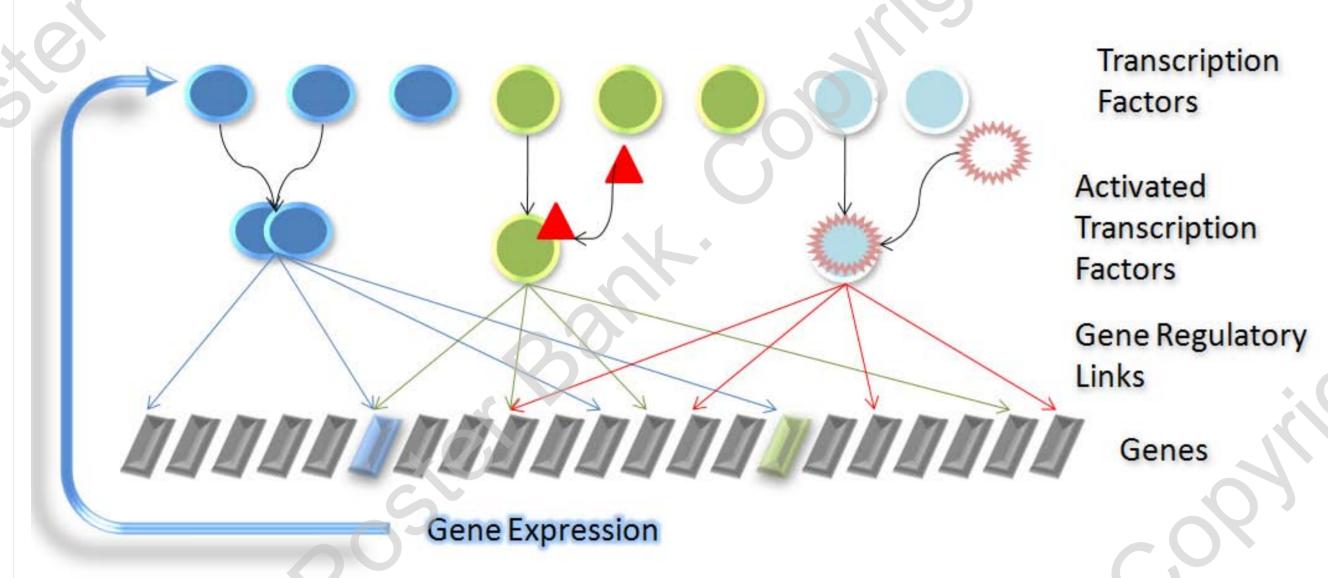
Gene Regulatory Network Reconstruction based on Gene Expression and Transcription Factor Activities

Yao (Al) Fu¹, Laura Jarboe^{1,2}, Julie A Dickerson^{1,3}, ^{1.} Bioinformatics and Computational Biology, ^{2.} Chemical Engineering, ^{3.} Electrical and Computer Engineering, lowa State University

Gene Regulatory Network Model Gene regulatory networks play essential roles in living organisms



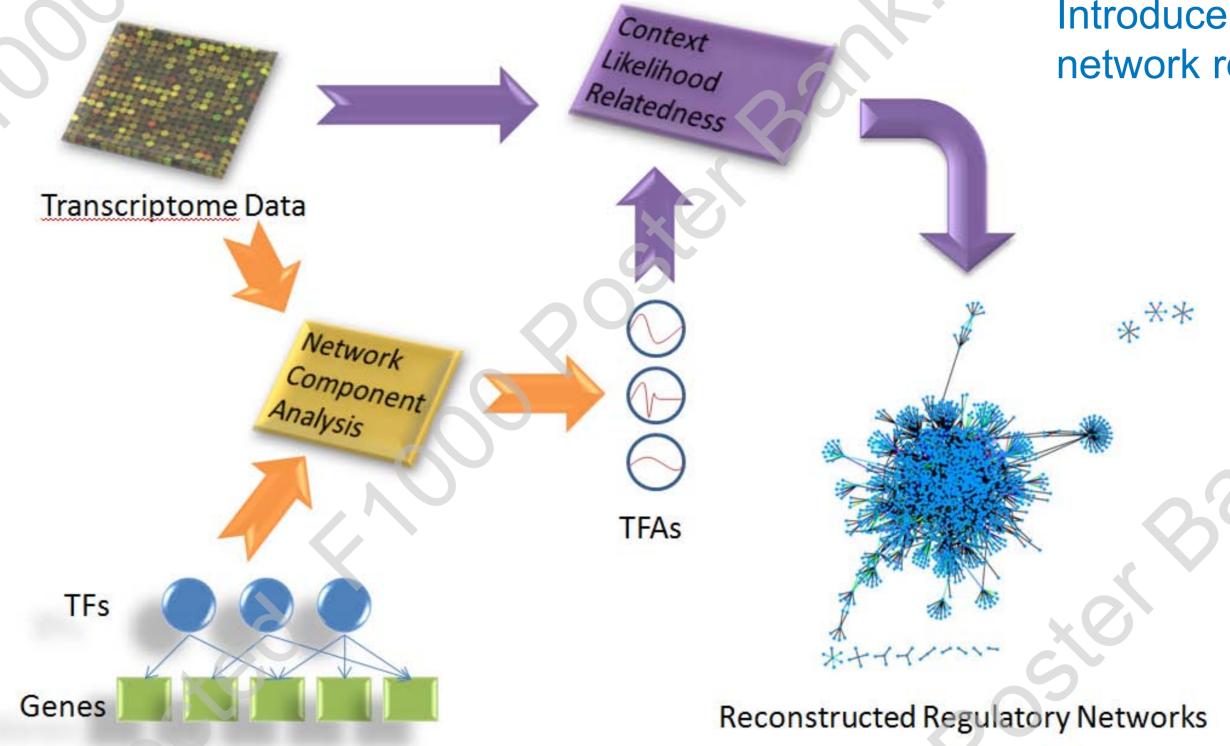
to control growth, keep internal metabolism running and make appropriate responses to external environmental changes. Understanding gene regulatory networks is a key step towards optimizing performance in biochemical engineering.

Most transcription factors (TFs) are regulated at the protein level, such as binding specific metabolites, RNA, or other TFs. Protein level regulation changes the TF's ability to induce or inhibit target gene. The ability of TF proteins to regulate target genes is known as transcription factor activity (TFA).

Context Likelihood Relatedness

 $z_i = f(cdf(MI))$

 $z_{ii} = sqrt(z_i^2 + z_i^2)$



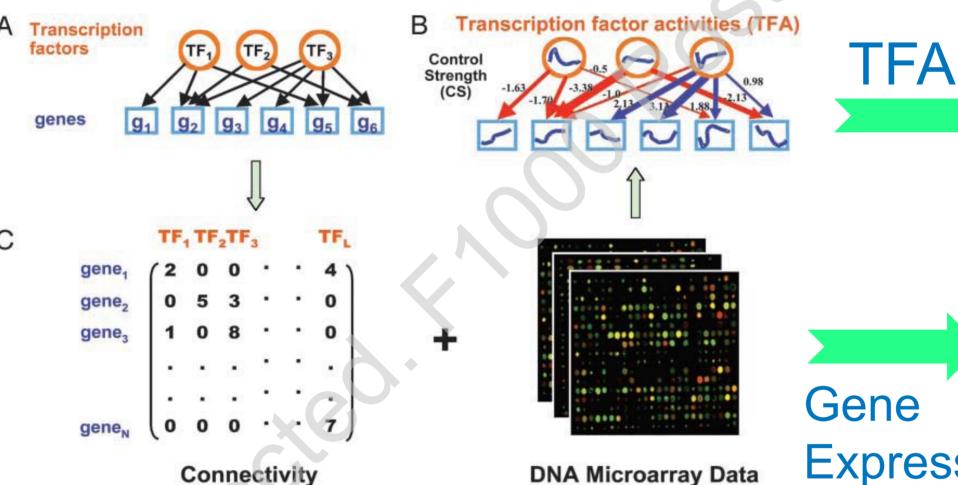
Introduce TFAs into Gene regulatory network reconstruction models

Generally, gene regulatory network reconstruction algorithms take two steps. First, a pairwise statistical relatedness matrix between all genes in the data set is computed and a minimum threshold for relatedness is chosen to get the gene co-expression network. Second, using known transcription factor information from databases or experiments, interesting gene regulatory networks are extracted from the co-expression

This work combines network reconstruction with TFA analysis to get a more complete picture of gene regulatory networks. Context Likelihood Relatedness with Gene expression and Transcription factor activities (CLR-GT). This novel gene regulatory network reconstruction algorithm introduces TFAs into gene regulatory network reconstruction models.

Algorithm Development and Test

Network Component Analysis

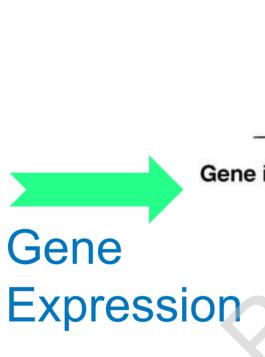


Katy C. Kao and etc., Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis (2004)

Network Component Analysis (NCA) is an algorithm to integrate known gene regulatory knowledge and gene expression data to predict TFAs under different experiment conditions. But this algorithm could not find new gene regulatory links, in other words, it could not be used to reconstruct gene regulatory networks.

Connectivity

 $z_i = f(cdf(MI_i))$



Fast NCA

NCA

NCA-r

Jeremiah J. Faith and etc., Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles (2007)

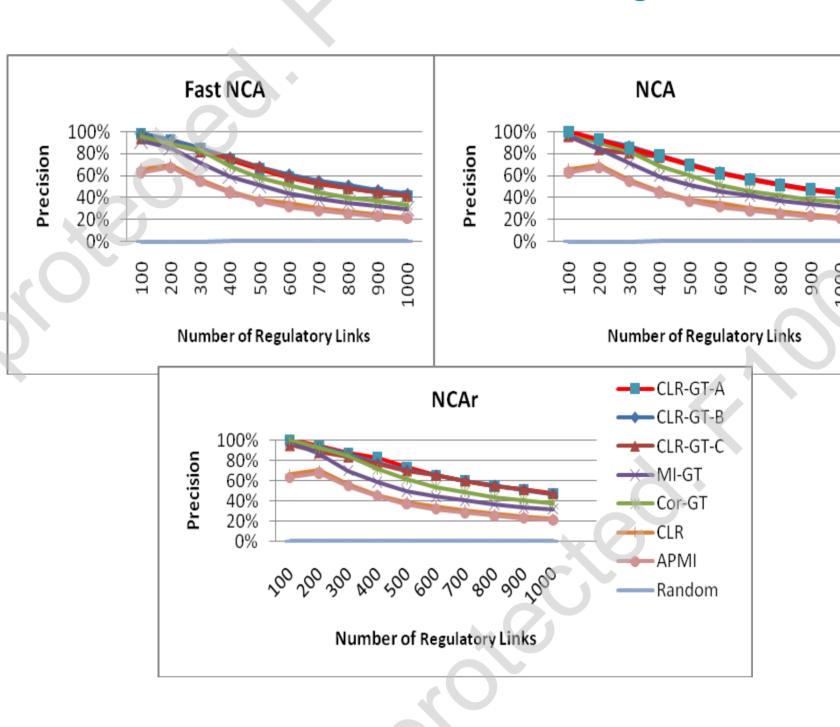
Correlation Based MI Based (B-Spline) MI Based (APMI)

M. - - - -



Known TF-Gene Network Topology

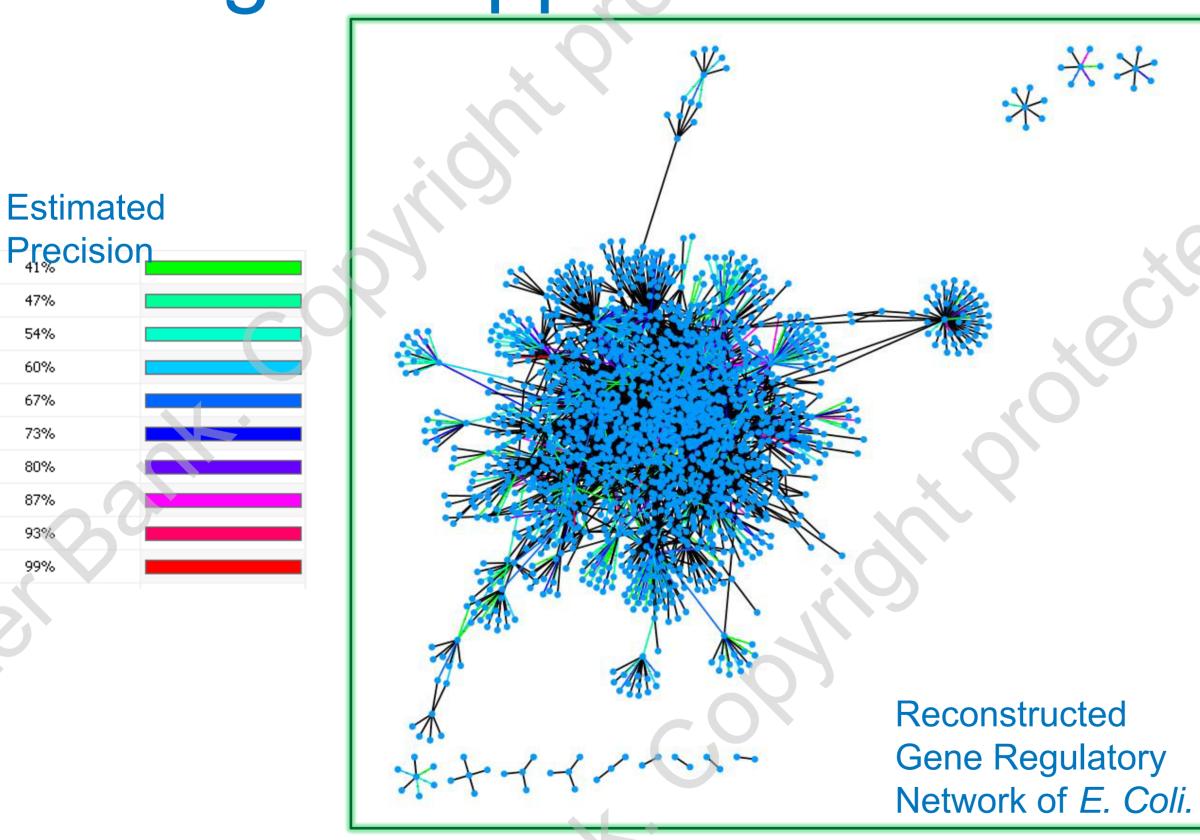
Test Results of Integrated Methods

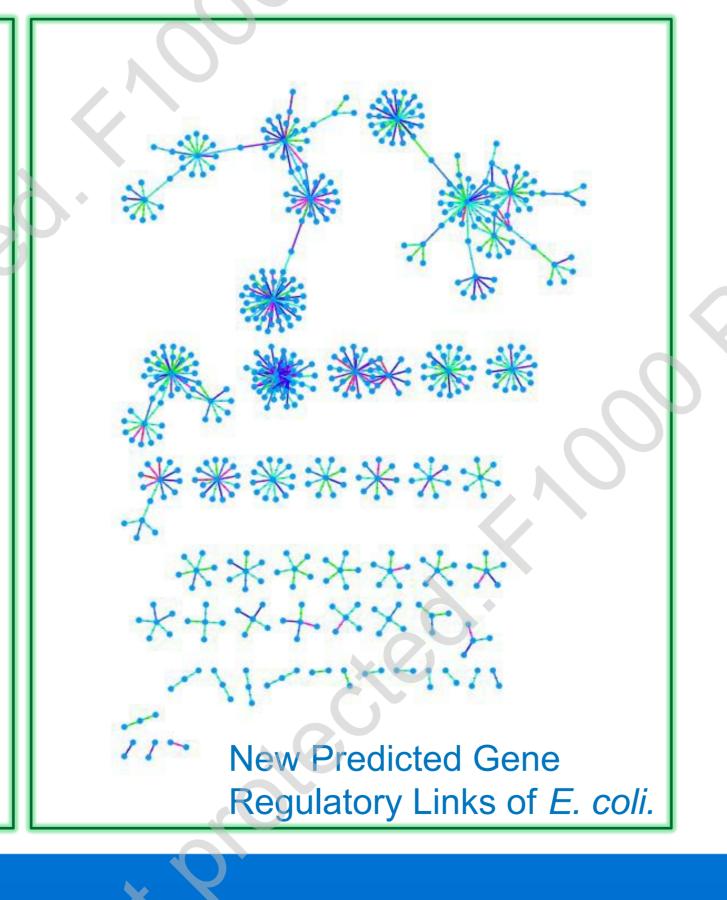


•CLR-GT-A: CLR-GT using adaptive partitioning MI as relatedness scores; •CLR-GT-B: CLR-GT using B-spline adjusted MI as relatedness scores; •CLR-GT-C: CLR-GT using correlation coefficient as relatedness scores; •MI-GT: B-spline MI between TFAs and gene expressions; •Cor-GT: correlation coefficient between TFAs and gene expressions; •CLR: Context Likelihood Relatedness: APMI: Adaptive Partitioning **Mutual Information** (GT: integrate Gene expression & TFA)

An E. coli gene expression experiment data set from the M3D database, which contains 4297 gene probes and 466 experiments, is selected as the gene expression input, and an old version of the gene regulatory links data from RegulonDB, which includes 1446 experimental verified gene regulatory links of 120 TFs of E. coli, is selected as the regulatory links input. The reconstructed gene regulatory networks are compared with the latest RegulonDB gene regulatory data (RegulonDB 6.4), which includes 3289 gene regulatory links of 166 TFs.

Biological Application





Literature Evidence of Top 12 High Precision Predictions Gene Estimated Validity Evidence

	Seme	25tillate G	, tandity 2 reduce
\sim ?		precision	
'CysB'	'yeeD'	91%	yeeD function unknown
'CysB'	'yeeE'	97%	yeeE: function unknown transport protein gene
'DcuR'	'pepE'	91%	Involve in anaerobic respiration related process (Ecocyc [18])
'Fur'	'ybdB'	91%	ybdB (entH) is proposed to be regulated by Fur (Ecocyc [18])
'Fur'	'yncE'	91%	YncE is de-repressed by Fur (Baba-Dikwa, 08 [19])
'Fur'	'yqjH'	91%	Candidate Fur box upstream of the yqjH ORF (Pania, 08 ^[20])
'GadX'	'slp'	91%	slp is positively regulated by Gad X (Tucker, 03 ^[21])
'IHF'	'sdhB'	91%	sdhB is proposed to be regulated by IHF (Ecocyc [18])
'LeuO'	'ilvH'	91%	Involve in leucine biosynthetic related processes (Ecocyc [18])
'MetR'	'metF'	91%	Involve in methionine biosynthetic process (Ecocyc [18])
'NrdR'	'fecI'	91%	Involve in regulation of transcription (Ecocyc [18])
'NrdR'	'fhuF'	91%	Unknown relationship

A set of E. coli gene regulatory networks with different numbers of regulatory links are reconstructed using CLR-GT-A. The latest RegulonDB (version 6.4) E. coli gene regulatory links are used as the initial regulatory links input, the M3D E. coli (v4, build 6) microarray experiment data is used as the gene expression input, and gNCA-r is used as the TFA inference algorithm. 566 new gene regulatory links with different estimated precisions are predicted. The biological rationale for 9 links from the first 12 precisely predicted new regulatory links is confirmed in the biological literature. And many predictions from these 566 predicted new gene regulatory links are biological significant and interesting.

