**STRUCTURAL AND FUNCTIONAL ANNOTATION OF GENOMIC SEQUENCES:ASSIGNMENT OF FOLD FAMILY AND SORTING OF PROTEINS WITH RESPECT TO SUBCELLULAR LOCALIZATION.**

EISENHABER FRANK1,2, BORK PEER1,2, HUYNEN MARTIJN1,2, ORENGO CHRISTINE3, SCHULTZ JORG1,4, SUNYAEV SHAMIL R.1, YUAN YANPING1,2

1EMBL Heidelberg, Meyerhofstr. 1, D-69012 Heidelberg, Fed. Rep. Germany;

2Max-Delbrueck-Centrum fuer Molekulare Medizin, Robert-Roessle-Str. 10, D-13122   Berlin-Buch, Fed. Rep. Germany;

3University College, Dept. Biochemistry & Molecular Biology, London, U.K.;

4Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilov Street 32, 117984, Moscow, Russia

Keywords: genomic sequences, proteins, functional annotation, fold assignment, cellular location

As a result of large-scale genomic sequencing, the gene sequences of many proteins are known but their structure and function is hardly understood. The major source of hypothetical information about such uncharacterized proteins is sequence comparison and the extrapolation of annotated information from homologous proteins. The information transfer to query proteins must be done carefully since (1) a similar sequence does not always imply similar protein structure and function and (2) the annotation of database proteins might be incomplete or even wrong. Usually only structural features and molecular functions can be transferred whereas cellular and phenotypic functions should be analyzed in the biological context of the organism considered.

The protein sequence and structure databases are a valuable tool for the prediction of structural features by homology. Iterative homology searches have been used to assign folds to the protein sequences derived from the complete genomes of*M. genitalium, E. coli, and M. jannaschii.*Protein sequence segments supposed to represent coiled coil and transmembrane regions have been excluded from the analysis. The procedure resulted in fold assignments for at least one globular domain in 30-40% of all proteins of*M. genitalium, E. coli,*and*M. janaschii.*Several proteins with multiple domains of known structure have also been identified. The accuracy of this prediction appears 98% as estimated from iterative homology searches for the 685 sequences of a maximal subset of non-homologous proteins extracted from the Brookhaven Protein Data Bank.

For the analysis of genome-phenotype relationships, the retrieval of complete sets of proteins with respect to functional categories, for example subcellular localization, signal transduction cascades, etc. is necessary. Annotations of proteins in databases are generally written for a human reader and use a wide variety in terminology for a detailed description of phenomena. Therefore, it is difficult to retrieve them with keyword searching engines such as SRS. For example, queries with stems of simple keywords such as "extracell" or "membrane" classify only 22% of all entries in SWISS-PROT, the currently best annotated protein sequence database.

The introduction of controlled vocabularies and hierarchical functional descriptions are a possible solution for the problem of categorized functional annotation but such systems require permanent efforts for adaption and updates and the subsequent rewriting of the sequence databases. An alternative way as exemplified by the META\_A (annotator) software system consists in a computer program able to evaluate the annotation with sets of biological rules encoded in form of regular expressions. Applied to the problem of subcellular localization, it was possible to assign useful cellular location attributes to more than 88% of all SWISS-PROT entries. The overwhelming majority of remaining entries does not have functional annotation at all. An application of the system to the*M. genitalium*sequences revealed the probable absence of purely extracellular proteins in this organism.

The determination of proteins which are at least partially localized extracellularly just with sequence-analytic methods is of great interest for the identification of protein targets, for example in pharmaceutical applications. A set of alignments of dozens of protein domains has been produced; their identification in protein queries and EST's can be used as marker for localization. The collection of profiles is also applicable in a sensitive search for domain detection. Many hitherto unannotated domains could be classified in database sequences as a comparison with the output of the META\_A program invoked with a set domain-recognition rules revealed. In addition, we found also database entries with probably completely wrong functional annotations. Our results allow also to estimate the number of human proteins with extracellular localization.