Martin Luther University Halle Wittenberg

Eberhard Karls Universität Tübingen

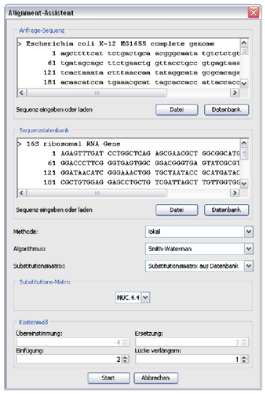
A Tool for Interactive Comparative Sequence Analysis

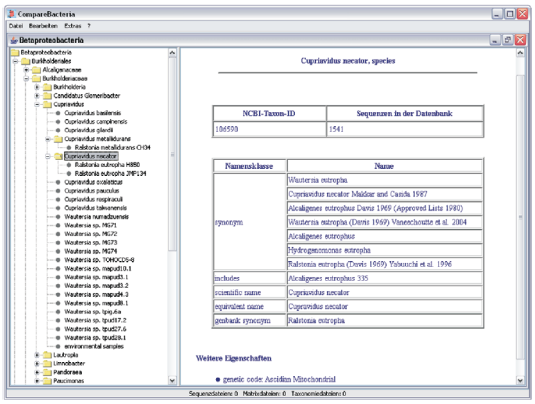
Andreas Dräger[1], Dietrich H. Nies[2] and Stefan Posch[3]

[1]Center for Bioinformatics Tübingen (ZBIT), [2]Institute for Microbiology Halle, [3]Institute for Computer Science Halle

Introduction

Sequenced species of the β-Proteobacteria show a high degree of diversity as this taxonomic group inhabits various ecological niches [1]. Therefore special biological features are needed. By comparison of proteins with essential functions in a taxonomic context a method was developed to normalize subsequent comparisons of these specialized proteins. To use online databases like NCBI, RDP, EMBL, JGI, SwissProt and Tigr, which provide genomic and proteomic sequence data from different species, to perform a comparative local analysis, an efficient way of storage has to be found. Problems of data storage occur due to different naming conventions of species, genes, proteins and other biological data. To maintain data consistently without redundancy a novel database application was implemented.





Materials and Methods

Implementation

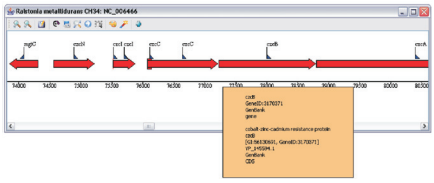
A MySQL database server was installed using the BioSQL relation scheme [2]. The client was implemented in Java using the open source library BioJava [3]. Global and local alignments were performed by the BLAST program, implementations of the Smith-Waterman- and Needleman-Wunsch-algorithm and a special Hidden Markov Model.

Taxonomic Proximity and Protein Similarity of Bacteria

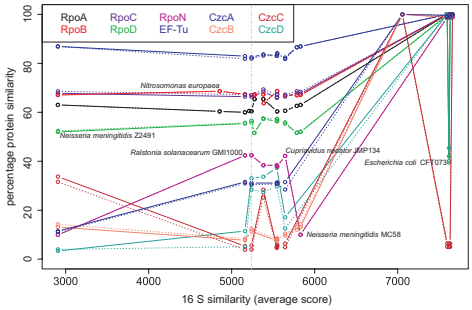
Taxonomic information can be derived from global alignments of the highly conserved 16 S rDNA [4], which is involved in the bacterial protein synthesis. Proteins are compared using local alignments to consider their functional domains.

Selection of Species and Proteins

Essential (DNA translation and transcription) and nonessential (heavy metal resistance) proteins of 15 β-Proteobacteria and 7 γ-Proteobacteria with known 16 S rDNA were used for the comparison relative to *E. coli* K12.



Similarity relative to *Escherichia coli* K12



Results

The client program contains a graphical user interface to faciliate easy interaction with the database for a non-expert. Necessary pre-processing steps to integrate data into the local database are performed automatically. Different file formats of biological sequence data can be imported and converted into each other. Global and local alignments can be performed interactively. Additionally, visualizations of the data, such as the taxonomic tree and genetic/proteomic annotations, are provided. The newly developed program provides convenient local storage, management and analysis of biological data.

A protein and gene comparison of selected Proteobacteria was performed. Plotting the protein similarity against the taxonomic neighborhood shows an almost linear increase for essential proteins and a higher variability for nonessential proteins with taxonomic proximity. This leads to the conclusion that specialized proteins with nonessential functions are less conserved and can be detected from the taxonomic context by normalization with essential proteins.

References

[1] M. Mergeay *et al. Ralstonia metallidurans*, a Bacterium Specifically Adapted to Toxic Metals: Towards a Catalogue of Metal-Responsive Genes. *FEMS Microbiology Reviews*, (27):285-410, May 2003.

[2] OBDA. Open Bioinformatics Foundation. `http://obda.open-bio.org/`.

[3] M. Schreiber, R. Holland, M. Heuer, T. Down, M. Pocock, K. James, and D. Huen. BioJava-1.4, `http://www.biojava.org`, July 2005.

[4] J. Wuyts *et al.* Distribution of Substitution Rates and Location of Insertion Sites in the Tertiary Structure of Ribosomal RNA. *Nucleic Acids Research*, 29(24):5017-5028, 2001.

Contact: Andreas Dräger, andreas.draeger@uni-tuebingen.de | www-ra.informatik.uni-tuebingen.de