A System for Automated Bacterial (genome) Integrated Annotation—SABIA

Luiz G. P. Almeida¹, Roger Paixão¹, Rangel C. Souza¹, Gisele C. da Costa¹, Frank J. A. Barrientos¹, M. Trindade dos Santos¹, Darcy F. de Almeida² and Ana Tereza R. Vasconcelos¹,*

¹Laboratório Nacional de Computação Científica LNCC/MCT, Rua Getúlio Vargas 333, Quitandinha, Petrópolis, RJ, 25651-070, Brazil and ²UFRJ, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil

Received on November 5, 2003; revised on February 13, 2004; accepted on March 29, 2004
Advance Access publication April 15, 2004

ABSTRACT

Summary: A web-based software suite, SABIA (System for Automated Bacterial Integrated Annotation), is described that provides a comprehensive computational support for the assembly and annotation of whole bacterial genomes from the data derived from sequencing projects.

Availability: Both SABIA and supplementary materials are available at http://www.sabia.lncc.br

Contact: atrv@lncc.br

Supplementary information: http://www.sabia.lncc.br/what.htm

The bioinformatics component of whole genome sequencing projects involves the receipt, storage and retrieval of large amounts of data that are manipulated in a highly interactive manner by those responsible for sequence assembly and annotation. There is a considerable amount of software available for genome annotation, the most recently published of which are GENDB (Meyer et al., 2003), Manatee (from http://www.tigr.org) and Genquire (Wilkinson et al., 2002).

In contrast, there is a relative scarcity of tools that support genome sequencing and assembly routines. We have compiled a software suite, SABIA (System for Automated Bacterial Integrated Annotation), which directly connects the assembly process with annotation thus streamlining the overall process. This connectivity allows an early start to the annotation process, by improving the sequence quality during data acquisition, and facilitates the use of the annotation data in the assembly process.

This software comprises an automated web environment and a set of Perl/CGI scripts that manipulate data within a Relational Database (MySQL). The scripts generate HTML reports and forms through which users can access both biological information and load the results of their analysis onto the database. SABIA also executes daily routines such as the monitoring of shotgun sequencing, assembly, finishing and the updating of local copies of public databases. The software is divided into two modules, assembly and annotation (Supplementary information, Figure 1, at http://www.sabia.lncc.br/what.htm).

CONTIG ASSEMBLY PIPELINE

During the shotgun sequencing phase of a whole genome project, SABIA assembles the data using the phred/phrap/consed package (http://www.phrap.org). The process is monitored by the provision of reports on the numbers of reads, bases and contigs generated together with statistical information (see Supplementary material). Users can also access a report describing the features of all genomic libraries utilized.

Subsequently, the system uses a scaffold generating program (http://www.ic.unicamp.br) to place contigs in the correct relative order and orientation. This process is most effective when different DNA library types, such as those constructed using cosmids and plasmids, are used in the same project.

The most time-consuming task in the assembly process is finishing, which involves gap closure and improvement of sequence quality. Specific problems include bases with low consensus quality (LCQ), high quality discrepancies (HQD), bases not confirmed on both strands (NCBS) and actual gaps. The goal is to automate the finishing process as much as possible using the following steps:

(i) Identification of assembly problems and storage in the database.

(ii) User choice between solutions including re-sequencing, primer walking and sub-library construction.
(iii) Generation of new reports with the additional data informing whether the problem is solved or not, and repeat of Step ii if necessary.

In order to deal with repetitive sequences larger than the average read length, SABIA provides a gap closure scheme that involves filtering repetitive sequences, individual repeat reassembly and re-insertion of complete repeats at the correct position within the assembly. To facilitate this task, a graphical tool permits navigation along scaffolds highlighting anchored clones, reporting consensus quality and indicating possible assembly errors. In addition, reads with a phrap secondary score higher than a given threshold are marked, indicating that further investigation is required.

ANNOTATION PIPELINE

All possible open reading frames (ORFs) within the contigs generated during assembly are predicted by Glimmer (http://www.tigr.org/software/glimmer/) and GeneMark (http://opal.biology.gatech.edu/GeneMark/). In addition, tRNAs are detected by tRNAscan (http://opal.biology.gatech.edu/GeneMark/). Each ORF is submitted to several databases for comparison and the results are made available on the screen for the assessment of expert users. Identification of bona fide ORFs and their possible function takes into account the presence of associated transcriptional regulation sequences, such as ribosome-binding sites and promoters, the results of similarity searches using both nucleotide and amino acid sequences (BLAST), identification of protein motifs using InterPro (http://www.ebi.ac.uk/interpro/), functional classification by the KEGG (http://www.genome.ad.jp/kegg) and COG (http://www.ncbi.nlm.nih.gov/COG/) databases, protein localization analysis as assigned by PSORT (http://psort.nibb.ac.jp/), possible membrane transport capacity as determined using TCDB (http://tcdb.ucsd.edu/tcdb/) and Gene Ontology (GO, http://www.ebi.ac.uk/interpro/). ORFs that are identified using this annotation procedure are displayed on a graphical interactive map allowing the evaluation and, if judged necessary, adjustment of the start codon position. Additional available procedures include the identification of tRNA sequences and frameshifts. During the whole process, it is possible to access the metabolic pathways found in the KEGG database. This is useful for early detection of missing steps in the pathways under study and can be used to determine that additional sequencing may be necessary at an early stage of assembly. In addition the identification of orthologs, which can be highly informative during the annotation process, is undertaken using the Bidirectional Best Hit (BBH) approach (Overbeek et al., 1999). All BBHs between genome pairs are identified and clustered by means of a greedy algorithm.

The database design and Perl scripts in SABIA are a simple way of implementing useful strategies for genome assembly and annotation. SABIA permits essentially one step de novo establishment of a comprehensive bioinformatics pipeline for whole bacterial genome sequencing projects. It was entirely developed during the assembly and annotation of the Chromobacterium violaceum genome, undertaken by the Brazilian National Genome Network (www.brgene.lncc.br/cviolaceum), and is now routinely used in our laboratories. The annotation and the assembly modules may be run independently and require the prior installation of MySql, Apache and Perl. SABIA was developed for both Linux and Solaris Platforms, and is distributed under licence. SABIA is an extremely colorful bird widely distributed in Brazil.

ACKNOWLEDGEMENTS

We thank Andrew Simpson for his help with the final preparation of this manuscript. This work was supported by the Brazilian Ministério de Ciência e Tecnologia (MCT) through the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERENCES

