Fact and Task Oriented System for Genome Assembly and Annotation

L.A. Digiampietri et al.

Instituto de Computação, Universidade Estadual de Campinas,
CP 6176, Campinas, SP 13084-971 BRAZIL
luciano@ic.unicamp.br

Abstract. We present a preliminary description and results of a system to help the curation of genome assembly and annotation. Standard tools are used for these tasks, and our methodology focuses on user guidance, data visualization and integration, and data browsing aspects.

1 Introduction

The usual concern of most of activities, tools and infrastructure related to genomic analyses is with computer systems functionality. Many systems are developed in an *ad hoc* way following only functional requirements. This development methodology pays little attention to characteristics like user interface and usability. We have developed a simple methodology to make the user-interaction part of genome assembly and annotation more user-friendly and therefore more effective. Based on this methodology we have implemented a web-based prototype. This prototype is being used as the main tool for the assembly and annotation of the *Xanthomonas axonopodis pv aurantifolii* strains *B* and *C* genomes at LBI [4] with the support of USP [1] and UNESP [5].

2 System Development Methodology

The system presented here was developed following a generic methodology specified by us at LBI. This methodology allows the development of any computational infrastructure which requires a flow of activities and that provides data mining and visualization mechanisms. This methodology has the following phases: (i) identification and description of tasks to be done; (ii) description of facts to be considered; (iii) development of fact analysis and visualization tools; (iv) development of examples or tutorials on how to execute each task; (v) development of tools for accomplishing the tasks. We have applied this methodology to improve a genome assembly and annotation system used at our laboratory.

Facts are characteristics observed in the set of available data. Facts are the basis for all the analysis and conclusions which will be made during assembly and annotation. *Tasks* are actions which must be executed (automatically or

J.C. Setubal and S. Verjovski-Almeida (Eds.): BSB 2005, LNBI 3594, pp. 238–241, 2005.

Table 1. Assembly tasks and facts

Task	Facts
Contigs management	set of reads, phrap and genscaff results
Links management	reads from the same insert found on different contigs
Contigs projection on the reference genome	alignment between the reference and the target genomes
Supercontigs management	contigs, links, gap closures and alignments to the reference genome
Management of inserts to be subcloned and sequenced	reads and links information

manually) with the objective of getting closer to the desired solution. For example, a set of facts can be observed in the result of the phrap [3] assembly and postprocessing by the genscaff program [6], such as a possible link between contig x and contig y. A task must obtain conclusions about the facts, for instance, to conclude whether contigs x and y are adjacent or not. For each kind of fact, data analyses and visualization tools were developed to ease the understanding and the making of a decision. Some examples of genome assembly tasks are: contigs management, links management, selection of clones to be subcloned and sequenced, comparison between the target and the reference genomes and *supercontig* management (supercontig is a set of linked contigs).

Figure 1 shows some of the graphical results of our tools (showing contig, supercontig, link and projection with reference genome information). All figures are automatically generated and have hyperlinks to allow easy data browsing.

One of the most complex tasks during genome assembly is to decide whether two contigs are linked or not. Our system used the following facts to help in decision making: (1) links between those contigs; (2) conservation of the order regarding the reference genome (based on alignment against a reference genome);

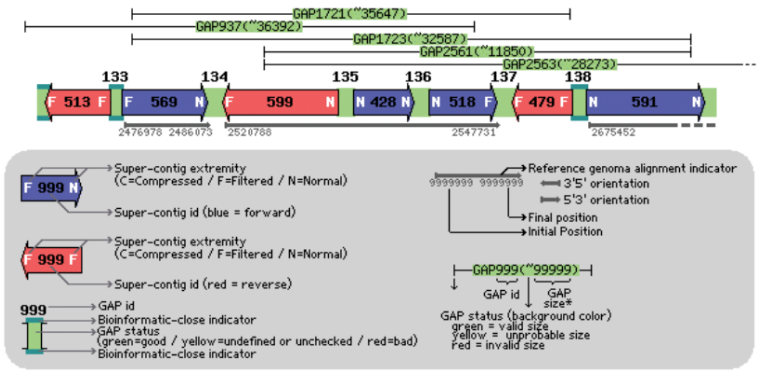

Fig. 1. Supercontig information: contigs, links, gaps and projection over the reference genome information

and (3) bionformatics gap closure (a sub-assembly using only reads in a particular region that successfully closes a gap). By integrating these facts, our system facilitates the curatorial part of the genome assembly process, decreasing the need for new sequencing.

3 Results, Conclusions and Future Work

Complex information systems that require intense user interaction deserve special care on user-related issues, such as usability and interface. Large-volume data processes, such as genome assembly and annotation, require special care on data presentation, through graphic visualizations, data summaries and data integration. We have briefly described a simple methodology that helped us create a web-based system that allowed us to achieve good results in a genome assembly process. The detailed description of each one of the tasks and facts, as well as the specification of tutorials or examples for each task, makes possible a more conscious, easy and systematic use of the system.

The system proposed is being used on the *Xanthomonas axonopodis pv aurantifolii* strains *B* and *C* genomes assembly and annotation. Before the work described in this paper, these two genomes were being assembled using a traditional system, which had no specific computational help for assembly curators. The use of our system showed quantitative and qualitative gains with respect to previous assembly results. The main gains were: (i) all data is integrated in a database management system (DBMS), making it possible to make efficient queries to every object involved in the project; (ii) low training cost of new assembly and annotation team, due to tutorials developed for the execution of each task; (iii) greater assembly efficiency through a better use of data. The most important practical conclusion of this case study was the reduction on the number of supercontigs without the need of new sequencing, causing greater genome coverage. Table 2 compares the results obtained by our system to the ones available before we put our system to use. This table shows that thorough our system we obtained better results on every analyzed characteristic, refining the assembly and being more efficient on the use of available data.

As a future step we intend to package tools (making them more generic and reusable), and extending the system for dealing also with comparative genomics.

Table 2. Comparison between previous results and results from our approach

Data	**Previous Results**	**Our results**	**Situation**
Number of supercontigs	45	35	Improved
Total number of contigs in the supercontigs	225	234	Improved
Average number of contigs by supercontig	5	6.69	Improved
Number of base pairs on supercontigs	4934046	5105624	Improved
Valid links on supercontigs	180	199	Improved
Number of new closed gaps	87	91	Improved

Another future work is the development and usage of ontology to publish data on the Web through XML [2], increasing interoperability.

More detailed descriptions and tools can be obtained through e-mail contact with LBI: lbi@ic.unicamp.br.

Acknowledgements. The work described in this paper was partially financed by CAPES, Fundecitrus and MCT-PRONEX SAI project. The Laboratory of Bioinformatics, in which most of this work was done, is partly funded by FAPESP. We acknowledge the others researchers that had worked in the sequencing and assembling of the genomes presented here.

References

1. Departamento de Bioquímica, Instituto de Química, University of Sao Paulo. http://www2.iq.usp.br/bioquimica/
2. Extensible Markup Language (XML) 1.0 (Third Edition) (2004). http://www.w3.org/TR/2004/REC-xml-20040204
3. Green P. Phrap: phragment assembly program. http://www.phrap.org.
4. Laboratory for Bioinformatics (LBI), Institute of Computing, University of Campinas. http://www.lbi.ic.unicamp.br
5. Laboratório de Bioquímica e Biologia Molecular (LBM), UNESP. http://www.lbm.fcav.unesp.br/
6. Setubal, J. and Werneck, R.: A program for building contig scaffolds in double-barreled shotgun genome sequencing. *Technical Report* IC-01-05, Institute of Computing, Unicamp, 2001.