Visualizing Multiscale Complexity and Features of Biological Sequence Datasets

Leonid Zaslavsky[1*], Boris Fedorov[1], Azat Badretdin[1], Yiming Bao[1], James R. Brister[1], Stacy Ciufo[1], William Klimke[1], Kathleen O'Neill[1], Alexander Souvorov[1] and Tatiana Tatusova[1]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 45 Center Drive, Bethesda, Maryland, 20894, USA
[*]Corresponding author: leonid.zaslavsky@nih.gov

Abstract

The number of genomic and protein sequences in public databases have grown tremendously in recent years. Sophisticated tools are required to provide meaningful overview and request-specific visual representation of data.

Users exploring NCBI sequence databases have different needs and benefit from request-centric data representations with multiple scales of resolution. This allows an overview of large-scale homology relationships and focusing on a relevant scale of resolution, with views of rearrangements and variations between the sequences. Thereby adaptively focused visualization is refined to the sequences most relevant to the request with the rest of the data represented more coarsely. Aggregation techniques also provide a mechanism to deal with the great variation in sequence population density in a database.

Aggregated representation can also be applied to metadata. In some cases (e.g., virus variation), an annotation is based on either season/year or geographic location, and can be easily aggregated for a group of sequences. In other cases, more complex approach is required, such as comparison to another tree with named hierarchy of nodes (e.g., genome tree or taxonomy) and keyword extraction.

Our algorithm for adaptive visual representation of large trees along with metadata aggregation has been implemented within the NCBI Virus Variation Resources and the NCBI 16S rRNA Resource, with ongoing work on a online visualization framework allowing for visualization of more general sets of protein and genomic sequences.