



## Vortragsankündigung

# Algorithms for Efficient Exact

# Motif Discovery

**Prof. Dr. Sven Rahmann**  
**Bioinformatics for High-Throughput Technologies**  
**Computer Science Department**  
**TU Dortmund**

The motif discovery problem consists of finding over-represented patterns in a collection of biosequences. It is one of the classical sequence analysis problems, but still has not been satisfactorily solved in an exact and efficient manner. This is partly due to the large number of possibilities of defining the motif search space and the notion of over-representation. Even for well-defined formalizations, the problem is frequently solved in an ad hoc manner with heuristics that do not guarantee to find the best motif.

We show how to solve the motif discovery problem (almost) exactly on a practically relevant space of IUPAC generalized string patterns, using the  $p$ -value with respect to an i.i.d. model or a Markov model as the measure of over-representation. In particular, (i) we use a highly accurate compound Poisson approximation for the null distribution of the number of motif occurrences. We show how to compute the exact clump size distribution using a recently introduced device called probabilistic arithmetic automaton (PAA). (ii) We describe an efficient algorithm to discover the optimal pattern with respect to the  $p$ -value. The method exploits monotonicity properties of the compound Poisson approximation and new bounds on the expected motif clump size and is two orders of magnitude faster than our previous algorithm presented at ISMB'09.

im  
**Zentrum für Bioinformatik**  
**Bundesstraße 43**  
**Raum 16**  
**am FREITAG, den 23.07.2010**  
**um 12:15 s.t.**

**Everyone is welcome!**