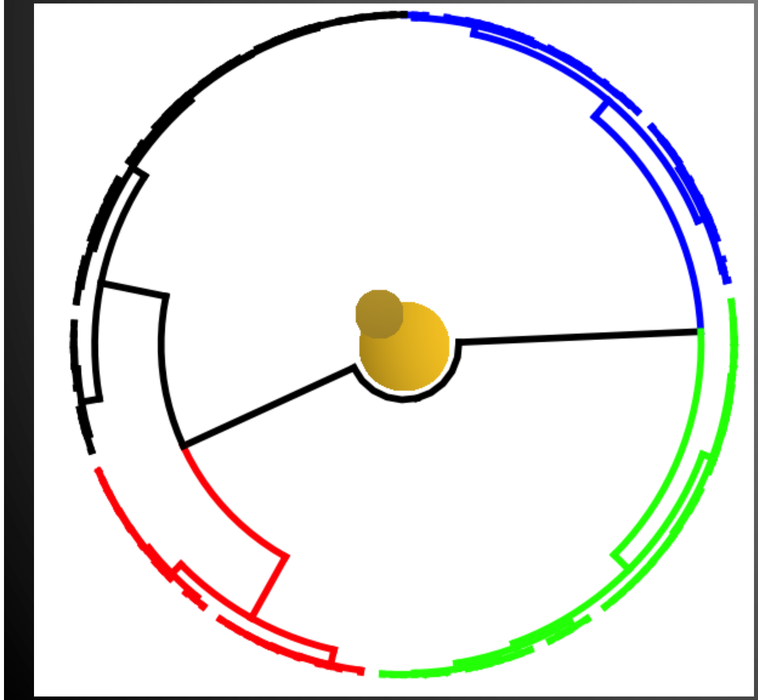
cgatatacgcgattacgcgatagcgcgagatctagcgctagcgggcgggcccctatatataaaaaaatctggctcctttaggccgcgcgcgatatctagctctct
ctagagagctctagctctagcgcttagctctcgagcgagcgagagatctagctctcttcagcgctagcgggcgggcccctatagcgattacgcgatagcgcgag
atctagcgctagcgggcgattacgcgatagcgcgagatctagcgctagcgggcgattacgcgatagcgcgagatctagcgctagcgggcgattacgcgatag
cgcgagatctagcgctagcggcgcttagctctcgagcgagcgagagatctagctctcttcagcgctagcgggcgggcccctatagcgattacgcgatagcgcg
agatctagcgctagcgggcgattacgcgatagcgcgagatctagcgctagcgggcgattacgcgatcgcttagctctcgagcgagcgagagatctagctctctt
cagcgctagcgggcgggcccctatagcgattacgcgatagcgcgagatctagcgctagcgggcgattacgcgatagcgcgagatctagcgctaaagctaga

Yeast Transcriptomics Profiler



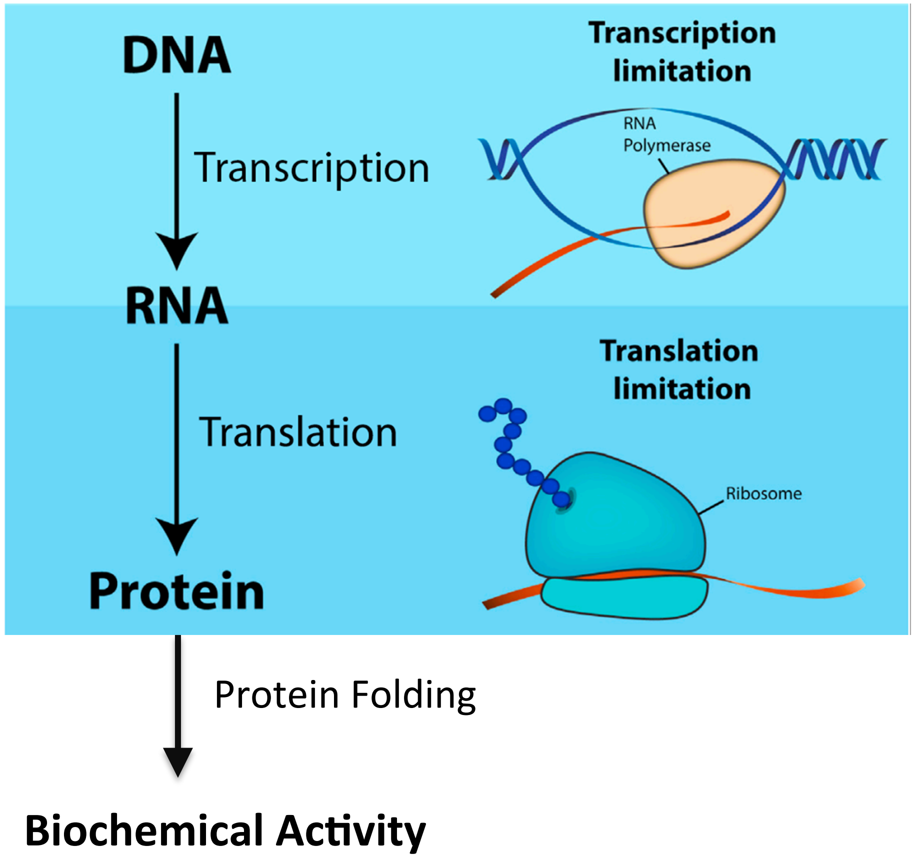
Shiny Project

NYC Data Science Academy

Ryan Willett

4/23/2019

Central Dogma of Molecular Biology



DNA

Transcription

RNA

Translation

Protein

Protein Folding

Biochemical Activity

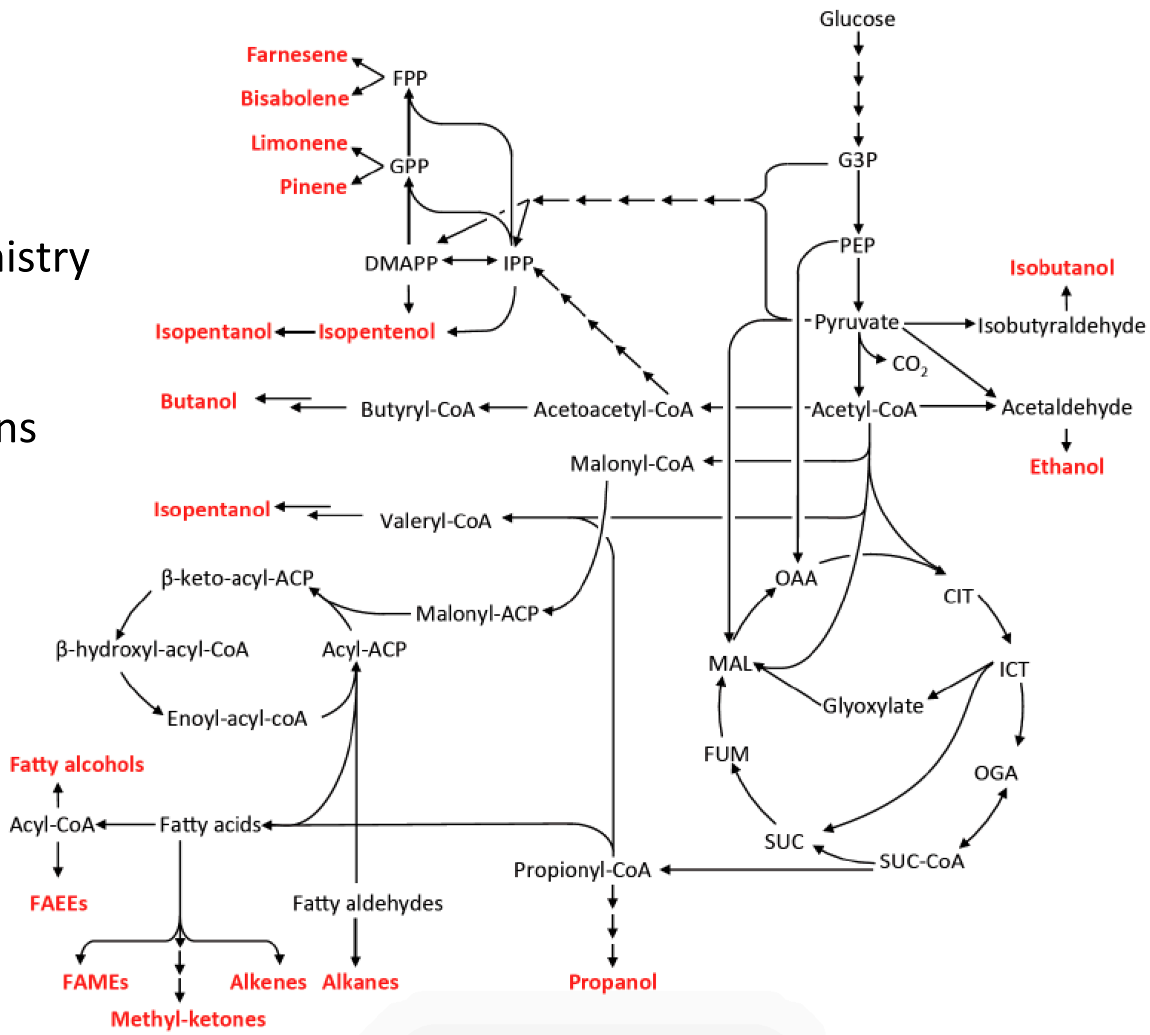
- **DNA** is a “hard-copy” template
- **mRNA** is a “temporary template” from which to produce proteins
- **Proteins** are the biochemical workhouses of cellular metabolism, catabolism and biochemistry

- There is an association between:
  - The number of RNA transcripts
  - The number of protein products of that transcript
  - The total biochemical activity of the gene products

Adapted from: Kafri, M., Metzl-Raz, E., Jona, G., & Barkai, N. (2016). The Cost of Protein Production. Cell Reports, 14(1), 22–31. doi:10.1016/j.celrep.2015.12.015

Utility of Understanding Gene Expression

Understanding the biochemistry of organisms enables optimization of their use in synthetic biology applications



Kang A, and Lee TS. *Bioengineering* 2015, 2, 184-203; doi:10.3390/bioengineering2040184

Project Goals

Analysis of RNA expression data on 6000 genes from 92 sets of yeast RNAseq next generation sequencing samples

- Display the significant differences in gene expression between experimental conditions or mutant backgrounds
- Clustering analysis to ascertain similarities in gene expression across experimental conditions or mutant backgrounds
- Identify features associated with enriched gene sets (e.g. subcellular localization, molecular function, etc)

Dataset Questions

Analysis of RNA expression data on 6000 genes from 92 sets of yeast RNAseq next generation sequencing samples

- Display the significant differences in gene expression between experimental conditions or mutant backgrounds
- Clustering analysis to ascertain similarities in gene expression across experimental conditions or mutant backgrounds
- Identify features associated with enriched gene sets (e.g. subcellular localization, molecular function, etc)

Data Analysis and Transformation

Raw Dataset Format

Sample	Gene 1	Gene 2	Gene 3	…	Gene n
Sample 1					
Sample 2					
Sample 2					
…					
Sample n					

Preprocessing ↓

$$E_{A,B} = log_2(B) - log_2(A)$$

$$X_p = -log_{10}(p - value)$$

Transformed Dataset

Genes	Category 1 Mean expression	Category 2 Mean expression	P-value	Expression ratio B/A	$E_{A,B}$	X_p
Gene 1						
Gene 2						
Gene 2						
…						
Gene n						

Manual Parsing of Gene Analysis

- High temperature:
  - HSP (YBR072W, YPL240C, YDR214W, YAL005C, YNL281W, YER103W), sporulation, cell wall metabolism, glucose metabolism (YFR015C)
- Low temperature:
  - Translation (YJL138C, YGR159C, YJL191W), transcription(YNL112W, YML043C), epigenetic regulation (chromatin methylation/acetylation), respiration, stress tolerance, hypoxia response, HSP (YPL106C), glycogen synthesis, trehalose synthesis (YML100W, YBR126C, YMR105C), unknown function, HSP (YCR021C, YDR258C), TCA (aconitase), citrate synthesis (YCR005C, YNR001C), mitochondrial function, metal ion transporters (Zn, Fe)
- Carbon source (glucose vs ethanol)
  - Gluconeogenesis, glyoxylate cycle, catabolite repression, iron homeostasis, HSP, aquaporins (water channels), TCA inhibition (YER175C), glycerol transporter, glucose transporter, TCA, acetate transporter, lipid transporter, ethanol metabolism (YMR303C), aldehyde metabolism (YOR374W), amino acid catabolism, glycolysis/glucose metabolism, lipid metabolism, mating pheromones
- Wildtype vs Biofuel Production Strain
  - Cyanimide detoxification, furfural detoxification (YNL134C), oxidative stress reduction, HSP, transcriptional regulation, alcohol dehydrogenases, unknown function, amino acid biosynthesis, amino acid uptake, amino acid biosynthesis, polyamine uptake, inositol phosphate synthesis, nucleotide synthesis, aldehyde reductase, sulfur metabolism, acetate metabolism, glycolysis, gluconeogenesis, TCA cycle
  - The upregulation of biosynthesis pathway seems to be centered on methionine production

Reference for yeast genome information

https://www.yeastgenome.org

Key Findings

Yeast in High Temperature

↑ Heat shock protein

↑ Cell wall

Yeast in Low Temperature:

↑ Transcription/Translation

↓ Metal ion uptake

Glycogen → Trehalose

Ethanol as a carbon source:

↑ Glucose production

↑ Glyoxylate cycle

↑ Water channels

↓ TCA

↓ Catabolism

↓ Metabolite uptake

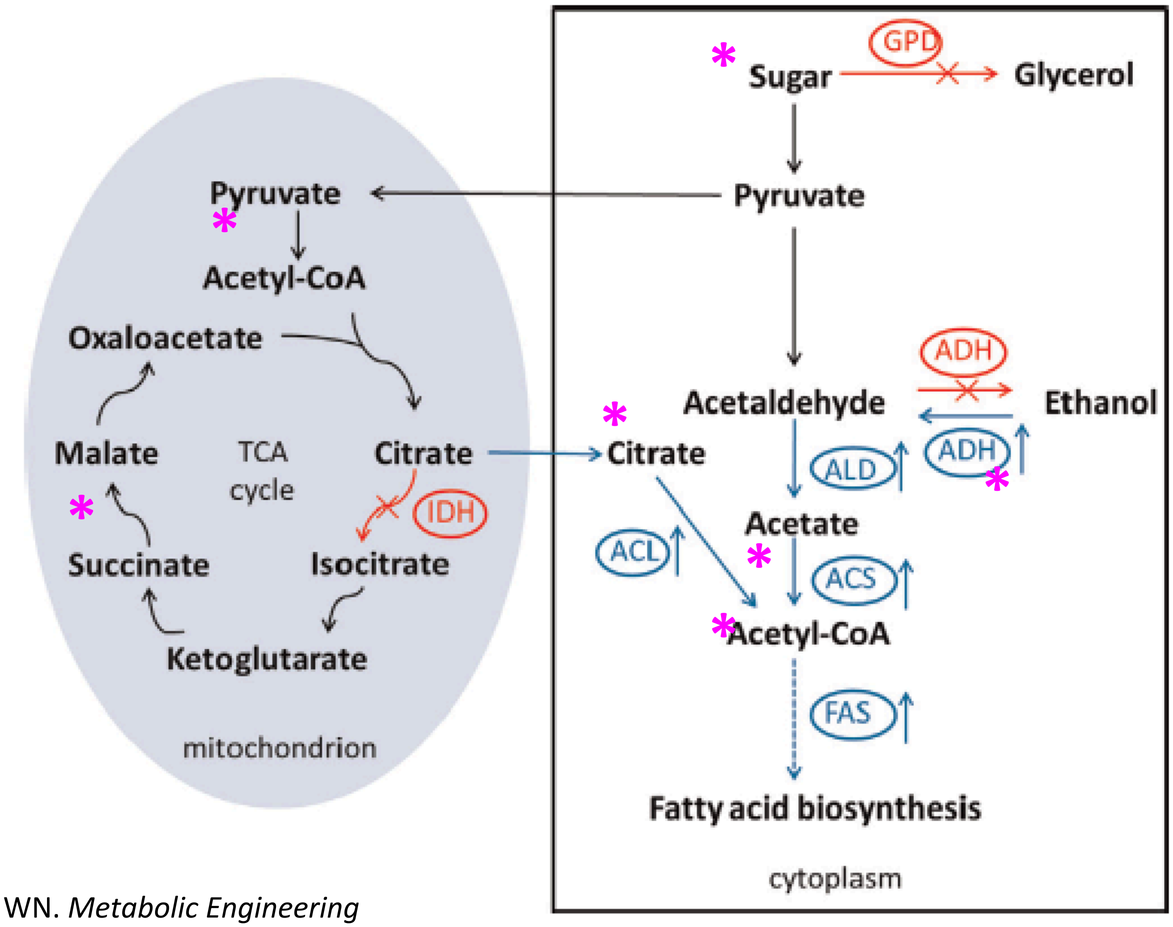
Biofuel Strains Compared to Wt

- Optimized for production of Actetyl-CoA
- Optimized for production of methionine

↑ Detoxification enzymes

↑ TCA and metabolite synthesis

Intuition about Biofuel Transcriptomics Output



Tang X, Lee J, Chen WN. *Metabolic Engineering Communications* 2(2015)58–66

Future Directions

- Include upload functionality so users can upload and browse their own data
- Gene-specific attribute data (genetic locations, molecular function, pathway analysis) will be enriched with data from web scraping gene repositories
- Relationships between significant gene expression and gene attributes will be extended by machine learning