

# Analysis of human promoters and gene expressions

Kihoon Yoon & Stephen Kwek  
Department of Computer Science  
University of Texas at San Antonio

# Overview

- Main Idea: Find distinctive sequence patterns on promoter regions and identify corresponding TFs
  - Prospective signals selected by non-position specific n-mer counting scheme (overlapping binning method)
  - Combining gene expression and known TF interaction data to narrow down either corresponding TF candidates or number of binding sites
  - Construct a promoter complexity index as the primary indication of tissue specificities and expression level.
- The goal is to develop a system that can be used to predict the normal expression patterns for a given gene.

# Preliminary Results

- For instance,
  - SELENBP1 – selenium binding protein1

1.

Signals	Positions
TACTA	-399 ~ -395
ATACA	-391 ~ -387
CATAC	-198 ~ -194
CATAC	-182 ~ -178
GTATAA	-47 ~ -42
TATAAA	-46 ~ -41
GGTAAG	+51 ~ +56

2. Construct an index to represent its expression patterns in different tissue types.

- Clustering genes by their expression information
- Identify common and rare signals in a cluster
- Assign proper complexity scores to genes

3. Build a model from the complexity and the cluster membership information

# Discussion

- Potential signal validation scheme
  - Check the distribution of each signal on entire human chromosome sequences
  - Compare to known TF binding sites
  - Check with other species
- Extend our approach to a multigenic disease research.