

Einladung zur Vortragsreihe *Algorithmische Bioinformatik*

Herr Dr. Hanjo Täubig, TU München

spricht über

Fast Structure Searching for Computational Proteomics

Datum: Dienstag, 23. Dezember 2008
Zeit: 14:00 Uhr s.t.
Ort: B-IT, Dahlmannstr. 2, Rheinsaal

In 2003 the Human Genome Project and Celera Genomics celebrated the completion of sequencing the human genome. Although the project was a great success, it has become apparent that knowledge of the sequence of amino acids alone would not allow one to make significant progress in curing any illness. The usefulness of the results and methods was mostly restricted to detecting predisposition to a variety of diseases, but the responsible gene did not provide much information on the real reason of the disfunction. The key to a better understanding of the functional relations must therefore be the structure of the agents that are participating in the respective processes. In the majority of cases, these reactions involve biochemical macromolecules like proteins or nucleic acids. Their structural diversity, created by alternative splicing and post-translational modifications, is a prerequisite for performing the vast number of different functions, that depend on the chemical specificity of the reactants.

Starting in the seventies, molecular structures of polypeptides and nucleic acids (determined by x-ray crystallography and NMR spectroscopy) were deposited in the Protein Data Bank (PDB), which is the primary source for structure information used in today's molecular biology and structural genomics research. Since its first days, the PDB has witnessed a rapid growth at exponential rates. Today, it holds more than 40.000 structures, and the size of the database exceeds twenty gigabytes. These huge numbers emphasize the urgent need for fast methods allowing one to search the database of existing structures.

Our work aims to exploit the advantages of text indexing methods commonly used in pattern matching for solving problems in structural genomics and computational proteomics. In particular, we apply suffix trees to problems related to structural databases of biopolymers like RNA and proteins. The main contribution is an approach for fast searching in huge databases like the PDB. While existing methods only have search times in the order of minutes, hours, or even days, our approach allows us to perform standard queries within seconds. The data structure is based on a generalized suffix tree which is extended by a method for approximate matching of special adapted alphabets. These alphabets rely on the discretization of translation- and rotation- invariant measures that represent the backbone conformation of the molecule. The method was evaluated by applying structural queries to the PDB and comparing the results to established tools in this area. On the one hand, the experiments demonstrate a significant reduction of the query time while comparable results are produced. On the other hand, several structures have been found by our approach that are missing in the result list of other tools because they are filtered out by sequence-based heuristics. Another contribution is a method for identifying frequent motifs and for partitioning a database of protein structures according to structural similarity which is, in contrast to currently used methods, fully automatic.