



BCB/IGERT Thesis Seminar

*A modular data analysis pipeline
for the discovery of novel RNA motifs*

JUSTIN SCHONFELD

Ph.D. Candidate

Bioinformatics and Computational Biology

Department of Mathematics

Major Professors: Dan Ashlock and Dan Voytas

2 p.m. Monday, April 10, 2006
390 Carver Hall

Abstract:

This dissertation presents a modular software pipeline that searches collections of RNA sequences for novel RNA motifs. In this case the motifs incorporate elements of primary and secondary structure. The motif search pipeline breaks up sets of RNA sequences into shorted segments of RNA primary sequence called bricks. The bricks are then folded to obtain low energy secondary structures. The distance estimation module of the pipeline then calculates distances between the folded bricks, and then analyzes the resulting distance matrices for patterns.

An initial implementation of the pipeline is applied to synthetic and biological data sets. This implementation introduces a new distance measure for comparing RNA sequences based on structural annotation of the folded sequence as well as a new data analysis technique called non-linear projection. The modular nature of the pipeline is then used to explore the relationships between several different distance measures on random data, synthetic data, and a biological data set consisting of iron response elements. It is shown that the different distance measures capture different relationships between the RNA sequences. The non-linear projection algorithm is used to produced 2-Dimensional projections of the distance matrices which are examined via inspection and k-means multiclustering. The pipeline is able to successfully cluster synthetic RNA sequences based only on primary sequence data as well as the iron response elements data set. The dissertation also presents a preliminary analysis of a large biological data set of HIV sequences in which crossover points were localized.