Trends in Bioinformatics: Bi-Clustering with Biological Context Information

Willi Gierke

Supervisor: Cindy Perscheid

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Motivation
Gene Expression Analysis

- Gene expression occurs by transcribing the DNA into RNA
- RNA is then translated into e.g. proteins
- “Interprets” genetic code as it differs between cell types
- Expression properties influence phenotype

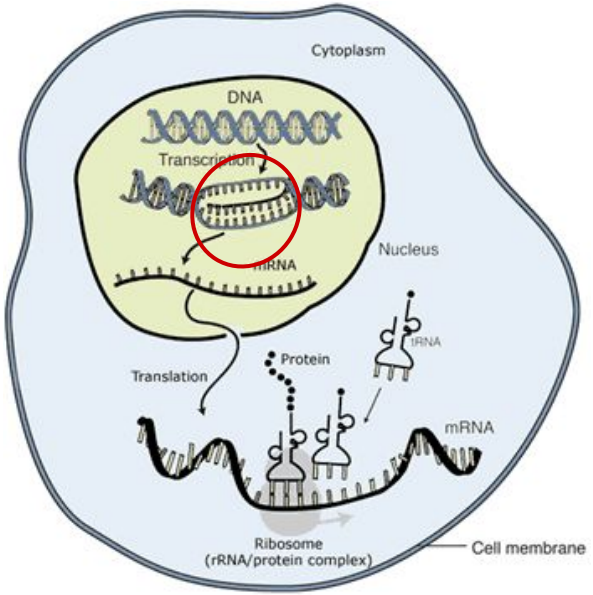


Image adapted from: National Human Genome Research Institute.

Motivation
Gene Expression Analysis

- How to find the task for each of the over 19.000 genes?

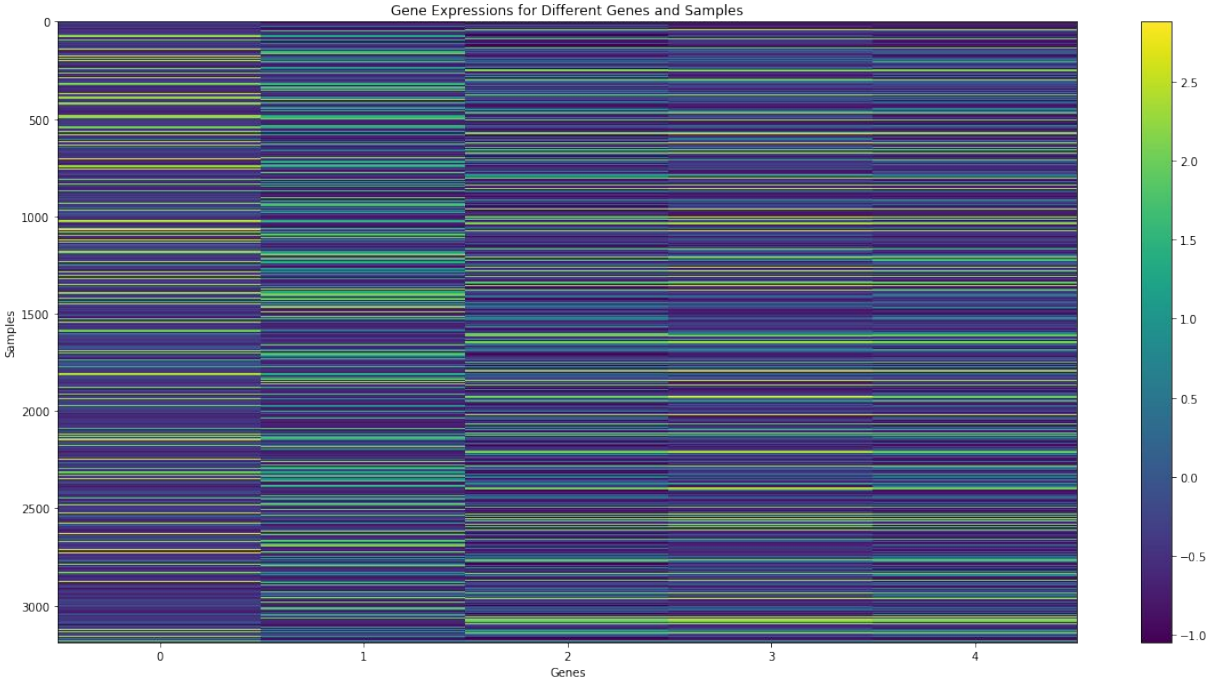
  → Infer from same expression levels same regulatory process (phylogenetic profiling)

- Idea: cluster genes based on expression levels

- Problem: gene contributes to ~10 processes

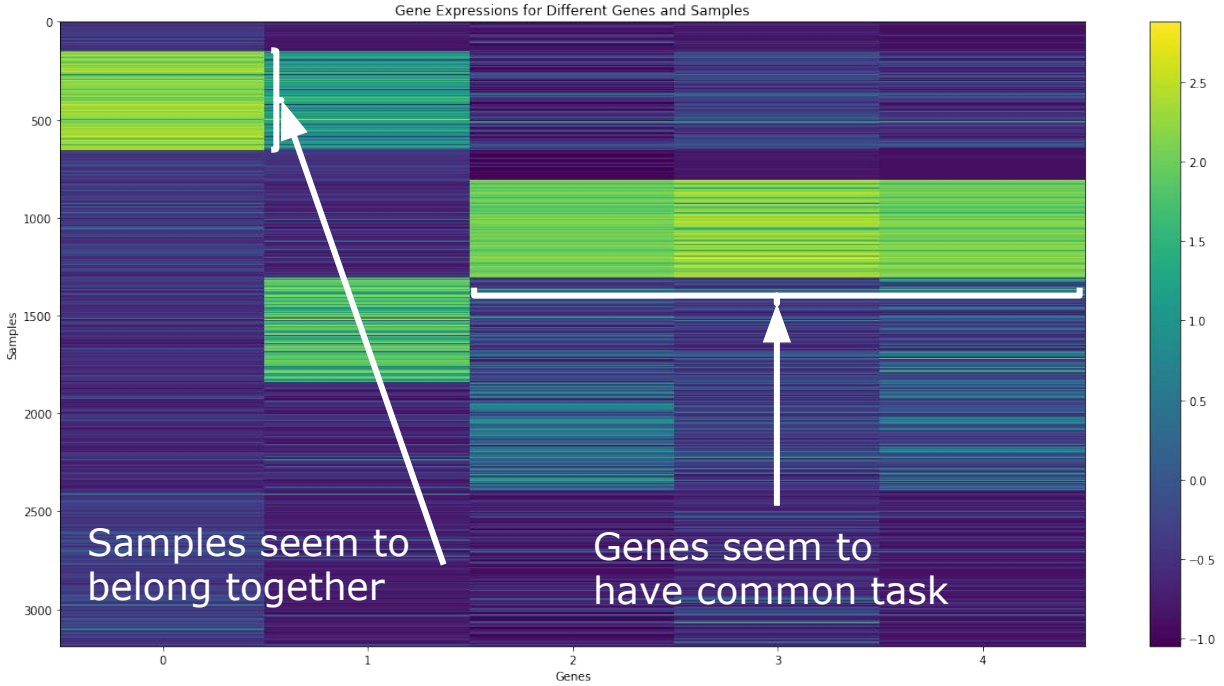
Motivation
Bi-Clustering

- Bi-Clustering: assign a data point to multiple clusters
- “Shuffle matrix to find (overlapping) groups”



Motivation
Bi-Clustering

- Bi-Clustering: assign a data point to multiple clusters
- “Shuffle matrix to find (overlapping) groups”



Why to include biological context information?

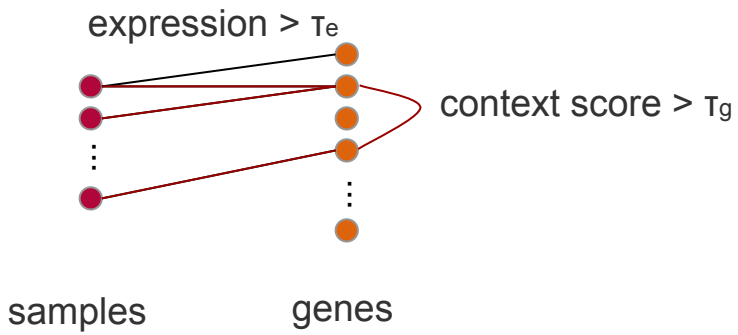
- Feature engineering: add features to support algorithm decisions

- Knowing somes genes co-occur in e.g. pathways could accelerate finding biclusters / improve the quality of the results

How to include the information?

- Bipartite graph between set of samples and set of genes
- Connection if gene expression exceeds threshold τ_e
- Gene connection if context score exceeds threshold τ_g

  → find biggest bi-cliques using Bron-Kerbosch algorithm



context score:

e.g. Are genes in the same pathway?

How to evaluate?

- Is bicluster G plausible?

  → proportion of pairs of genes in G for which

    - there exists a connecting path (precision)
    - there exists no connecting path (recall)

- StringDB offers various scores and combines them:
  - experimental score (derived from experimental data, e.g., affinity chromatography)
  - database score (derived from curated data of various databases)
  - textmining score (derived from the co-occurrence of gene/protein names in abstracts)

Based on: A systematic comparison and evaluation of biclustering methods for gene expression data by Prelić et al.

Discussion

Do you have ideas on how to improve further or evaluate better?

e.g. include other information than pathways, …

e.g. evaluate based on result of team A4:
Verification of Gene Expression Patterns in Public Knowledge Bases, ...