Statistical Data Visualization of Two Important Proteins as Phylogenetic Tools in Chloroplast Genomes

Beatrice Kilel
George Mason University
School of Computational Sciences
Fairfax, Virginia
Motivation

- With recent and continuing advances in bioinformatics, many whole genomes now available
- Lack of good visualization techniques for rapid identification of regions of possible structural and/or functional importance [4]
- Using tools that go beyond 2D views
- Viewing data simultaneously, identifying clusters, frequencies, relationships and patterns in data, identifying outliers and anomalous data with the techniques used depending on data and objective.
- Extrapolate information from the fully sequenced genomes to those with economical importance
- Evolutionary reconstruction
Visualization tools used in the study

- **Scatter plots** [1] show how much one variable is affected by the other through correlations.
- **Parallel Coordinates** [2] represent a collection of data points as y-axis coordinate values arrayed along the x-axis.
- **Principal Component Plots** [6] for data dimensionality reduction.
- **Density Estimate** [5] for valuable indication of such features as skewness and multimodality in the data.
- **Star Coordinates** for examining the relative behavior of all variables in a multivariate data set.
Important genes used

- Rubisco - ribulose-1,5-bisphosphate carboxylase \( (\text{rbCL} \text{ gene}) \) plays an important role in Calvin-Benson-Bassham cycle for the conversion of inorganic atmospheric \( \text{CO}_2 \) into organic cellular constituents.

- \( \text{Rps8} \) is an important ribosomal protein found in most complete chloroplast genomes and plays an important role in translation of all the mRNA to proteins.

- Both proteins are found ubiquitously in archea, bacteria, and eucaryotes. Have been widely sequenced, used in several studies as model genes, well conserved hence can be useful in phylogenetic studies.
Conclusions

- Visual tools allow for the creation of complex views of large amounts of inter-related data, presentation of various types of evidence in required context (e.g., similar genes together – [3]), and the productivity of data mining [8].
- By mapping the results of comparative genomics analysis onto a phylogenetic framework, a foundation for future molecular investigations are made.
- Statistical data visualization can provide one of the more powerful means of analyzing sequence conservation across multiple sequence alignment, condensing the mass of information present [10].
- By applying the BRUSH-TOUR and TOUR-PRUNE techniques with visualization tools like Crystal Vision, we can resolve some of the problems of 2D graphs like dendrograms for showing evolutionary relationships.