DIPLOMARBEIT

UNIVERSITÄTSZENTRUM INFORMATIK

Martin-Luther-Universität Halle-Wittenberg

Automatische und vergleichende Analyse bakterieller Genome mit Schwerpunkt auf *RalstonialCupriavidus* Arten sowie verwandten Proteobakterien

Andreas Dräger

(2005)

Allgemeine Angaben

Die Diplomarbeit wurde am Lehrstuhl für Bioinformatik von Prof. Dr.-Ing. Stefan Posch, Institut für Informatik, Fachbereich Mathematik und Informatik, sowie am Lehrstuhl für Molekulare Mikrobiologie von Prof. Dr. Dietrich H. Nies, Institut für Mikrobiologie, Fachbereich Biologie, der Martin-Luther-Universität angefertigt.

Zur Kontaktaufnahme benutzen Sie bitte die Email-Adresse
direktor@uzi.uni-halle.de.

Summary

Knowledge generation by comparative genomics become more and more meaningful due to the increased availability of genomic sequence data. This thesis compares the genomes of β-*Proteobacteria,* a taxonomic group of bacteria showing a high degree of diversity. Sequenced species of this group inhabit various ecological niches. Therefore, they need special features to surviving toxic heavy metal concentrations, to act as pathogens of plants or animals or to degrade organic substances, which are normally difficult to break down. These characteristic features are mediated by special proteins, leading to the question, if characteristic proteins can be detected automatically by comparing the sequences of related proteins of different species. By comparison of highly conserved proteins with essential life functions in a taxonomic context a measurement should be developed to normalize later on comparison of these specialized proteins. The taxonomic context of different bacterial species can be derived from comparisons of the highly conserved genes for the 16 S rRNA, which is involved in the protein synthesis of bacteria.

To perform those comparisons different algorithms have been released. For global alignments the Needleman-Wunsch-Algorithm or a special Hidden Markov Model can be used. Local alignments can be done by the Smith-Waterman-Algorithm or its heuristic approximation BLAST. Global alignments compare whole sequences, whereas local alignments find longest conserved regions in two sequences. Proteins contain functional regions (domains) surrounded by less important regions, so that these can be compared by local alignments. The complete 16 S rRNA sequence is

essential for its functionality, so that for taxonomic analysis global alignments are needed. To evaluate the quality of alignments different substitutions matrices like BLOSUM, PAM, NUC contain scores for the substitution of a symbol with another one. The sum of these scores is the score of the alignment.

Several online database servers like NCBI, RDP, EMBL, JGI, SwissProt, Tigr and many others provide genomic and proteomic sequence data of different species. These databases grow at an exponential rate, because new techniques of sequencing proteins and genetic elements in high throughput analysis are used and data can be uploaded by individual authors, laboratories or other scientific institutions. However, to use these data to perform comparative local analysis, an efficient way of storage has to be found.

Problems of data storage occur due to the different naming conventions of species, genes, proteins and other biological data. To maintain the data consistently without redundancy, a database server was installed using MySQL. As a relation scheme BioSQL was used. To integrate downloaded data into the local database different pre-processing steps (data cleaning) were necessary.

A client program to compare the locally stored data was implemented in Java, using the open source library BioJava. The resulting database application also provides interactive visualizations of the data in the database such as the taxonomic tree and genetic annotations. It contains a graphical user interface to interact with BioSQL without knowledge of databases. In addition, different file formats of biological sequence data can be converted into each other. With a special dialog global and local sequence alignments can be performed interactively. The BioJava library had to be extended to provide the full functionality of that program.

This application was used on a selection of essential proteins (involved in DNA transcription and translation) of 15 β-*Proteobacteria* and 7 γ-*Proteobacteria* and non essential proteins (involved in heavy metal resistance) and the 16 S rRNA genes. All comparisons are relative to the well investigated species *Escherichia coli* K12.

Plotting the protein similarity against the taxonomic neighborhood shows an almost linear increase for essential proteins with taxonomic nearness. The other proteins show higher variability. This leads to the conclusion that proteins with special features are less conserved than highly essential proteins, so that these can be detected from the taxonomic context by normalization with essential proteins.