MetaLook & MetaMine: Data Mining and Analysis of Gene Patterns in an Environmental Context

Uta Bohnebeck*, Thierry Lombardot†, Renzo Kottmann‡ and Frank O. Glöckner*#

*Technology Transfer Centre Bremerhaven, Germany, bohnebeck@ttz-bremerhaven.de
†Max Planck Institute for Marine Microbiology, Germany, {tlombard | rkottman | fog}@mpi-bremen.de
‡Jacobs University Bremen, Germany

MetaLook and MetaMine are Java-based software tools developed in the context of the EU project MetaFunctions [1]. The goal is to detect and assign functions to habitat-specific gene patterns. The underlying database contains prokaryotic genome and metagenome sequences of marine origin together with information about their environmental context. Habitat parameters like water and sediment depths, temperature, salinity, and other physical-chemical properties are extracted from the literature or extrapolated [2] based on global ocean data sets (WOA, WOD [3]) and remote sensing information (SeaWiFS [4]).

Two basic types of gene patterns are currently under investigation: (1) genes which are present or over/under-represented under specific environmental conditions (MetaLook [5]) and (2) patterns consisting of a set of genes occurring in specific genomic neighbourhoods (MetaMine). If such gene patterns are found the respective genomic context helps in assigning a potential function. In addition, if the gene patterns correlate with distinct environmental parameters or processes potential functions may again be inferred.

MetaLook is a search tool for marine ecological genomics. It offers a three-dimensional toolbox for visualisation and data mining of DNA sequence information in a geographical and ecological context. The core component allows the definition of several gene containers related to specific environmental parameters. These containers can than be searched against each other to detect shared genes or genes which are only present in a single container. This container represents a set of relevant environmental “key” genes, which can be used as the starting point for gene pattern detection in MetaMine.

MetaMine is a graphical oriented data mining tool for the detection of gene patterns. The following user controlled process steps can be performed: (1) definition of a key gene and import / retrieval of the key gene sequence, (2) BLAST search of the key gene sequence against the local database, (3) determination of the k neighbouring genes at each side of the key gene, (4) all against all BLAST searches of the k neighbouring genes, (5) determination of orthologous genes using the reciprocal best hit approach and minimal triangle relationship, (6) determination of gene patterns with a given minimal length and presence in at least q (quorum) different genomes providing systematic and heuristic search strategies, (7) clustering and visualisation of gene pattern instances concerning gene order and direction together with their environmental parameters. The results of each process step are stored in the database and can be reiterated at each step with other parameters.

First analysis show interesting patterns of genes involved in methane-related metabolisms like Methanogenesis and the anaerobic oxidation of methane.

References: