Advanced Journal of Microbiology Research Vol. 2004
Available online at http://internationalscholarsjournals.org/journal/ajmr

Full Length Research Paper

Comparative analysis and relationships of six important crop species chloroplast genomes using whole genome web-based informatics tools

Beatrice Kilel

School of Computational Sciences and Informatics, George Mason University, Fairfax, VA. 20110. USA. Email: bkilel@gmu.edu.

Accepted 25 March 2004

Using informatics tools to compare important species is now feasible as structural genomics continue in importance and establishment of structure-function relationships become a common way of comparative analysis. Currently, many of the technical issues involved in sequencing complete genomes have been solved. The smallness in size of chloroplast genomes facilitates being used for the discovery of disease resistance genes, introgression of important traits in transgenic plants, quantitative trait analysis and phylogenetic studies. Knowledge from this can be extrapolated to important crops like sorghum, millet, taro, and cassava that have not been fully sequenced. This study compared six important crop species using GeneOrder3.0 and CoreGenes2.0 web-based informatics tools using complete chloroplast genomes. Results obtained depict cases of major genome rearrangements, translocation, duplication, inversion and deletion of genes. Members of the poaceae family indicate a close relationship in the nature of conserved sequences while *Oryza sativa* and *Chlorella vulgaris,* which are not members of poaceae indicate no synteny. Gene content indicates that there are common sets of putative orthologs across the different species. *Zea mays*, *O. sativa*, *Nicotiana tabacum*, *Spinacea oleracea*, *Triticum aestivum* had 71 rows of putative orthologs (355 total) with one hypothetical protein (GI:11465969) in *N. tabacum*, which is homologous to cemA and ORF230 protein in *O. sativa* and *Z. mays*, respectively. There was a clear indication from these sets of putative orthologs that maturase-encoding genes were found only in the terrestrial plants and not in the unicellular organisms.

Key words: Chloroplast genomes, comparative analysis, informatics tools, whole genome, synteny.

INTRODUCTION

Inference of relationships from proteins of known function to proteins of unknown function that are structurally similar can be accomplished through comparative analysis. Plant species that have not been fully sequenced can be compared on whole genome level using chloroplast genomes. This is an important aspect in the quest to decipher more the plant characteristics for ensured food security in the developing economies (Herdt, 1998; Kishore and Shewmaker, 1999). Chloroplasts are multifunctional plant organelles that are used for critical functions such as photosynthesis, starch synthesis, nitrogen metabolism, sulfate reduction, fatty acid synthesis, DNA, and RNA synthesis (Zeltz et al., 1993). These organelles are generally small in size that ranges from about 120-220 kb with 120-150 genes (Stoebe et al., 1998) that are both unique and irreplaceable and involved in the energetic processes in plants (Maier et al., 1995). Crop species under this study are found in different shapes and sizes and they can be classified into distinct groups depending on their habit with the important families being poaceae and fabaceae.

The conservation in gene order is an informative measure, which may provide information about gene function and interactions of proteins that are encoded by these genes (Overbeek et al., 1999; Huynen et al., 2000). Tamames et al., 1997 indicated that gene order is conserved and well preserved at phylogenetic distances for closely related plant species. However, this conservation in gene order is low or lost during evolution in the distantly related species (Huynen and Bork, 1998).

Table 1. Complete chloroplast genomes of selected plant species.

Botanical Name	Common Name	Accession #	Bp	Genes	Family	Phylum	Kingdom
L. japonicus	Japanese lotus	NC_002694	150519	74	Fabaceae	Anthophyta	Plantae
N. tabacum	Tobacco	NC_001879	155939	107	Solanaceae	Anthophyta	Plantae
O. sativa	Rice	NC_001320	134525	92	Poaceae	Anthophyta	Plantae
S. oleracea	Spinach	NC_002202	150725	77	Chenopodiaceae	Anthophyta	Plantae
T. aestivum	Wheat	NC_002762	134545	75	Poaceae	Anthophyta	Plantae
Z. mays	Maize	NC_001666	140384	123	Poaceae	Anthophyta	Plantae

A recent divergence of the species with the gene order still intact, clustering of genes for the cell integrity and presence of lateral gene transfer (LGT) as a block of genes are some of the reasons for this closeness in gene order conservation. These conserved features are exhibited as orthologs or paralogs, which may also help to explain the phylogenetic distance of the species considered. Snel et al. (1999) have indicated that using gene order is a better comparative measure of phylogeny since it is not influenced by the presence of any particular sets of genes in individual chloroplast genomes. Correlated presence or absence of genes constitutes phylogenetic profiles, which is an important aspect in predicting functions in individual genomes and across genomes (Kilel et al., 2004). Gene content (the ratio between the numbers of orthologs between the two species compared and the maximum number of possible orthologs) may not be as conserved as gene order as reported by Wakasugi et al. (2001). It is therefore of essence to compare global features from gene order, gene content and phylogeny studies for a more conclusive comparative analysis between and among these chloroplast genomes. Results of this study will serve as a basis in the analysis of important crop species characteristics so as to come up with tangible solutions in disease control, herbicide resistance, and introgression of novel genes into transgenic crops.

MATERIALS AND METHODS

Complete chloroplast genomes for *Oryza sativa*, *Triticum aestivum*, *Nicotiana tabacum*, *Zea mays*, *Spinacea oleracea* , and *Lotus japonicus* were obtained from the GenBank (Benson et al., 1998) (Table 1). GeneOrder3.0 (Mazumder et al., 2001; Zafar et al., 2001; Kundeti et al., 2003) and CoreGenes2.0 (Zafar et al., 2002) interactive web-based informatics tools were used to do comparative analyses of these genomes. GeneOrder3.0 output on Microsoft Excel shows each point on the graph generated as a coding gene. Any linear arrangements suggest presence of synteny or identity, which is resultant from groups of orthologous or paralogous sequences. CoreGenes2.0 was used to identify related sequences through the core set of genes, cataloguing them and classifying the species with shared conserved sequences (Mazumder et al., 2001). Output from these core genes are subjected to PSI-BLAST or Clustal analyses (Thompson et al., 1994). A phylogenetic tree was constructed using WINCLADA (Nixon, 1999a,b) and NONA (Goloboff, 1994) to see how these organelles are evolutionarily divergent and also to use as a guide on the gene order pairwise combinations.

RESULTS AND DISCUSSION

Results indicate cases of major genome rearrangements, translocation, duplication, inversion and deletion of genes. Members of the Poaceae family indicate a close relationship in the nature of conserved sequences while *O. sativa* and *Chlorella vulgaris* (divergent comparative analysis) indicate no conserved gene order between them (Figure 1). Gene content indicates that there are common sets of putative orthologs across the different species. *Z. mays*, *O. sativa*, *N. tabacum*, *S. oleracea*, *T. aestivum* had 71 rows of putative orthologs (355 total) with one hypothetical protein (GI:11465969) in *N. tabacum*, which is homologous to cemA - a heme binding protein and ORF230 protein in *O. sativa* and *Z. mays* respectively. There was a clear indication from these sets of putative orthologs that maturase - encoding genes that are implicated in CO_2 transport were found only in the terrestrial plants and not in the unicellular organisms. Figure 1 indicates that *O. sativa* and *Z. mays, T. aestivum* and *N. tabacum* seem to be have undergone some genomic rearrangement through inversion with some gene deletions occurring among all of them. This is a good indication of the close relationship between these members of poaceae. A very different relationship is seen with *O. sativa* and *C.vulgaris* where there is no indication of a conserved gene order between the two. Figure 2 is a comparison of species from different families but share a common nature in the conservation of syntenic regions. This may provide insight in the extrapolation of information across the species divide. Figure 3 shows that *T. aestivum* and *Z. mays* seem to have undergone some major genomic rearrangements through translocation and inversion whereas *Arabidopsis thaliana* – a model plant and *N. tabacum* share similarities in gene conservation. This may provide information on novel genes in the other members of solanaceae that have not been fully sequenced even though they may be in different clades (Figure 4).

Other comparisons using *Pinus koraiensis*, *Pinus thunbergii* [Pinaceae] and *T. aestivum* [Poaceae] (results not shown), indicate that the information from species can be extrapolated across the species divide. Also gene

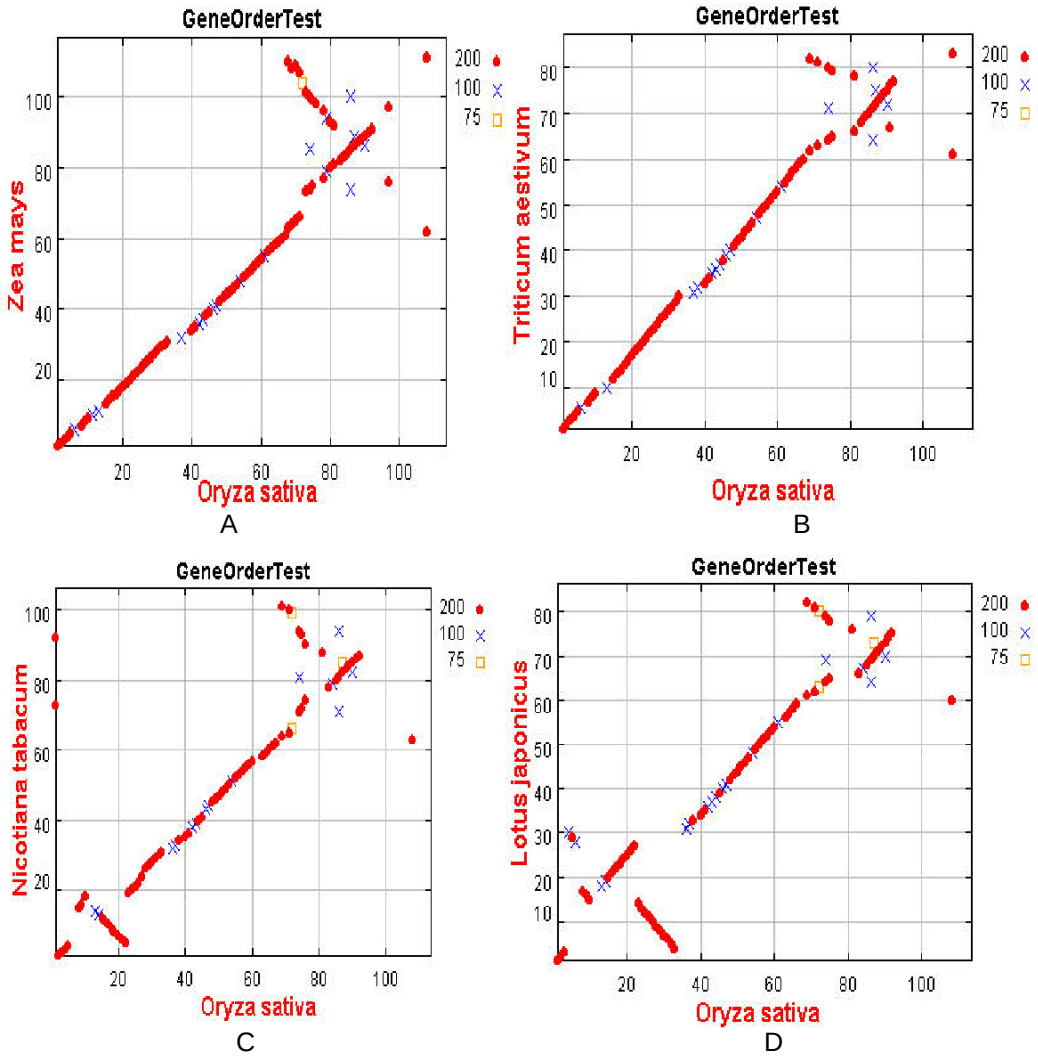

Figure 1. The plots generated using GeneOrder2.0 for pairwise genomes comparison. A diagonal dot indicates synteny and lack of co-linearity is indicated as dots away from the diagonal. A red dot indicates a coding gene as a point and any co-linear arrangements suggest synteny between the compared genomes. The X and the Y axis denote the gene numbers of the reference and query organisms. BlastP score parameters are ranged from 75 to 200. (A) Gene order between *O. sativa* and *Z. mays*. (B) Gene order between *O. sativa* and *T. aestivum*. (C) Gene order between *O. sativa* and *N. tabacum*. (D) Gene order between *O.sativa* and *C. vulgaris* (for a divergent comparative analysis).

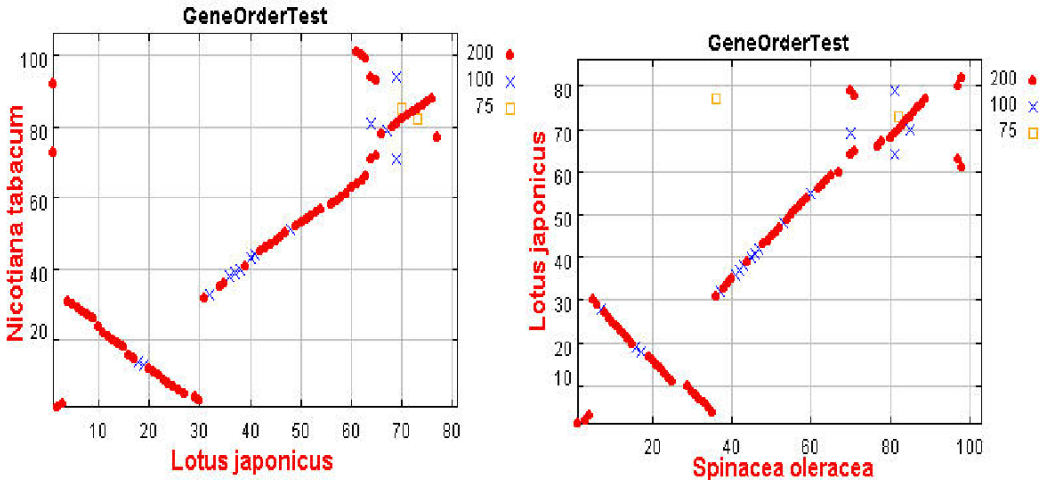

Figure 2. Plots generated for gene order in *N. tabacum* versus *L. japonicus* and *S. oleracea*.

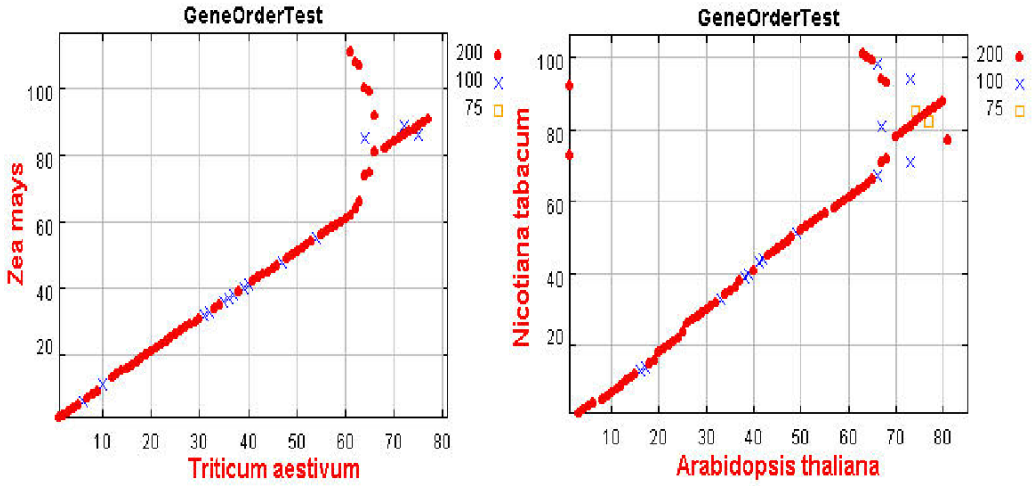

Figure 3. Plot generated for gene order in *T. aestivum* versus *Z. mays and A. thaliana* versus *N. tabacum.*

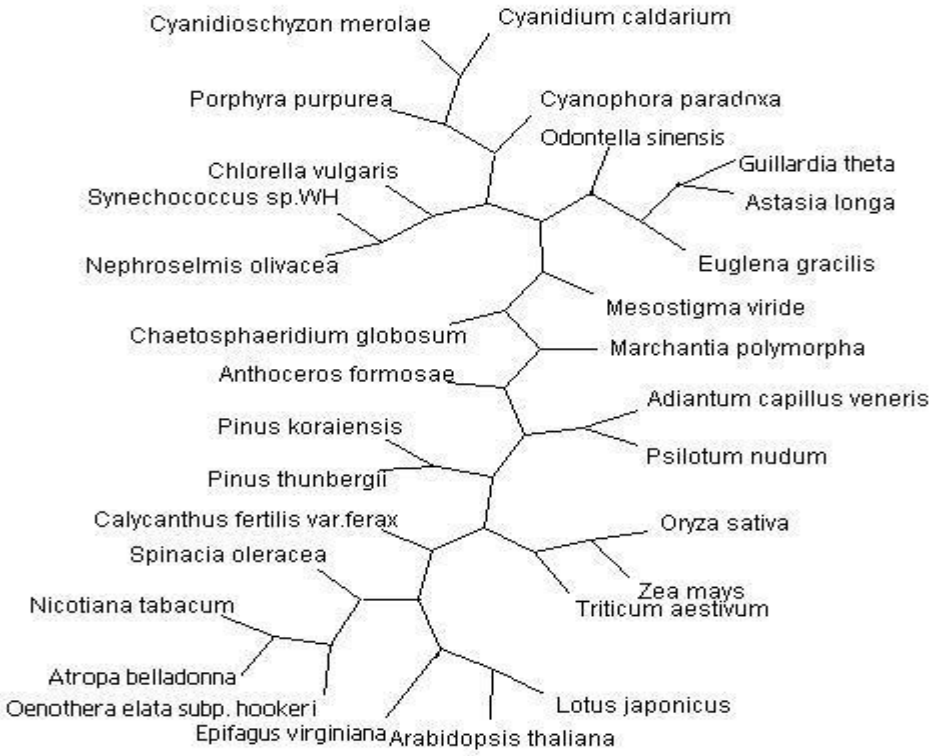

Figure 4. Network tree of common genes in complete chloroplast genomes. The poales are closely placed in a common clade while unrelated species such as *C. vulgaris* is in a clade of green algae. Generating a good tree is important in predicting gene function based on evolutionary distance.

order between *Epifagus virginiana* and *Calycanthus fertilis* var. ferax (results not shown), seem to have had a case of gene duplication and a case of paralogy.

CoreGenes2.0 is primarily used to find putative orthologs in the two to five compared species, which is further subjected to PSI-BLAST or Clustal analyses. The genomes were divided into the following groups. Group #1 *Z. mays*, *O. sativa*, *N. tabacum*, *S. oleracea*, *T. aestivum*; Group #2 *L. japonicus*, *S. oleracea*, *N. tabacum*, and *Z. mays*. Parsimony jackknife scores (phylogenetic tree not shown) indicate *T. aestivum*, *O. sativa* and *Z. mays* [Poaceae] with values at 100 and *L. japonicus* gave values at 96. Besides the clusters that did not receive greater than 50% support values were

observed between *O. sativa* and *Z. mays* with values at 56 weak support values and a large clade at the base of the tree. This could be a result of a long conserved geneorder and content and only recent changes noticed at the tips, indicative of a recent divergence between the two species (Figure 1).

Comparisons performed between any two genomes for GeneOrder3.0 and any two to five genomes for CoreGenes2.0 is easily visualized graphically. In order to fully understand the correlations that exist among the genomes, the combination of these two tools is important. As more and more complete genomes are sequenced, conservation of gene order between different organisms is emerging as an informative property of the genomes. Any rearrangements in the genomes could be a result of flipping of entire sets of genes or just a subset of these genes. Earlier studies have shown that even if the loss in gene order may be faster than loss in the similarity of sequences, there would still be some conservation, which is at medium phylogenetic distance. Common gene content between two organisms is an indication of genomic estimation of distances between them.

Differences in gene order are a result of chromosome changes like translocation, deletion, inversion and any major sequence flipping. Compared to gene co-linearity, gene content is not influenced by the environment of the organism. As a result, gene content and organization tend to be highly conserved in most of the terrestrial plants. The presence of conserved structure can be applied in an interspecific manner in search for any functional genes and their gene organization. If this information is known and made available, then it becomes easy during molecular analysis to introduce foreign genes into chloroplast DNA with more control and precision (Daniell et al., 1998; Heifetz, 2000). In order for comparative analysis to provide meaningful deductions, the conserved functional sequences have to stand out as distinct from the nonfunctional sequences that were not conserved (Fraser and Eisen, 2000). This is important in sequence annotation and gene prediction. That degree of distinction requires the passage of time in order for mutations and the lack of selection pressures to cause the nonfunctional sequences in the two genomes to drift apart. To this end, there is a greater need to develop more robust algorithms to study the organization of these genomes and also improve on the visualization and interpretation of the outputs.

REFERENCES

Benson DA, Boguski M, Lipman, DJ, Ostell J (1998). GenBank. Nucleic Acids Res. 22(17): 3441-3444.

Daniell H, Datta R, Varma S, Gray S, Lee SB (1998). Containment of herbicide resistance through genetic engineering of the chloroplast genome. *Nature Biotechnology* 16:345-348.

Dubchak I, Brudno M, Loots GG, Mayor C, Pachter L, Rubin EM, Frazer KA (2000). Active Conservation of Noncoding Sequences Revealed by 3-way Species Comparisons. Genome Res. 10:1304.

Fraser CM, Eisen J, Fleischmann RD, Ketchum KA, Peterson S (2000). Comparative genomics and understanding of microbial biology. Emerg Infect Dis. 6(5):505-12.

Goloboff P (1994). Nona: A tree search program. Program and Documentation that is available from www.cladistics.com.org.

Heifetz P (2000). Genetic engineering of the chloroplast. *Biochimie* 82:655-666.

Herdt RW (1998). Assisting developing countries toward food selfreliance. Proc. Natl. Acad Sci. USA 95: 1989-1992.

Huynen MA, Bork P(1998). Measuring genome evolution. Proc. Natl. Acad. Sci. USA. 95:5849-5856.

Huynen MA, Snel B, Lathe W, Bork P (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. 10:1204 - 210.

Kilel et al. (2004). In preparation. Re-annotation and phylogenetic characterization of common genes found in complete chloroplast genomes.

Kishore GM, Shewmaker C (1999). Biotechnology: Enhancing human nutrition in developing and developed worlds. Proc. Natl. Acad. Sci. USA 96: 5968-5972.

Kundeti et al. (2003). In preparation. GeneOrder3.0.

Maier RM, Neckermann K, Igloi GL, Kössel H (1995). Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J. Mol. Biol. 251:614-628.

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000). VISTA: Visualizing Global DNA Sequence Alignments of Arbitrary Length. Bioinformatics 16:1046.

Mazumder R, Kolaskar A, Seto D (2001). GeneOrder: Comparing the order of genes in small genomes. Bioinformatics 17:162-166.

Nixon KC (1999a). Winclada (beta) ver. 0.9.99m24. Pubished by the author, Ithaca, NY.

Nixon KC (1999b). The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* 15: 407-414.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999). The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA. 96:2896-2901.

Snel B, Bork P, Huynen MA (1999). Genome phylogeny based on gene content. Nat. Genet. 21:108-110.

Stoebe B, MartinW, Kowallik KV (1998). Distribution and nomenclature of protein-coding genes in 12 sequenced chloroplast genomes. Plant Mol. Biol. Rep. 16:243.

Tamames J, Ouzounis C, Casari G, Valencia A (1997). Conserved clusters of functionally related genes in two bacterial genomes. J. Mol. Evol. 44:66-73.

Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

Wakasugi T, Tsudzuki T, Sugiura M (2001). The genomics of land plant chloroplasts: Gene content and alternation of genomic information by RNA editing. Photosynthesis Res. 70: 107-118.

Zafar N, Mazumder R, Seto D (2001). Comparisons of gene co-linearity in genomes using GeneOrder2.0. Trends Biochem. Sci. 26:514-516.

Zafar N, Mazumder R, Seto D (2002). Coregenes: A computational tool for identifying and cataloguing "Core" genes in a set of small genomes. BMC Bioinformatics 3:12.

Zeltz P, Hess WR, Neckermann K, Borner T, Kossel H (1993). Editing of the chloroplast rpoB transcript is independent of chloroplast translation and shows different patterns in barley and maize. EMBO J. 12:4291-4296.