UNIVERSITÀ DEGLI STUDI DI SALERNO
DIPARTIMENTO DI INFORMATICA

# Avviso di Seminario

I giorni 7 e 8 Gennaio 2016 alle ore 11:00 il Prof. Burkhard Morgenstern dell' "Institut für Mikrobiologie und Genetik" dell' "Universität Göttingen" terrà un seminario diviso in due parti presso la sala dei consigli del Dipartimento di Informatica, stecca 7, piano 2, stanza 11.

## Spaced Words for alignment-free sequence comparison and an improved hill-climbing algorithm for seed optimization

La prima parte sarà dedicate all'introduzione del tema Alignment Free Comparison nell'ambito dell'analisi delle sequenze biologiche e nella seconda parte saranno affrontati nel dettaglio alcuni risultati scientifici prodotti dal prof. Morgenstern.

## Abstract

With the huge amount of sequence data that are now available, pairwise and multiple alignment have become too slow for many sequence-analysis tasks. Therefore, alignment-free methods are increasingly used for genome comparison and phylogeny reconstruction. Most alignment-free algorithms work by comparing the word composition of sequences. Sequences are represented by word-frequency vectors, and standard distance measures on vector spaces can be applied to define a pairwise distance matrix for a set of input sequences. Phylogenetic trees can then be calculated from these distance matrices with standard distance-based methods such as UPGMA or Neighbour-Joining.

Database search programs such as BLAST originally used word matches of a fixed length k as seeds to search for local homologies. Later, it has been shown that the sensitivity and speed of these programs can be substantially improved if spaced seeds – i.e. word matches with possible mismatches at certain pre-defined don't-care positions – are used instead of contiguous word matches. Considerable efforts have been made since then to optimize patterns for this spaced-seed approach.

Inspired by these approaches, we proposed to use spaced words for alignment-free sequence comparison, i.e. words containing wildcard characters at fixed positions, according to one or several underlying patterns of match and don't-care positions. In a benchmark study, we could show that, if multiple patterns are used, our spaced-words approach produces better phylogenies than previously proposed alignment-free methods that are based on contiguous words.

In subsequent studies, we proposed to estimate evolutionary distances between DNA sequences more accurately, based on the number N of spaced-word matches between them, and we showed

how the variance of N can be estimated. We showed that the variance of N is closely related to the so-called overlap complexity of a set of patterns that has been proposed by Ilie and Ilie in the context of database searching. We proposed an improved hill-climbing algorithm to optimize sets of patterns (seeds) for alignment-free sequence comparison and database searching.

Our software is freely available as source code and through a user-friendly WWW interface at Gottingen Bioinformatics Compute Server (GOBICS) http://spaced.gobics.de/.

**Prof. Burkhard Morgenstern - *Short bio***

since 2002: Full professor, Univ. Göttingen

2001 - 2002: Group leader, Int. Grad. School of Bioinformatics and Genome Research, Univ. Bielefeld

2000 - 2001: MIPS, Max-Planck-Institut, Martinsried, and GSF Research Center, Neuherberg

1998 - 2000: RPR / Aventis Pharma, Dagenham, UK

1997 - 1998: Visiting scientist with Prof. W.R. Atchley, North Carolina State Univ., USA

1996 - 1998: Postdoc, GSF Research Center, Neuherberg

1996: PhD, Univ. Bielefeld. Thesis work on multiple sequence alignment. Supervisor: Prof. A. Dress

1993: Diplom (Mathematics), Univ. Munich. Thesis work on partial differential equations. Supervisor: Prof. J. Batt