

# Motif discovery in conserved regions of chloroplast genomes

Beatrice Kilel

School of Computational Sciences and Informatics, George Mason University, 4400 University Dr., MSN 5C3, Fairfax, VA. 22030. USA. Email: [bkilel@gmu.edu](mailto:bkilel@gmu.edu)

## 1 Introduction

As more chloroplast genomes are fully sequenced, it is becoming increasingly easy to identify genes believed to be co-regulated through comparative genomics, and then look for patterns in their regulatory regions [5]. Some of the important genes in these genomes are involved in photosynthesis, metabolism, transport, transcription/translation, and protein kinases or phosphatases. Ribosomal and transfer RNAs for instance are important genes and play an important role in translation of all the mRNA to proteins using the universal genetic code. The objective of this study is to cluster ribosomal RNA and protein kinase sequences into different families and find conserved motifs from sequence families using different informatics tools. Results obtained indicated that based on sequence similarity, ribosomal RNAs and phosphatases can be individually clustered and hence specificity in function could also be deduced. These were evidenced by some new conserved sites found.

## 2 Method and Materials

Motif search was performed on two ribosomal proteins using MEME software [9] and a comparison of results done with that from PROSITE. A perl script was used to extract annotated sequences from the GenBank [4]. Blast [2, 10] sequence alignment tool was used to do pairwise comparisons and ClustalX [8] for multiple sequence alignment. GeneRage [3] was used to perform sequence clustering. WebLogo was used to find the sequence logos [7] in the two ribosomal proteins.

## 3 Results and Discussion

*Rps8*, a cytosolic ribosomal protein [1] seem to have been involved in a nuclear/cytosolic origin in a common ancestor of angiosperms and gymnosperms (results not shown). This gene is absent in *Medicago sativa* and a *Toxoplasma gondii* (false chloroplast) which implies that loss of the gene has occurred across different plant species. Most of the motifs discovered in several chloroplast genes were found in the same promoter region (results not shown) and are conserved for the *rps8* gene (Fig.1). A similar pattern is seen in *rpl14* (results not shown). These motifs can be used to perform molecules as well any cross structural comparisons The type of proteins lost in *Epifagus* and *Toxoplasma*, as well as *Medicago* genomes clearly shows that these organisms have non photosynthetic plastids and that they also exhibit a much reduced ability for independent translation. Results obtained from sequence logos generated for *rps8* (Fig. 2) and *rpl14* support these findings.

1 50 PGLRIYSNYQEIPKVLGGMGIAILSTSQGIMTDREARQEGIGGEILCYIW  
 2 50 DTIADMLTSIRNADMAKHGTVQVPATNITENIVQILLQEGFIENVREHQE  
 3 21 YFLVLTLRHQRNKKRPYITTL

Fig.1. Motifs with the best possible match for *rps 8* in promoter regions of 38 chloroplast genes with an E value less than 10.

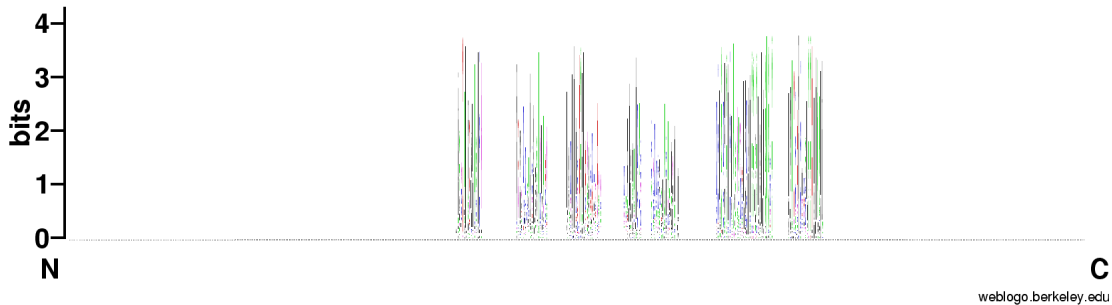


Fig. 2. Sequence logo for *rps8* gene. The highest peaks indicate sequence conservation at that position and the symbol at that is the relative frequency of each amino or nucleic acid at that position.

#### 4 References

- [1] Adams, K.L, Daley, D.O, Whelan, J. and Palmer J.D. 2002 Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts. *Plant Cell* 14: 931-943.
- [2] 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.
- [3] Anton J., Enright, O. and Ouzounis, C.A. 2000. GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics* 2000 16: 451-457.
- [4] Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., and Ouellette, B.F. 1998. GenBank. *Nucleic Acids Research* 26:1-7.
- [5] Dean, C. and Schmidt, R. 1995. Plant genomes: a current molecular description. *Ann. Rev. Plant Physiol. Plant Molec.* 46: 395-418.
- [6] Gorodkin, J., Heyer, L.J, Brunak, S., Stormo, G.D. 1997. Displaying the information contents of structural RNA alignments: the structure logos. *Comput. Appl. Biosci.*, Vol. 13, no. 6 pp 583-586.
- [7] Schneider TD, Stephens RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.* 18: 6097-6100.
- [8] Thompson, J. D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25: 4876-82.
- [9] Timothy L.B. and Michael, G. 1998. Combining evidence using p-values: application to sequence homology searches, *Bioinformatics*, 14(48-54).
- [10] Zhang, J., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.