The Department of Computer Science and Engineering

University of Nevada, Reno

cordially invites you to a PhD defense

Fuzzy Sequence Classification and Assembly of Environmental Genomes

A colloquia submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
with a major in Computer Science and Engineering.

by

Sara Nasser

Abstract: Traditional methods obtain an organism's DNA by culturing it individually. Recent advances in genomics have lead to the procurement of DNA of more than one organism from its natural habitat. The population contains fragments from several closely related species. Assembling these genomes is a crucial step irrespective of the method of obtaining the DNA. This dissertation will present our proposed methods for fuzzy multiple genome sequence assembly.

An optimal alignment of fragments is based on several factors, such as the quality of bases and the length of overlap. Factors such as quality indicate if the data is an actual read or an experimental error. Current sequencing methods tend to reject sequences that do not match with a high degree of similarity. This can lead to large amounts of data being rejected that otherwise may be important. To address this challenge we propose a sequence assembly solution based on fuzzy logic, which allows for tolerance of inexactness or errors in fragment matching and can be used to create optimal assembly.

Assembly of a single organism's genome is presented using a modified dynamic programming approach with fuzzy characteristic functions. The characteristic functions are used to select optimal alignments of sequence fragments. Assembly of environmental genomes starts with the classification of mixed fragments from different organisms into homogeneous groups. Separating closely related species is a dissertationcult task because the fragments contain many similarities. We propose fuzzy classification using modified fuzzy weighted averages to classify fragments belonging to different organisms within an environmental genome population. Our proposed approach uses DNA-based signatures such as GC content and nucleotide frequencies as features for the classification. This divide-and-conquer strategy also improves performance on larger datasets. We evaluate our method on artificially created environmental genomes to test various combinations of organisms and on environmental genomes obtained from acid mine drainage available at National Center for Biotechnology Information.

4:00 pm, Monday, March 17^{th}, 2008

Scrugham Engineering and Mines (SEM) 201

for more information contact Dr. Fred Harris @ 784-6571 (Fred.Harris@cse.unr.edu)