Bioinformatics Tools for Assembling and Analysis of Chloroplast Genomes

Jesus Mena-Chalco[1], Henrique S. Alves[2], Helaine Carrer[2], Roberto M. Cesar-Jr[1]

[1] Depto. Ciência da Computação, IME-USP. Rua do Matão, 1010. São Paulo-SP 05508-090.

[2] Depto. Ciências Biológicas, ESALQ-USP. Av. Pádua Dias, 11. Piracicaba-SP 13418-900.

Chloroplasts are organelles found only in plant and algae cells. They are responsible for photosynthesis and for the synthesis of key molecules required for the basic architeture and functioning of plant cells. These organelles have their own genetic machinery and together with the nucleus and mitochondrial genomes are responsible for celular coordenation activity. At the moment 29 higher plant plastid genomes (plastomes) have been sequenced (http://ncbi.nlm.nih.gov/). The plastome sequences are conserved among species but the genes arrangements are different for divergent plant groups. The knowledge of the nucleotide sequence of chloroplast genomes is important for evolution studies and for biotechnology applications. The chloroplast organelle being used as a model in this study was isolated from *Eucalyptus grandis*, an important economical tree for the production of paper and cellulose and in Brazil is located the main germoplasm collection of *Eucalyptus* outside Australia.

We have sequenced 3500 sequences from an *Eucalyptus* DNA library. These sequences represent so far, 50% of the total plastome sequence of *Eucalyptus grandis*. These sequences are stored through a special pipeline at the bioinformatics servers at URL http://malariadb.ime.usp.br:8026/pipeline/. Once this phase is accomplished, the next step is the search for similar sequences in other related organisms. Some tentative results towards this direction have been already obtained.

In this study, we apply digital signal processing (DSP) techniques [1, 2, 3] on the genomic data sequences in order to identify and compare DNA and protein sequences of *Eucalyptus grandis* to the other available higher plant plastomes. We have chosen different approaches to identify protein coding DNA regions and to compare protein sequences. In particular, traditional Fourier analysis and the wavelet transform will be evaluated [4, 5].

REFERENCES

[1] Jie Chen, Huai Li, Kaihua Sun and Bill Kim, "How will bioinformatics impact signal processing research", *IEEE Signal Processing Magazine*, November 2003.

[2] Xin-Yu Zhang et al. "Signal processing techniques in genomic engineering", *Proceedings of the IEEE*, Vol 90. Nro 12, December 2002.

[3] Dimitris Anasstassiou, "Genomic signal processing", *IEEE Signal Proc. Mag*., pp. 8-20, July 2001.

[4] Pietro Lio, "Wavelets in bioinformatics and computacional biology: state of art and perspectives". *Bioinformatics Review*, 19(1), 2003.

[5] Chafia H. Trad, Qiang Fang and Irena Cosic. "Protein sequence comparison based on the wavelet transform approach", *Protein engineering*, 15(3):193-203, 2002.