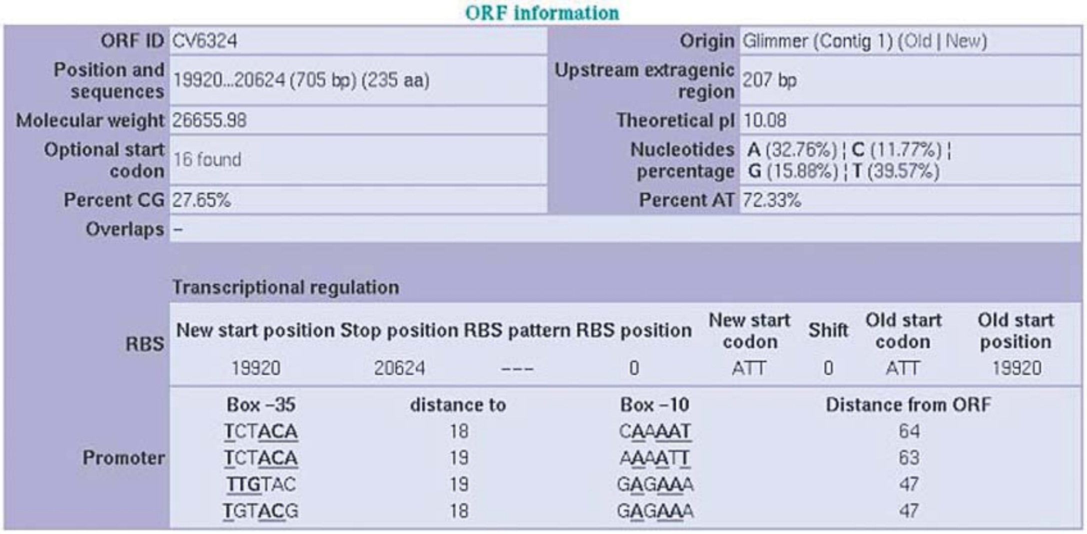
Automatic annotation

The annotation module carries out the identification and functional categorization of all ORFs found in the genome (Figure 9).

ORFs identification

The annotation process begins through the FASTA format contigs, with or without the respective qualities. The first annotation phase consists in an automatic search for ORFs and tRNAs. tRNAScan-SE was used for the tRNAs prediction. The programs used for ORF prediction were Glimmer, which uses Markov's interpolated models, and GeneMark, which uses heuristic models. The annotation module allows only one of these programs to be used. These ORFs prediction programs must train their models with data from other organisms, preferably those situated phylogenetically close. Models extracted from *E. coli* were first used in this project, due to its well-known extensive annotation process; in a second phase, the ORFs of the genome itself were used. The RBSfinder program (www.tigr.org), which searches for ribosome-binding sites in the extragenic regions was also used, in order to increase the reliability of the Glimmer and GeneMark results. To accomplish this, the module that manages the ORFs identification filters the results, generating a single coordinate file, which is then used as the input of the RBSfinder program.

ORF identification was performed automatically, taking into account the coordinates produced by the prediction programs and the output file of the RBS finder. After this procedure, information, such as the RBS position in the genome, new options for the initial codon, and the suggested shift for the RBS correction are stored in the database (Figure 10).



ORF information

ORF ID	CV6324	Origin	Glimmer (Contig 1) (Old | New)
Position and sequences	19920...20624 (705 bp) (235 aa)	Upstream extragenic region	207 bp
Molecular weight	26655.98	Theoretical pI	10.08
Optional start codon	16 found	Nucleotides percentage	A (32.76%) ¦ C (11.77%) ¦ G (15.88%) ¦ T (39.57%)
Percent CG	27.65%	Percent AT	72.33%
Overlaps	–		

Transcriptional regulation

RBS

New start position	Stop position	RBS pattern	RBS position	New start codon	Shift	Old start codon	Old start position
19920	20624	---	0	ATT	0	ATT	19920

Promoter

Box –35	distance to	Box –10	Distance from ORF
TCTACA	18	CAAAAT	64
TCTACA	19	AAAATT	63
TTGTAC	19	GAGAAA	47
TGTACG	18	GAGAAA	47

Figure 10. ORF's data generated by the automatic annotation.

The next step is the identification of the extragenic regions for each of the ORFs, with the purpose of: i) looking for other possible initiation codons (optional start codons) in this region and in the 99 initial bases of each ORF; the purpose of this procedure is to reduce the overlaps between ORFs and to find the correct position of initiation codons; ii) looking for promoter boxes similar to the consensus sequence - 35 (TTGACA) and - 10 (TATAAT), with acceptance of up to three mismatches in each box and of 16 to 19 bp as the distance between them.

Information about all the ORFs identified by the SABIA and stored in the database includes their nucleotide and protein sequences, associated with their phrap quality, as well as the nucleotide percentages, isoelectric points (IP) and molecular weights (MW).

Also, using this module, genomic maps were generated, allowing the visualization of the ORF localizations. SABIA provides two types of maps, one showing all identified ORFs and another showing only the categorized ORFs. The size of each one of these maps may be configured to best suit the project's needs, thereby allowing a group of annotators, for example, to have a particular map under its direct responsibility. The ORFs and other structures are represented by rectangles of different colors, and are functionally classified according to the KEGG or COG. An inscription describes each functional classification and its respective color in

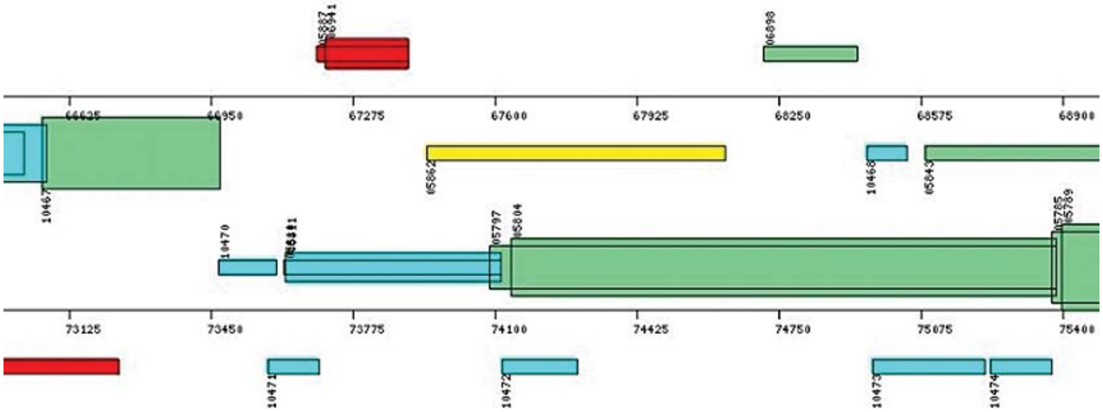


Figure 11. Map of a specific genome region showing the identified ORFs. Colors according to KEGG's functional classification. Overlapped ORFs represented by large rectangles.

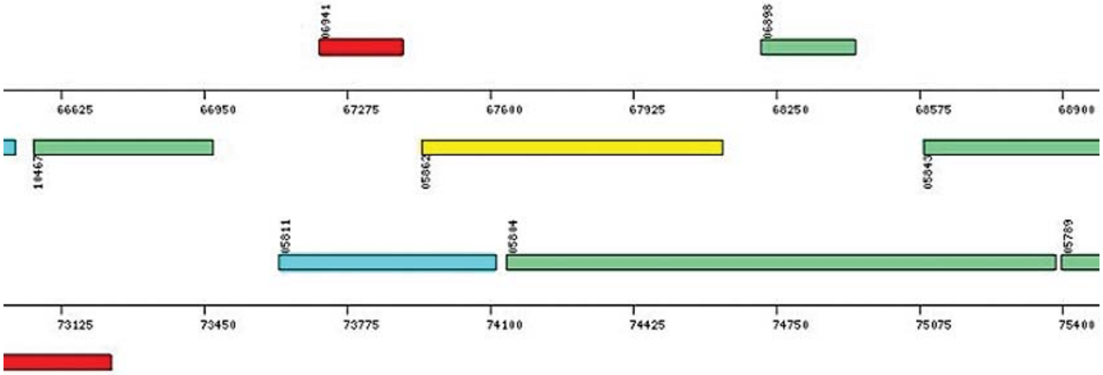
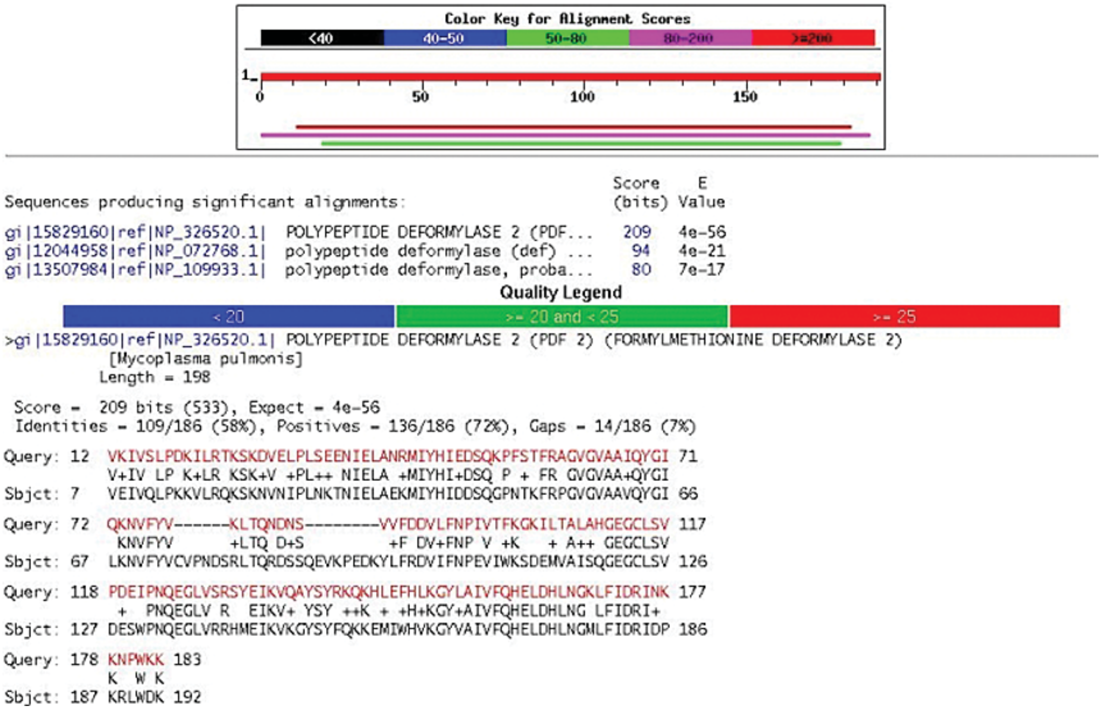


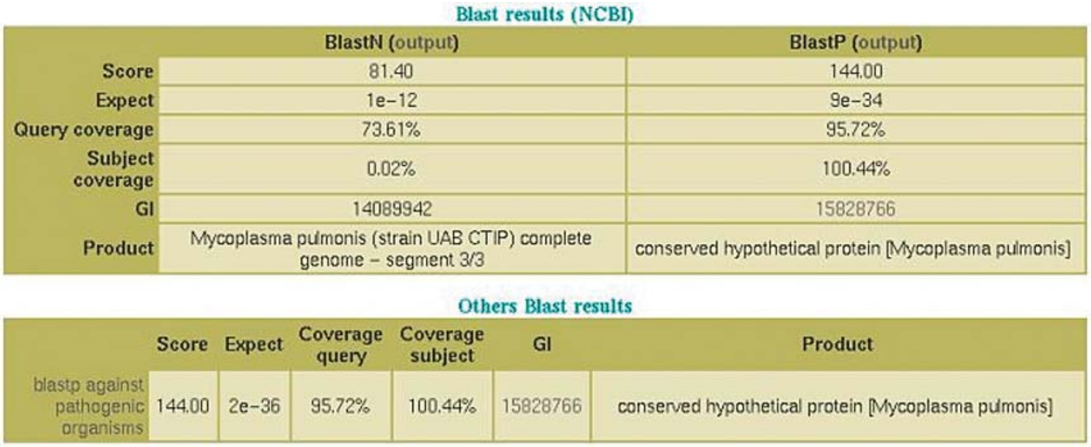
Figure 12. Map of a specific genome region showing the categorized ORFs.

the maps. The height of the rectangles is proportional to the number of overlapping bases between two or more ORFs. All ORFs in the map are “clickable” and take the annotator to the annotation page of the corresponding ORF. If the browser allows the use of java script, by moving the mouse over an ORF the annotator obtains its functional description, and its start and end positions. Besides showing the distribution of and information about the ORFs, the maps allow the visualization of the tRNAs, mRNAs, rRNAs and frameshifts (Figures 11 and 12).



```
                                                           Score    E
Sequences producing significant alignments:                (bits) Value

gi|15829160|ref|NP_326520.1|  POLYPEPTIDE DEFORMYLASE 2 (PDF...   209   4e-56
gi|12044958|ref|NP_072768.1|  polypeptide deformylase (def) ...    94   4e-21
gi|13507984|ref|NP_109933.1|  polypeptide deformylase, proba...    80   7e-17

>gi|15829160|ref|NP_326520.1| POLYPEPTIDE DEFORMYLASE 2 (PDF 2) (FORMYLMETHIONINE DEFORMYLASE 2)
          [Mycoplasma pulmonis]
         Length = 198

 Score =  209 bits (533), Expect = 4e-56
 Identities = 109/186 (58%), Positives = 136/186 (72%), Gaps = 14/186 (7%)

Query: 12  VKIVSLPDKILRTKSKDVELPLSEENIELANRMIYHIEDSQKPFSTFRAGVGVAAIQYGI 71
            V+IV LP K+LR KSK+V +PL++ NIELA +MIYHI+DSQ P + FR GVGVAA+QYGI
Sbjct: 7   VEIVQLPKKVLRQKSKNVNIPLNKTNIELAEKMIYHIDDSQGPNTKFRPGVGVAAVQYGI 66

Query: 72  QKNVFYV------KLTQNDNS---------VVFDDVLFNPIVTFKGKILTALAHGEGCLSV 117
             KNVFYV      +LTQ D+S         +F DV+FNP V +K    + A++ GEGCLSV
Sbjct: 67  LKNVFYVCVPNDSRLTQRDSSQEVKPEDKYLFRDVIFNPEVIWKSDEMVAISQGEGCLSV 126

Query: 118 PDEIPNQEGLVSRSYEIKVQAYSYRKQKHLEFHLKGYLAIVFQHELDHLNGKLFIDRINK 177
             +  PNQEGLV R  EIKV+ YSY ++K + +H+KGY+AIVFQHELDHLNG LFIDRI+
Sbjct: 127 DESWPNQEGLVRRHMEIKVKGYSYFQKKEMIWHVKGYVAIVFQHELDHLNGMLFIDRIDP 186

Query: 178 KNPWKK 183
            K  W K
Sbjct: 187 KRLWDK 192
```

Figure 13. blastp result for an ORF.



Blast results (NCBI)

	BlastN (output)	BlastP (output)
Score	81.40	144.00
Expect	1e-12	9e-34
Query coverage	73.61%	95.72%
Subject coverage	0.02%	100.44%
GI	14089942	15828766
Product	Mycoplasma pulmonis (strain UAB CTIP) complete genome - segment 3/3	conserved hypothetical protein [Mycoplasma pulmonis]

Others Blast results

	Score	Expect	Coverage query	Coverage subject	GI	Product
blastp against pathogenic organisms	144.00	2e-36	95.72%	100.44%	15828766	conserved hypothetical protein [Mycoplasma pulmonis]

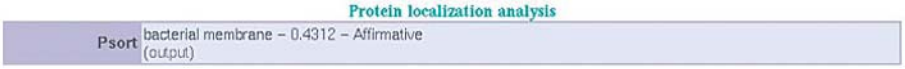
Figure 14. Results of blastn, blastp for NCBI database and blastp for pathogenic organism database.

Functional classification

In the analysis of the nucleotide and amino acid sequences, SABIA manages the use of five programs of the Blast family: blastn, blastp, blastx, tblastn, and tblastx, which run through the server version (WWWBlastServer), allowing the alignment images to be generated, classified according to their scores, making the visualization of the results easier. An additional database was used in the *C. violaceum* project, dealing exclusively with pathogenic organism sequences. When the system accesses the base quality file, it automatically alters the final file, indicating the quality of each one of the bases in the alignment by means of a color pattern. In addition, the system automatically informs the score values, the expectation value (e-value), query coverage and subject coverage (Figures 13 and 14).

The amino acid sequence is also used as an input for the PSORT (Nakai and Kanehisa, 1991) program, which predicts the location of the protein in the cell. ORFs classified by PSORT as membrane proteins are automatically aligned in relation to the sequence of the TCDB bank (Saier, 1999), by means of the BLASTP program. In this way it is possible to classify these proteins according to the information from this bank, and according to the already known transport protein number.

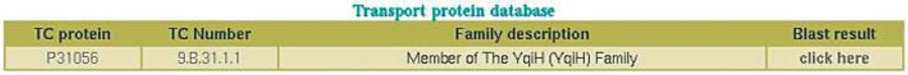
The blastp and blastn programs are executed for each generated ORF, using the database of the KEGG (Kyoto Encyclopedia of Genes and Genomes), which contains more than 120 organisms. SABIA selects the best general result and also shows the results obtained for the *E. coli* genome. The information from these results, such as the organism, gene name (usually a four-letter annotation), synonyms, links to external sites (containing further data on the gene), metabolic pathways, unique functional classification, and EC number, is stored in the database (Figures 15 and 16).



Protein localization analysis

Psort	bacterial membrane – 0.4312 – Affirmative (output)

Figure 15. PSORT result.



Transport protein database

TC protein	TC Number	Family description	Blast result
P31056	9.B.31.1.1	Member of The YqiH (YqiH) Family	click here

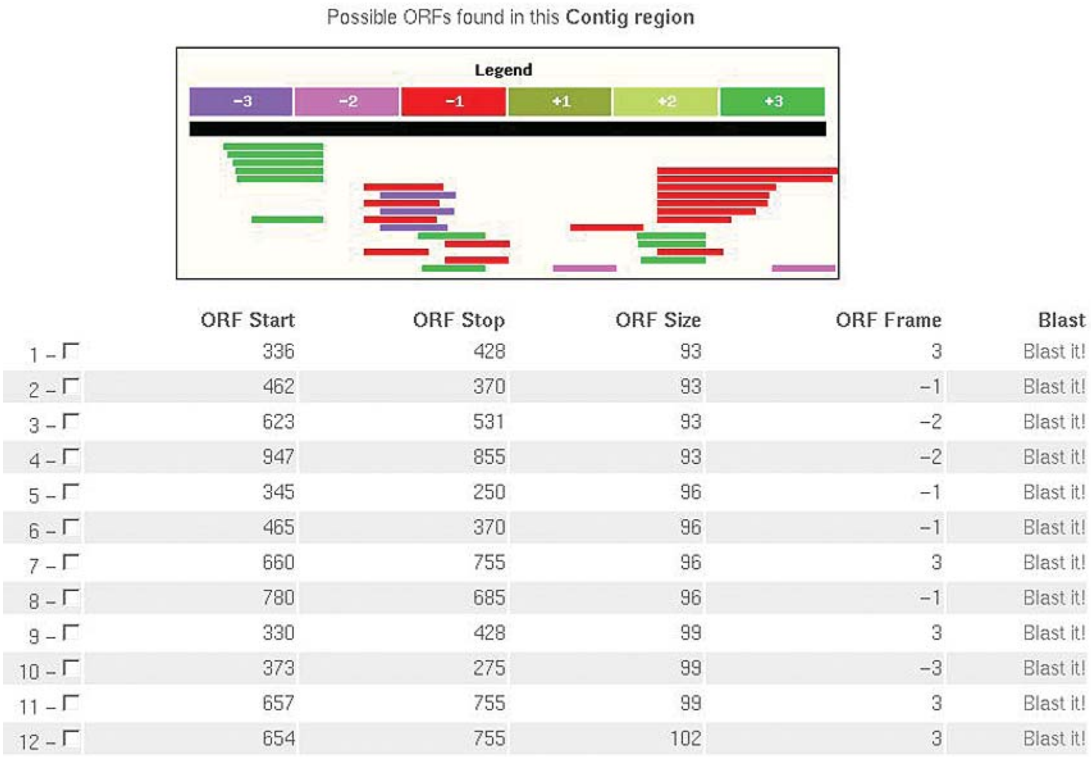
Figure 16. TCDB result.

The amino acid sequence is used as an input for the local execution of InterPro. The information generated by this program is stored in the database: ID, Name, InterPro ID, InterPro name, GO, besides providing external links. The protein sequence of each ORF is used for local consultation in the COG database.

Whenever the KEGG blast provides a result for an *E. coli* gene, it is used for the ORF functional classification, as recommended by Riley (1998).

SABIA allows the insertion of new ORFs in the genome, by means of a tool denominated “pick a sequence”, which identifies six possible ORFs; these are graphically shown, in a given region of the genome, each with a link for the execution of a Blast program (Figure 17). Furthermore, SABIA allows structures, such as mRNAs, rRNAs and frameshifts, to be manually inserted.

The backup of all tables in the SABIA database can be scheduled for periodic execution (for *C. violaceum*, a daily schedule was adopted). All tables are stored in a single file. After



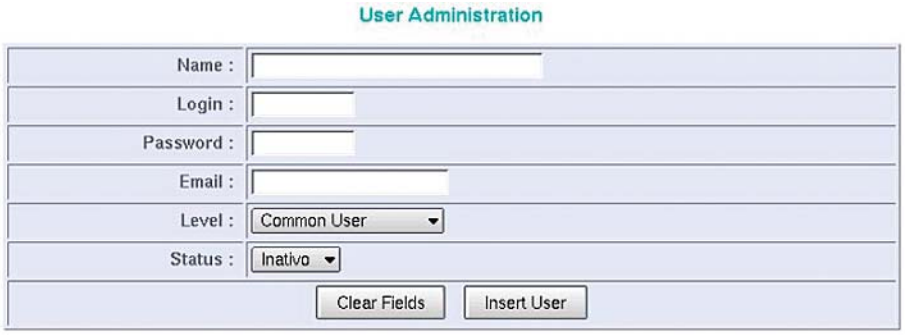
Possible ORFs found in this **Contig region**

	ORF Start	ORF Stop	ORF Size	ORF Frame	Blast
1 – □	336	428	93	3	Blast it!
2 – □	462	370	93	–1	Blast it!
3 – □	623	531	93	–2	Blast it!
4 – □	947	855	93	–2	Blast it!
5 – □	345	250	96	–1	Blast it!
6 – □	465	370	96	–1	Blast it!
7 – □	660	755	96	3	Blast it!
8 – □	780	685	96	–1	Blast it!
9 – □	330	428	99	3	Blast it!
10 – □	373	275	99	–3	Blast it!
11 – □	657	755	99	3	Blast it!
12 – □	654	755	102	3	Blast it!

Figure 17. "Pick a Sequence" tool, showing all six possible frames.

"batch" processing, which performs the automatic annotation, the system loads all the information that is produced into the database. This information is available through a simple and intuitive web interface.

Access to the web interface is limited to registered users authorized by the system administrator. There are three levels of access: i) the annotator, who may annotate and request new ORFs to the system; ii) the coordinator, who is able to end the annotation process for a specific group of ORFs; iii) the user, who is only allowed to examine the data and annotation through a web page (Figure 18).



User Administration

Name :	
Login :	
Password :	
Email :	
Level :	Common User
Status :	Inativo

Figure 18. Screen of administrative user's attributes.

In the earlier stages of the *C. violaceum* project, the annotator could visualize two graphs (Figure 19) on the web page, the first containing the genomic localization of the ORF and the second showing a summary of the information provided by the annotation module. Later on, two information blocks about the ORF were presented; the first had the ORF identification, the program used for its identification, its contig number, its position in the genome, the nucleotide and amino acid sequences, with their respective qualities, besides the information regarding the extragenic region: promoters, RBS and optional initiation codons, with links for the blastn or blastp programs.

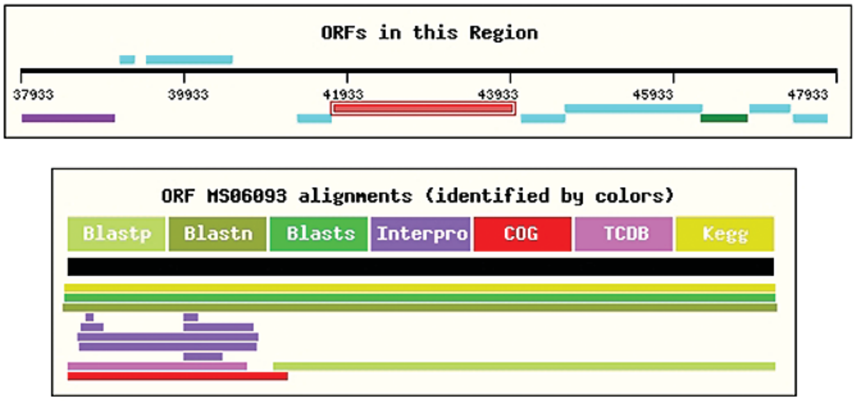


Figure 19. Several annotation illustrations.

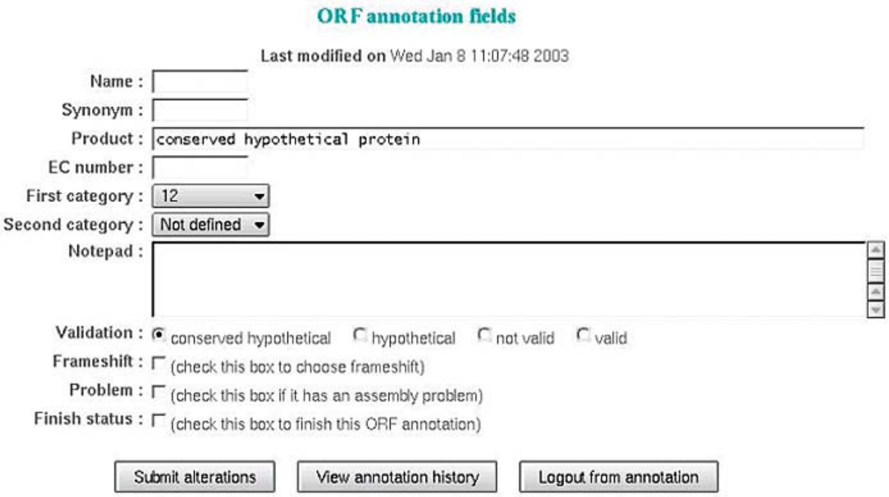
The next block informed the best alignment derived from the blastn and blastp programs. Furthermore, information such as score, expectation value, query coverage, subject coverage, GI, and the product was available to the annotator. Finally the best results of the COG, KEGG and InterPro programs were shown.

Annotator

The annotation block is the part of the system where the annotator inserts the results of his final analysis, after evaluating all available information (Figure 20). The annotator is expected to insert the name of the gene, with eventual synonyms, EC number and primary and secondary categories. The annotator may describe useful details about the sequence under scrutiny in a notepad. This block permits access to the annotation report, where all the modifications can be visualized, as well as the time of annotation, the user's name, and the product description. There is also an option of automatic annotation request to start optional initial codons and identified ORFs through the "pick a sequence" tool.

The annotator classifies the ORF based on all the information generated by the automatic annotation. The following categories were adopted in the Brazilian genome project:

- Valid ORF: whenever there was an extremely well-defined product.
- Hypothetical conserved ORF: with similarities to other conserved ORFs or little similarity with valid ORFs in other organisms.



ORF annotation fields

Last modified on Wed Jan 8 11:07:48 2003

Name :	
Synonym :	
Product :	conserved hypothetical protein
EC number :	
First category :	12
Second category :	Not defined
Notepad :	
Validation :	● conserved hypothetical ○ hypothetical ○ not valid ○ valid
Frameshift :	□ (check this box to choose frameshift)
Problem :	□ (check this box if it has an assembly problem)
Finish status :	□ (check this box to finish this ORF annotation)

Figure 20. Screen for inserting annotation data.

- Hypothetical ORF: with no significant results in the Blast program.
- Invalid ORF: i) with an overlap greater than 10 amino acids with other ORFs or ii) size below 50 amino acids.

Additional functions include:

- Submit alterations: new information provided by the user is kept in the database.
- View annotation history: presents a page containing all previous annotations on the ORF.
- Logout from annotation: to exit the annotation phase.
- Optional first start: makes the ORF first start option available, selected during automatic annotation.

Assembly updating

The annotation module allows the update of the assembly already loaded in the bank, without losing existing information. The assembly update process is carried out safely and in a coordinated manner by a group of scripts. All new sequences are compared with the sequences downloaded in the database by using the Crossmatch program. The system will process three different situations: i) update the quantity of ORFs perfectly aligned with the ones found in the database; ii) accomplish automatic annotations for the new ORFs; iii) mark the ORFs that are no longer present in the new assembly or had some modification made in their base sequences. After the assembly updating process the system displays two reports through the web interface: a report of the ORFs found in the new version and a report of the ORFs that no longer exist.

Verification of the ORFs in the extragenic region

To determine whether all coding structures (ORFs, mRNAs, tRNAs) were identified, a group of scripts examines all the extragenic regions, with the help of the blastn and blastp programs, which search for such sequences. The process provides reports on the possible structures found and opens a link for the “pick a sequence” tool to be applied wherever needed.

RESULTS AND DISCUSSION

Some programs and report forms were developed to make the analysis of annotation easier, and also to correct eventual mistakes; they are available in the project home page.

Comprehensive research in the annotation database (search)

SABIA provides a search system that allows detailed searches in the annotation database. These searches may start from product, EC number, gene name, synonyms, PSORT, sequence, conserved or hypothetical sequence in the PSORT, ID, GI, InterPro ID or name, COG ID, product or functional classification, EC number, definition, classification, KEGG organism or gene name, and *E. coli* gene or products. These searches allow filtering through strings that differ from the pattern informed by the annotator, so that ORFs with similar and relevant characteristics are rapidly found.

Overlapping of ORFs

To prevent large overlaps, a report is produced showing ORFs with overlapping bases, ordered according to the total number of common bases (Figure 21).

Repeated gene names

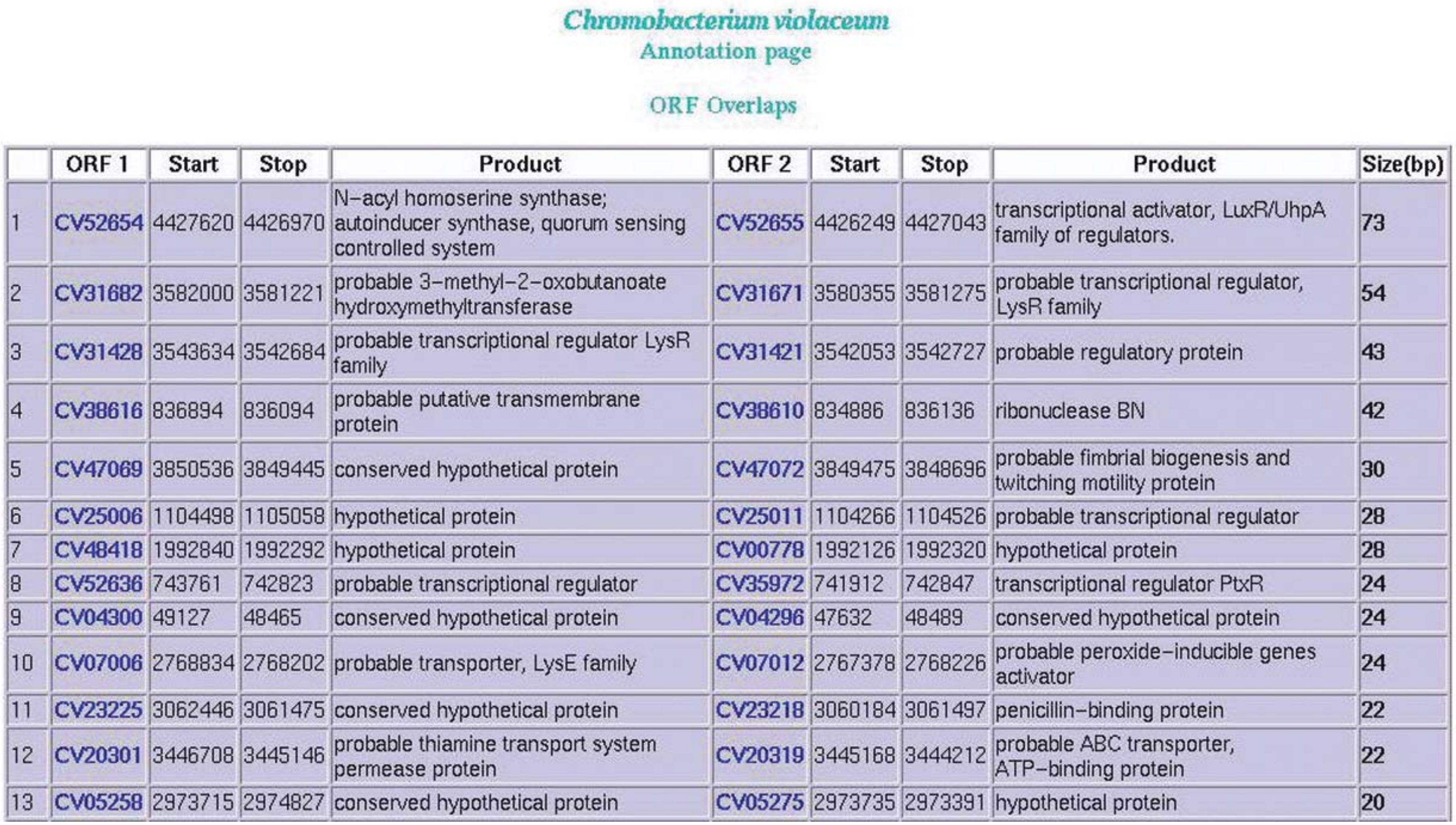
ORFs are grouped by the gene name. Names common to two or more ORFs are highlighted and a revision in the annotation is suggested.

KEGG and EC number

The EC list generated by SABIA is used to improve the annotation quality, by comparing the product name suggested by the annotator with the name recommended by the IUBMB (International Union of Biochemistry and Molecular Biology). The EC number is also used, during the automatic annotation process, to overview the detection of ORFs participating in the numerous steps of metabolic pathways (Figures 22 and 23).

Distribution of ORFs based on similarity

SABIA presents the ORFs distribution by organism, based on the best KEGG hits. For each organism there is a total listing of ORFs and the percentage of the total, compared to the one currently annotated. High correlations suggest a greater similarity between organisms (Figure 24).

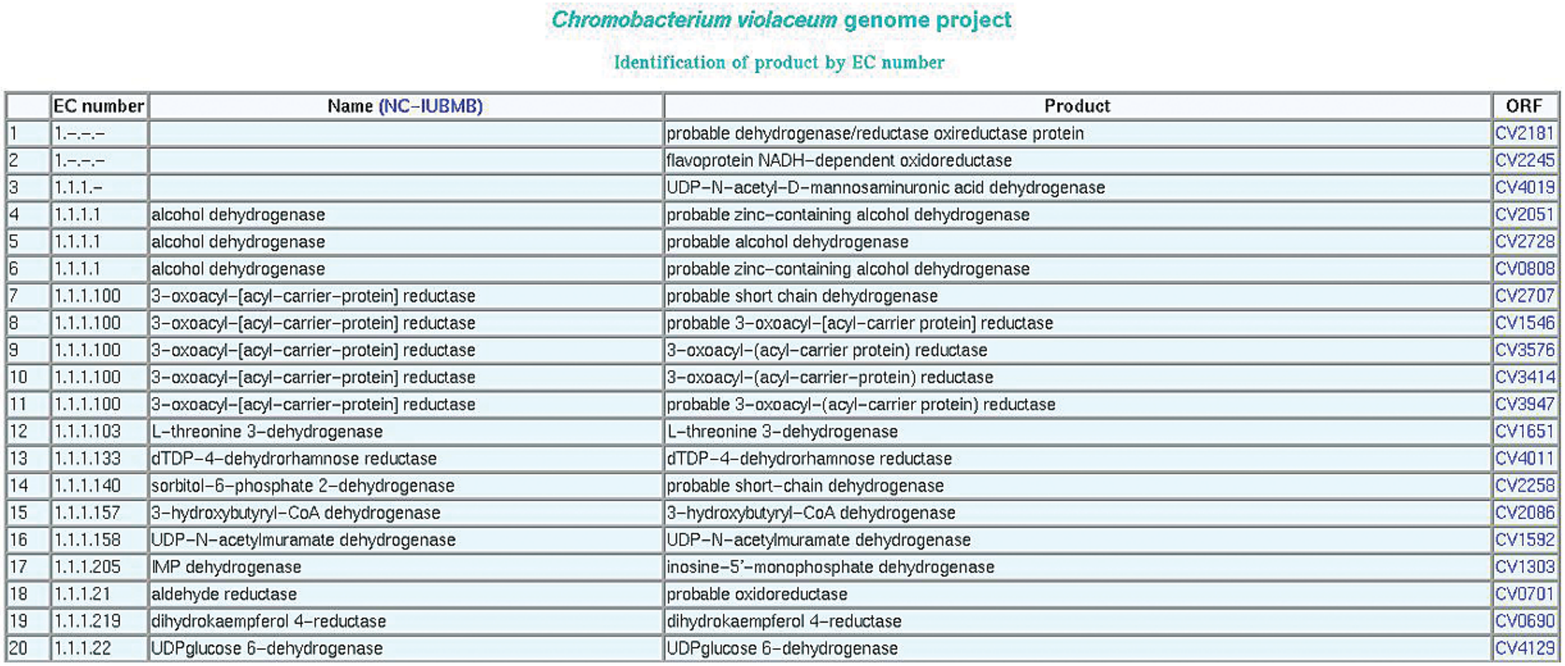


Chromobacterium violaceum
Annotation page

ORF Overlaps

	ORF 1	Start	Stop	Product	ORF 2	Start	Stop	Product	Size(bp)
1	CV52654	4427620	4426970	N−acyl homoserine synthase; autoinducer synthase, quorum sensing controlled system	CV52655	4426249	4427043	transcriptional activator, LuxR/UhpA family of regulators.	73
2	CV31682	3582000	3581221	probable 3−methyl−2−oxobutanoate hydroxymethyltransferase	CV31671	3580355	3581275	probable transcriptional regulator, LysR family	54
3	CV31428	3543634	3542684	probable transcriptional regulator LysR family	CV31421	3542053	3542727	probable regulatory protein	43
4	CV38616	836894	836094	probable putative transmembrane protein	CV38610	834886	836136	ribonuclease BN	42
5	CV47069	3850536	3849445	conserved hypothetical protein	CV47072	3849475	3848696	probable fimbrial biogenesis and twitching motility protein	30
6	CV25006	1104498	1105058	hypothetical protein	CV25011	1104266	1104526	probable transcriptional regulator	28
7	CV48418	1992840	1992292	hypothetical protein	CV00778	1992126	1992320	hypothetical protein	28
8	CV52636	743761	742823	probable transcriptional regulator	CV35972	741912	742847	transcriptional regulator PtxR	24
9	CV04300	49127	48465	conserved hypothetical protein	CV04296	47632	48489	conserved hypothetical protein	24
10	CV07006	2768834	2768202	probable transporter, LysE family	CV07012	2767378	2768226	probable peroxide−inducible genes activator	24
11	CV23225	3062446	3061475	conserved hypothetical protein	CV23218	3060184	3061497	penicillin−binding protein	22
12	CV20301	3446708	3445146	probable thiamine transport system permease protein	CV20319	3445168	3444212	probable ABC transporter, ATP−binding protein	22
13	CV05258	2973715	2974827	conserved hypothetical protein	CV05275	2973735	2973391	hypothetical protein	20

Figure 21. Overlap of a pair of ORFs.

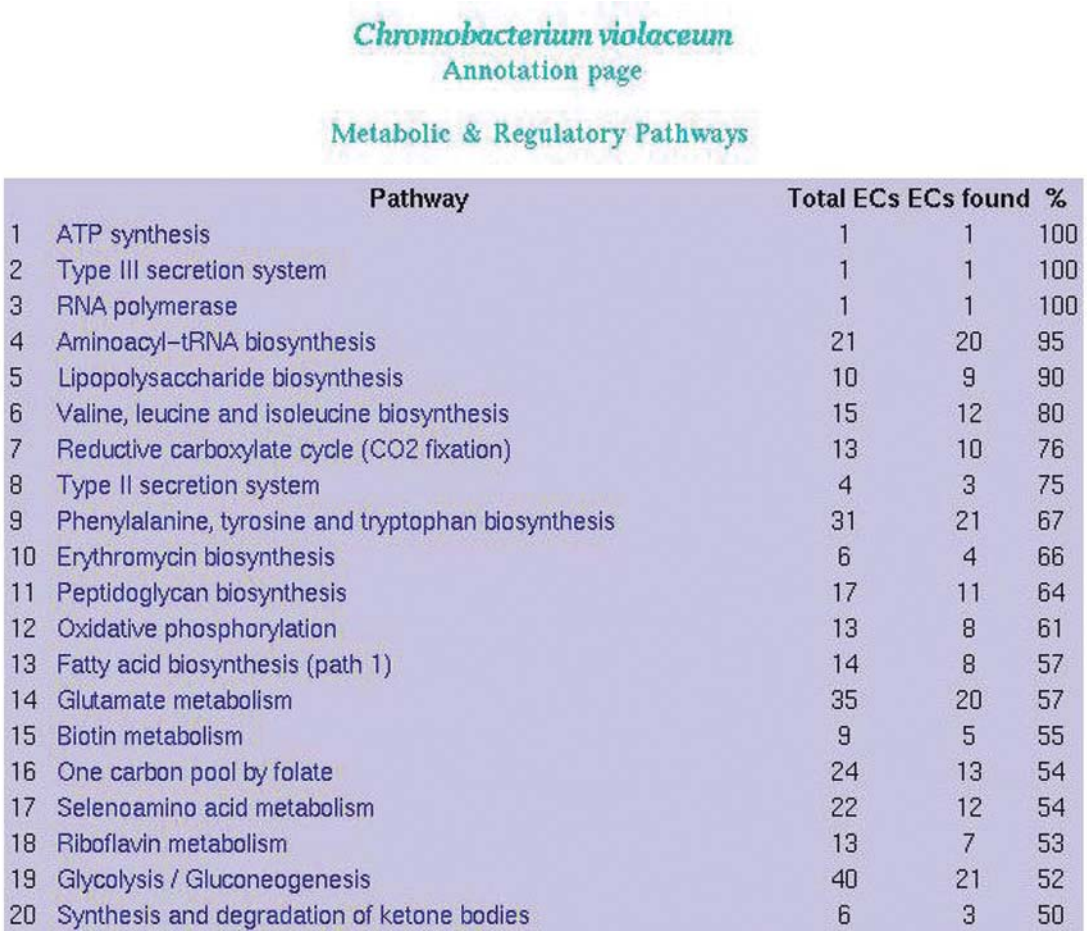


Chromobacterium violaceum genome project

Identification of product by EC number

	EC number	Name (NC–IUBMB)	Product	ORF
1	1.–.–.–		probable dehydrogenase/reductase oxireductase protein	CV2181
2	1.–.–.–		flavoprotein NADH–dependent oxidoreductase	CV2245
3	1.1.1.–		UDP–N–acetyl–D–mannosaminuronic acid dehydrogenase	CV4019
4	1.1.1.1	alcohol dehydrogenase	probable zinc–containing alcohol dehydrogenase	CV2051
5	1.1.1.1	alcohol dehydrogenase	probable alcohol dehydrogenase	CV2728
6	1.1.1.1	alcohol dehydrogenase	probable zinc–containing alcohol dehydrogenase	CV0808
7	1.1.1.100	3–oxoacyl–[acyl–carrier–protein] reductase	probable short chain dehydrogenase	CV2707
8	1.1.1.100	3–oxoacyl–[acyl–carrier–protein] reductase	probable 3–oxoacyl–[acyl–carrier protein] reductase	CV1546
9	1.1.1.100	3–oxoacyl–[acyl–carrier–protein] reductase	3–oxoacyl–(acyl–carrier protein) reductase	CV3576
10	1.1.1.100	3–oxoacyl–[acyl–carrier–protein] reductase	3–oxoacyl–(acyl–carrier–protein) reductase	CV3414
11	1.1.1.100	3–oxoacyl–[acyl–carrier–protein] reductase	probable 3–oxoacyl–(acyl–carrier protein) reductase	CV3947
12	1.1.1.103	L–threonine 3–dehydrogenase	L–threonine 3–dehydrogenase	CV1651
13	1.1.1.133	dTDP–4–dehydrorhamnose reductase	dTDP–4–dehydrorhamnose reductase	CV4011
14	1.1.1.140	sorbitol–6–phosphate 2–dehydrogenase	probable short–chain dehydrogenase	CV2258
15	1.1.1.157	3–hydroxybutyryl–CoA dehydrogenase	3–hydroxybutyryl–CoA dehydrogenase	CV2086
16	1.1.1.158	UDP–N–acetylmuramate dehydrogenase	UDP–N–acetylmuramate dehydrogenase	CV1592
17	1.1.1.205	IMP dehydrogenase	inosine–5'–monophosphate dehydrogenase	CV1303
18	1.1.1.21	aldehyde reductase	probable oxidoreductase	CV0701
19	1.1.1.219	dihydrokaempferol 4–reductase	dihydrokaempferol 4–reductase	CV0690
20	1.1.1.22	UDPglucose 6–dehydrogenase	UDPglucose 6–dehydrogenase	CV4129

Figure 22. List of ORFs and their respective EC_numbers.



Chromobacterium violaceum
Annotation page

Metabolic & Regulatory Pathways

	Pathway	Total ECs	ECs found	%
1	ATP synthesis	1	1	100
2	Type III secretion system	1	1	100
3	RNA polymerase	1	1	100
4	Aminoacyl–tRNA biosynthesis	21	20	95
5	Lipopolysaccharide biosynthesis	10	9	90
6	Valine, leucine and isoleucine biosynthesis	15	12	80
7	Reductive carboxylate cycle (CO2 fixation)	13	10	76
8	Type II secretion system	4	3	75
9	Phenylalanine, tyrosine and tryptophan biosynthesis	31	21	67
10	Erythromycin biosynthesis	6	4	66
11	Peptidoglycan biosynthesis	17	11	64
12	Oxidative phosphorylation	13	8	61
13	Fatty acid biosynthesis (path 1)	14	8	57
14	Glutamate metabolism	35	20	57
15	Biotin metabolism	9	5	55
16	One carbon pool by folate	24	13	54
17	Selenoamino acid metabolism	22	12	54
18	Riboflavin metabolism	13	7	53
19	Glycolysis / Gluconeogenesis	40	21	52
20	Synthesis and degradation of ketone bodies	6	3	50

Figure 23. List of all metabolic pathways using KEGG.

Paralogous families

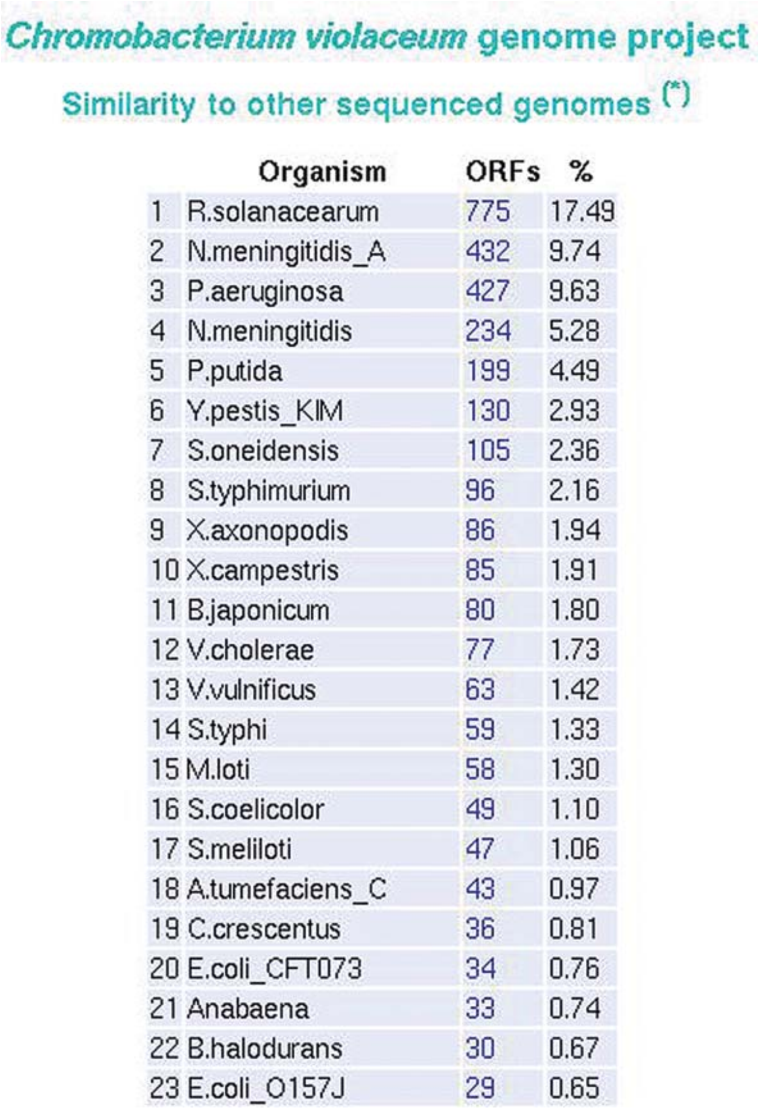
To find ORFs with a high degree of identity (paralogous) a blastp is executed among all ORFs, with a expected default value of E-05, a minimum identity percentage of 50%, and 60% query coverage. ORFs with the best hits are grouped.

Motifs in hypothetical and conserved hypothetical ORFs

InterPro motifs and COG-defined products arising from automatic annotation are recovered for hypothetical and conserved hypothetical ORFs. In case the definitions in these two blocks are similar, the annotator may review his annotation (Figure 25).

COG - clusters of orthologous groups of proteins

SABIA produces a report based on the ORFs functionally classified by the COG. A



Chromobacterium violaceum genome project

Similarity to other sequenced genomes (*)

	Organism	ORFs	%
1	R.solanacearum	775	17.49
2	N.meningitidis_A	432	9.74
3	P.aeruginosa	427	9.63
4	N.meningitidis	234	5.28
5	P.putida	199	4.49
6	Y.pestis_KIM	130	2.93
7	S.oneidensis	105	2.36
8	S.typhimurium	96	2.16
9	X.axonopodis	86	1.94
10	X.campestris	85	1.91
11	B.japonicum	80	1.80
12	V.cholerae	77	1.73
13	V.vulnificus	63	1.42
14	S.typhi	59	1.33
15	M.loti	58	1.30
16	S.coelicolor	49	1.10
17	S.meliloti	47	1.06
18	A.tumefaciens_C	43	0.97
19	C.crescentus	36	0.81
20	E.coli_CFT073	34	0.76
21	Anabaena	33	0.74
22	B.halodurans	30	0.67
23	E.coli_O157J	29	0.65

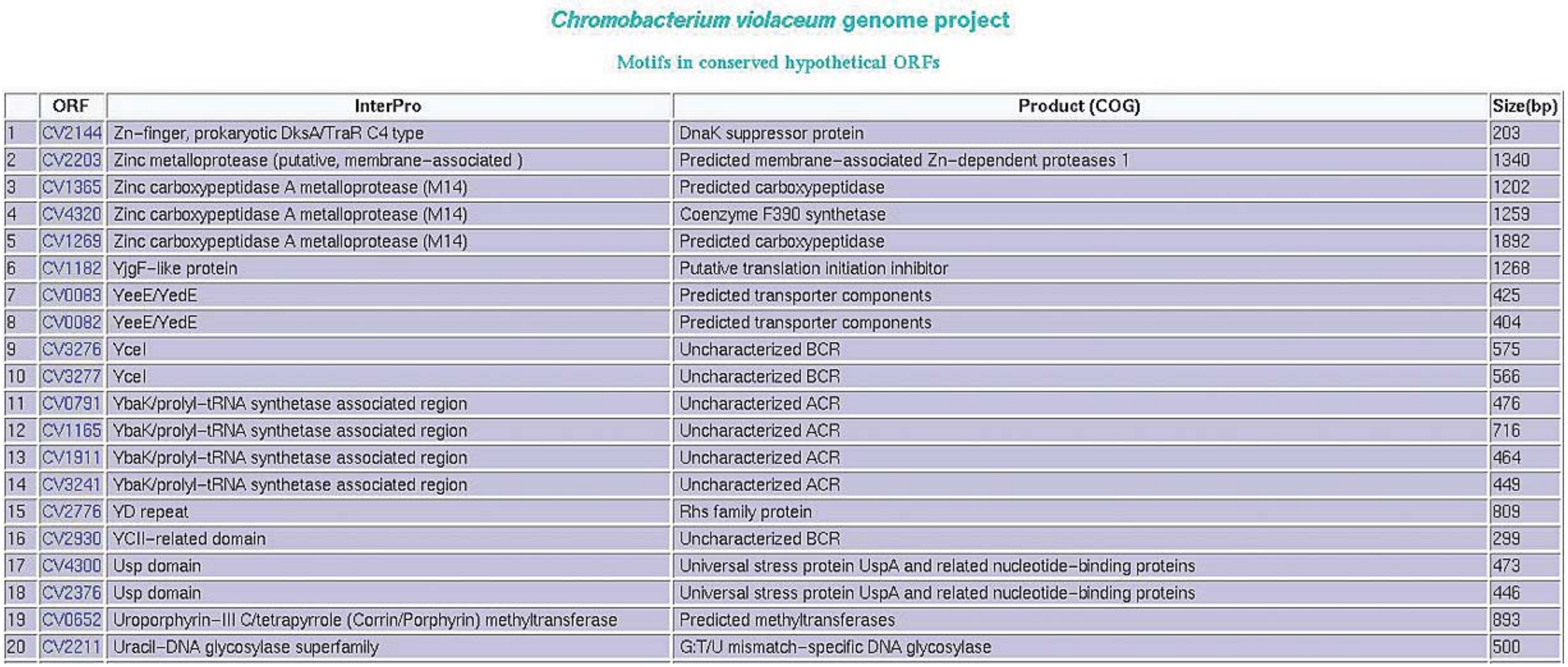
Figure 24. ORF distribution based upon KEGG hits.

general vision of the distribution and the percentage of total for classified ORFs is provided after the categorization of each ORF (Figure 26).

ORF table

The annotator may navigate selectively using the ORF list, ordered by their genome coordinates, containing their ID, gene names and products (Figure 27).

SABIA has been shown to be a useful tool for the management, assembly and annotation of genomes. The information made available daily on the home page allowed strategies to be adopted and decisions to be made in an efficient manner, during the course of the project.The software was able to extract the main information needed for the assembly and closure of the

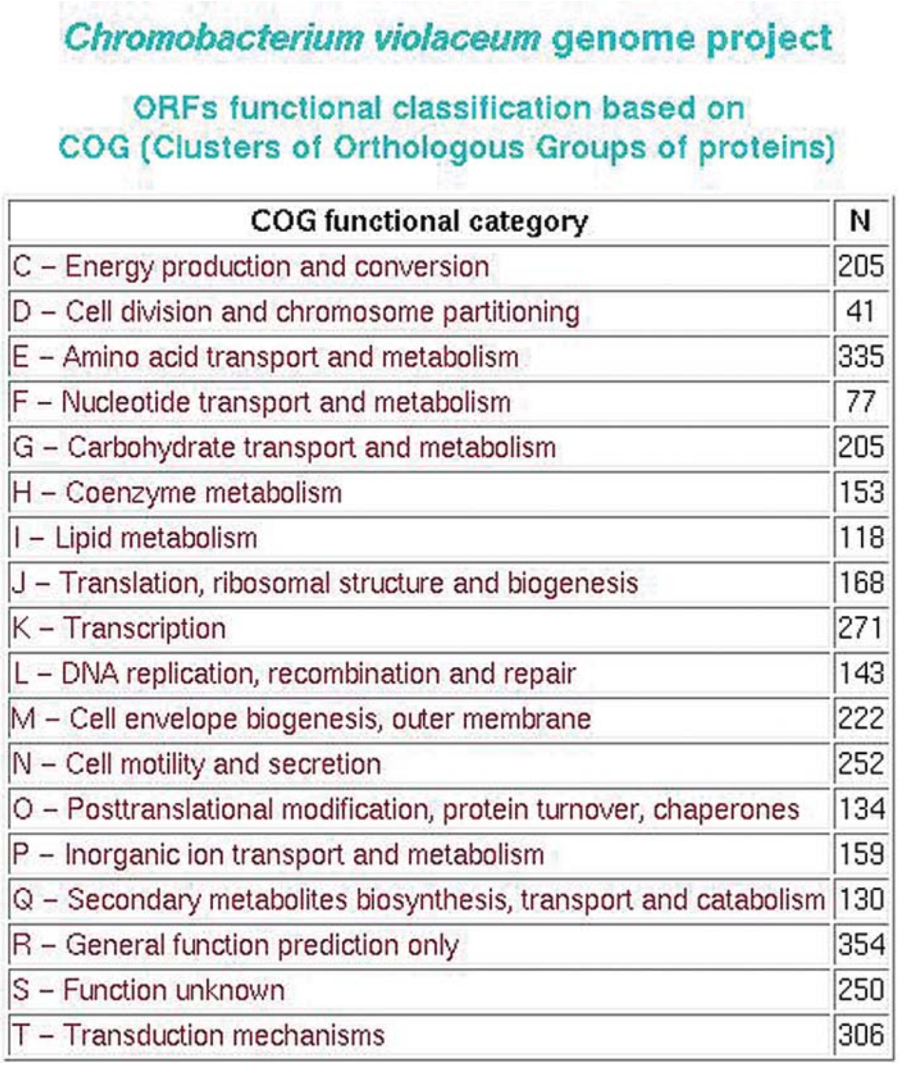


Chromobacterium violaceum genome project

Motifs in conserved hypothetical ORFs

	ORF	InterPro	Product (COG)	Size(bp)
1	CV2144	Zn–finger, prokaryotic DksA/TraR C4 type	DnaK suppressor protein	203
2	CV2203	Zinc metalloprotease (putative, membrane–associated)	Predicted membrane–associated Zn–dependent proteases 1	1340
3	CV1365	Zinc carboxypeptidase A metalloprotease (M14)	Predicted carboxypeptidase	1202
4	CV4320	Zinc carboxypeptidase A metalloprotease (M14)	Coenzyme F390 synthetase	1259
5	CV1269	Zinc carboxypeptidase A metalloprotease (M14)	Predicted carboxypeptidase	1892
6	CV1182	YjgF–like protein	Putative translation initiation inhibitor	1268
7	CV0083	YeeE/YedE	Predicted transporter components	425
8	CV0082	YeeE/YedE	Predicted transporter components	404
9	CV3276	YceI	Uncharacterized BCR	575
10	CV3277	YceI	Uncharacterized BCR	566
11	CV0791	YbaK/prolyl–tRNA synthetase associated region	Uncharacterized ACR	476
12	CV1165	YbaK/prolyl–tRNA synthetase associated region	Uncharacterized ACR	716
13	CV1911	YbaK/prolyl–tRNA synthetase associated region	Uncharacterized ACR	464
14	CV3241	YbaK/prolyl–tRNA synthetase associated region	Uncharacterized ACR	449
15	CV2776	YD repeat	Rhs family protein	809
16	CV2930	YCII–related domain	Uncharacterized BCR	299
17	CV4300	Usp domain	Universal stress protein UspA and related nucleotide–binding proteins	473
18	CV2376	Usp domain	Universal stress protein UspA and related nucleotide–binding proteins	446
19	CV0652	Uroporphyrin–III C/tetrapyrrole (Corrin/Porphyrin) methyltransferase	Predicted methyltransferases	893
20	CV2211	Uracil–DNA glycosylase superfamily	G:T/U mismatch–specific DNA glycosylase	500

Figure 25. List of conserved hypothetical ORFs and their InterPro and COG products.



Chromobacterium violaceum genome project

ORFs functional classification based on COG (Clusters of Orthologous Groups of proteins)

COG functional category	N
C – Energy production and conversion	205
D – Cell division and chromosome partitioning	41
E – Amino acid transport and metabolism	335
F – Nucleotide transport and metabolism	77
G – Carbohydrate transport and metabolism	205
H – Coenzyme metabolism	153
I – Lipid metabolism	118
J – Translation, ribosomal structure and biogenesis	168
K – Transcription	271
L – DNA replication, recombination and repair	143
M – Cell envelope biogenesis, outer membrane	222
N – Cell motility and secretion	252
O – Posttranslational modification, protein turnover, chaperones	134
P – Inorganic ion transport and metabolism	159
Q – Secondary metabolites biosynthesis, transport and catabolism	130
R – General function prediction only	354
S – Function unknown	250
T – Transduction mechanisms	306

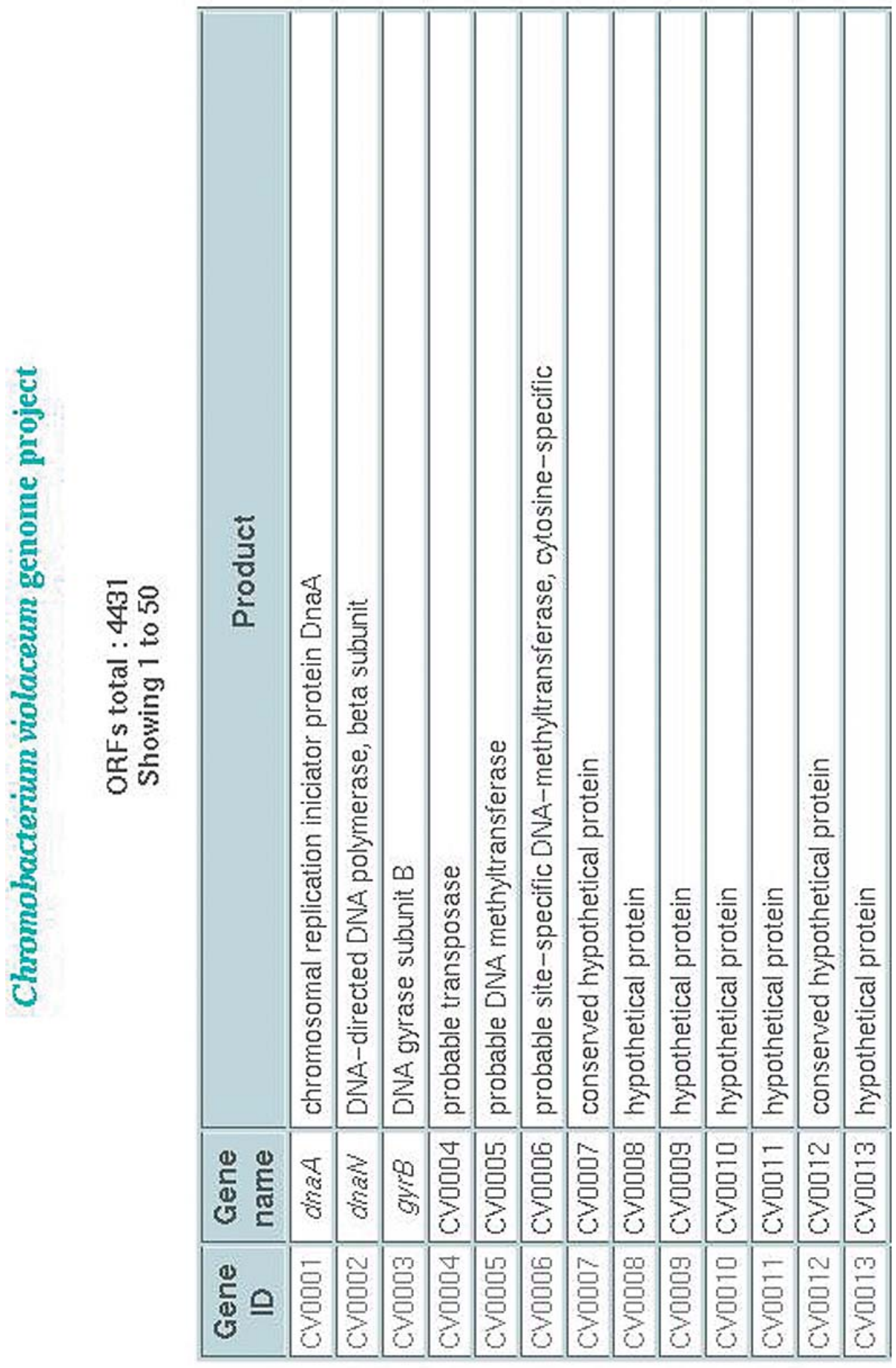
Figure 26. COG table.

genome from the various programs, making these tasks less difficult. For annotation, this tool was able to integrate information held in the best available database, and presented them to the users in an easy to use and gracefully intuitive format.

SABIA proved to be a flexible and easily extensible system. It is being currently used in other genome projects under our coordination. Future work using SABIA will serve to test ever more sophisticated annotation methods.

License

We distribute the complete system (including source code) to non-commercial users under an open source license, as a resource for the academic community. Special commercial licenses are available on request.



Chromobacterium violaceum genome project

ORFs total : 4431
Showing 1 to 50

Gene ID	Gene name	Product
CV0001	*dnaA*	chromosomal replication iniciator protein DnaA
CV0002	*dnaN*	DNA-directed DNA polymerase, beta subunit
CV0003	*gyrB*	DNA gyrase subunit B
CV0004	CV0004	probable transposase
CV0005	CV0005	probable DNA methyltransferase
CV0006	CV0006	probable site-specific DNA-methyltransferase, cytosine-specific
CV0007	CV0007	conserved hypothetical protein
CV0008	CV0008	hypothetical protein
CV0009	CV0009	hypothetical protein
CV0010	CV0010	hypothetical protein
CV0011	CV0011	hypothetical protein
CV0012	CV0012	conserved hypothetical protein
CV0013	CV0013	hypothetical protein

Figure 27. ORFs table.

ACKNOWLEDGMENTS

Research supported by the Ministério de Ciência e Tecnologia (MCT) through the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). We are indebted to L.N.C.C. for administrative and technical support.

REFERENCES

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.

Borodovsky, M. and **McIninch, J.** (1993). GeneMark: parallel gene recognition for both DNA strands. *Comput. Chem. 19*: 123-133.

Carraro, D.M., Camargo, A.A., Salim, A.C., Grivet, M., Vasconcelos, A.T. and **Simpson, A.J.** (2003). PCR-assisted contig extension; stepwise strategy for bacterial genome closure. *Biotechniques 34*: 626-628, 630-632.

Delcher, A.L., Harmon, D., Kasif, S., White, O. and **Salzberg , S.L.** (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res. 27*: 4636-4641.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M., McKenney, K., Sutton, G.G., FitzHugh, W., Fields, C.A., Gocayne, J.D., Scott, J.D., Shirley, R., Liu, L.I., Glodek, A., Kelley, J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O. and **Venter, J.C.** (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science 269*: 468-470.

Kanehisa, M. (1996). Toward pathway engineering: a new database of genetic and molecular pathways. *Sci. Technol. 59*: 34-38.

Lowe, T. and **Eddy, S.R.** (1997). tRNAscan-SE: a Program for Improved Detection of Transfer RNA genes in genomic sequence. *Nucleic Acids Res. 25*: 955-964.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J.A., Vaughan, R. and **Zdobnov, E.M.** (2003). The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res. 31*: 315-318.

Nakai, K. and **Kanehisa, M.** (1991). Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins: Struct. Funct. Genet. 11*: 95-110.

Riley, M. (1998). Genes and proteins of *Escherichia coli* K-12 (GenProtEC). *Nucleic Acids Res. 26*: 54.

Saier, M.H. (1999). Genome archeology leading to the characterization and classification of transport proteins. *Curr. Opin. Microbiol. 2*: 555-561.

Setubal, J. and **Werneck, R.** (2001). A program for building contig scaffolds in double-barreled shotgun genome sequencing. (www.lbi.ic.unicamp).

Simpson, A.J.G. (2001). Genome sequencing networks. *Nat. Rev. Genet. 2*: 79-83.

Tatusov, R.L., Koonin, E.V. and **Lipman, D.J.** (1997). A genomic perspective on protein families. *Science 278*: 631-637.