CLEAR: A Cluster Editor for Phylogenetic Purposes

Carlos Norberto Fischer[1]
Celso Shiguemi Takimoto[1]
Ivan Rizzo Guilherme[1]
Maurício Bacci Júnior[2]
[1]Departamento de Estatística, Matemática Aplicada e Computação - IGCE/UNESP
[2]Centro de Estudos de Insetos Sociais - IB/UNESP
Av. 24-A, 1515, 13506-700, Rio Claro, SP.

ABSTRACT

Laboratories working on Molecular Phylogeny develop a work routine that involves the following basic tasks: collection of different organisms, obtaining of molecular data (through DNA sequencing), alignment of sequenced fragments [5], and analysis of results [2]; they are addressed to the study of phylogeny and evolution of the molecules of interest. Most of these tasks are done using computation programs. Thus, it is possible to automate great part of the working routine, so that the molecular phylogeny work would become faster and track-controlled. Two tasks that can be included in an automated process are *Alignment of Sequenced Fragments*, that uses algorithms of multiple alignments, and the *Analysis Task* [1].

The fragment alignment, which is carried out by specialized software, consists in combining, for example, sequenced fragments of DNA, called *Reads*, aiming at the reconstruction of either the complete sequence of the molecule or long fragments. In this process, reads are combined to generate *Contigs*. Contigs related to the same sequencing are then considered grouped in what is treated as *Cluster of Aligned Sequences*.

However, contigs may contain reads with low quality regions, i.e. regions containing errors that occurred either during the sequencing or the automated base calling phases. Therefore, it is important that each formed cluster can be analyzed in order to correct sequencing mistakes.

Whether possible faults in the generation of contigs are noticed during the analysis, it is essential that who makes the analysis also makes the corrections. One way to modify a contig is through the direct change of nucleotide bases in some reads. Another solution is the exclusion of bad reads or read regions. In both cases, if a read is modified or simply excluded, it is necessary to redo the alignment process for generating different contigs and then analyze them again. However, this process is very laborious and should be automated.

In this work, we present a *Cluster Editor software (CLEAR)* that allows the visualization, analysis and edition of a cluster of contigs. During the edition, it is possible to delete one or more reads used to generate a contig, in order to eliminate possible failures of alignment. Whether a read exclusion is performed, the software automatically redoes the alignment process. All the modifications are recorded in successive editions, a process that allows the recovering of the earlier stages of work.

CLEAR also permits that an individual contig is exported to be analyzed in the most commonly used sequence editor, like BioEdit (running in the client machine), in order to make modifications in selected regions of a given read. After the edition, the modified reads are transferred back to the server computer in order to be included in a new alignment process.

CLEAR works incorporated to a Bioinformatics Environment [4] that integrates several automated tools, such as chromatogram reception (via Web) and conversion to organic bases [3], read trimming, and alignment of trimmed read [5]. The environment is currently implemented at the Laboratório de Bioinformática at Universidade Estadual Paulista - UNESP, Rio Claro, SP (http://lbi.rc.unesp.br/), and it has been used for researchers working at the Laboratório de Evolução Molecular of the Centro de Estudos de Insetos Sociais, UNESP, Rio Claro.

REFERENCES

[1]Adams, M.D., Kelley, J.M., Gocayne, J.D., Dudnick, M., Polymeropoulod, M.H., Xiao, H., Merril,C.R., Wu, A., Olde, B.,Moreno, R., Kerlavage, A.R., McCombie, W.R., Venter, J.C."Complementary DNA Sequencing: Expressed Sequence Tags and the Human Genome Project". Science, n.252, p.1651-1656, 1991.

[2]Andrade, M.A. "Automated Genome Sequence Analysis and Annotation". Bioinformatics, n.15, p.391-412, 1999.

[3]Ewing, B., Hillier, L., Wendl, M.C., Green, P. "Base-Calling of Automated Sequencer Traces Using Phred I. Accuracy Assessment". Genome Research, n.8, p.175-185, 1998.

[4]Fischer, C.N.; Guilherme, I.R.; Bacci Jr., M. – "Uma Arquitetura de Bioinformática para Automação de Laboratórios de Evolução Molecular". Anais (em CD-ROM) do I2TS'2002 – International Information Technology Symposium, Florianópolis, SC, Brasil, 2002.

[5]Huang, X., Madan, A. "CAP3: A DNA Sequence Assembly Program". Genome Research, n.9, p.868-877.