Speaker: Paul Medvedev, Penn State University

Title: "Data structures to represent sets of *k*-mers."

Abstract:

The analysis of biological sequencing data has been one of the biggest applications of string algorithms. The approaches used in many such applications are based on the analysis of k-mers, which are short fixed-length strings present in a dataset. While these approaches are rather diverse, storing and querying k-mer sets has emerged as a shared underlying component and there have been many specialized data structures for their representation. In this talk, I will describe the applications of k-mer sets in bioinformatics and motivate the need for specialized data structures. I will give an overview of known approaches and lower bounds, with a focus on unitig-based representations.

Finally, I will describe a data structure for representing sets of k-mer sets, called the HowDe Sequence Bloom Tree.

A brief bio is:

Paul Medvedev is an Associate Professor in the Department of Computer Science and Engineering and the Department of Biochemistry and Molecular Biology and the Director of the Center for Computational Biology and Bioinformatics at the Pennsylvania State University. His research focus is on developing computer science techniques for analysis of biological data and on answering fundamental biological questions using such methods. Prior to joining Penn State in 2012, he was a postdoc at the University of California, San Diego and a visiting scholar at the Oregon Health & Sciences University and the University of Bielefeld. He received his Ph.D. from the University of Toronto in 2010, his M.Sc. from the University of Southern Denmark in 2004, and his B.S. from the University of California, Los Angeles in 2002.