Feature Characterization of sequence location in genome analysis

Presented by **Megha Sinha**

Oregon State University

Project Overview

- **Goal:** Identify features of sequence locations in a genome that distinguish two different types of genomic entities.

- **Selected Genome:** Armadillo genes (Dasnov3.0)
  - **Class1: 5' UTR Class 2: 3' UTR**

- **Classification method:** Support-vector machine (SVM)

Project Steps

Fasta Splitter

Splits the Fasta files into even "chunks" of sequences

Input: Fasta file, a chosen number F (number of splits)
Output: F number of splitted fasta files

Parallel Feature Extractor

Generates features from Fasta files. Runs on Sun Grid Engine (SGE), where we submit each splitted Fasta files as jobs, for fast processing

Inputs: F Fasta files (splitted)
Output: tab-delimited feature tables, merged into one

SVM classifier

Analyses and classifies data by marking them to one of the two class categories (5' or 3' UTR)

Input: Class 1 & 2 feature table
Output: Classification Accuracy

Features

AT content

% of nitrogenous bases that are either Adenine or Thymine

GC content

% of nitrogenous bases that are either Guanine or Cytosine

AT/GC ratio

The ratio of the sum of the adenine plus thymine bases to the sum of the guanine plus cytosine bases

AT cumulative skew

measure of over or under abundance of Adenine and Thymine

GC cumulative skew

measure of over or under abundance of Guanine and Cytosine

Z curve

Bioinformatics algorithm to measure distribution of nucleotides. It has three components as:
$x = (A + G) - (C + T)$
$y = (A + C) - (G + T)$
$z = (A + T) - (C + G)$
measured over $n = 0, 1, 2, \ldots, N$

N content

% of N's in the sequence

Result

```
optimization finished, #iter = 5617
Objective value = -1797.754244
nSV = 10389
COPY MODEL TO WEIGHT VECTOR
FREE SPACE
FREED SPACE
[1] "confusion matrix:"
           actuals
predictions    0    1
          0 1636  395
          1   77  903
[1] "accuracy:"
[1] 0.8432414
```

Accuracy ~ 82 - 85 %

Time ~ reduced from ~2 mins to 42 secs

THANK YOU

Oregon State
University