Data Mining and Bioinformatics

Faculty Member: Yi Lu Tel: 936-261-9862 Email: yilu@pvamu.edu

Focus Areas of Interest

- Data Mining
- Bioinformatics
- Cloud/Service Computing

Active Research Projects

1. Correlation mining to reveal the transcription network.
3. Scientific workflow driven bioinformatics pipeline to predict the flexible C-Terminal Tethers.

Research Facilities

2. Six Dell Precision Workstation 390, Intel® Core™ 2 Duo E4300.

Major Journal Publications Since 2004

1. Yi Lu and Jeffrey L. Ram, “Predictions of flexible C-terminal tethers of bacterial proteins with the FLEXTAIL bioinformatics pipeline”, International Journal of Data Mining and Bioinformatics (IJDMB), in press
Research project: Correlation mining to reveal the transcription network.

In this research, we will implement a new data mining algorithm to generate correlation rules between motifs and the expression profiles of genes with significant expression change through the time course of gene expression. Preliminary result of algorithm shows our prediction confirmed the result of the cell cycle in CDC28 data from literature and showed above figure, for each stage, only subset of the promoter will be turned on for transcription.

Research project: A Map-Reduced genetic K-Means algorithm for gene expression analysis
Cloud computing service provides us the great opportunity to solve data mining problem which was not solvable before due to the extensive computing needs. In this research, we will develop the map-reduce based clustering algorithm to reveal the knowledge from the gene expression data. As showed in the above figure, the gene expression data will be separated into different mappers and then combined and shuffled in the reducers.

**Research project: Scientific workflow driven bioinformatics pipeline to predict the flexible C-Terminal Tethers**

As showed in the figure below, proteins use conserved binding motifs associated with relatively unconserved flexible amino acid sequences as mobile tethers for interacting molecules, as exemplified by C-terminal tether sequences of bacterial chemotaxis receptors. We describe here a bioinformatics pipeline (as showed the figure) to discover new instances of flexible tethers and their binding motifs. C-terminal regions of all proteins in subsets of bacteria and archaea were analyzed for flexibility with DisEMBL hotloops, grouped with BLASTCLUST, aligned with ClustalW, analyzed for conserved 5-mers, and then putative tethers were identified as sequences with highly conserved (>80%) 5-mers that were at least 20% more conserved than the rest of the flexible region. The algorithm identified previously known flexible binding domains, as well as >100 other putative flexible tether sequences that should be further investigated for binding targets and flexibility.