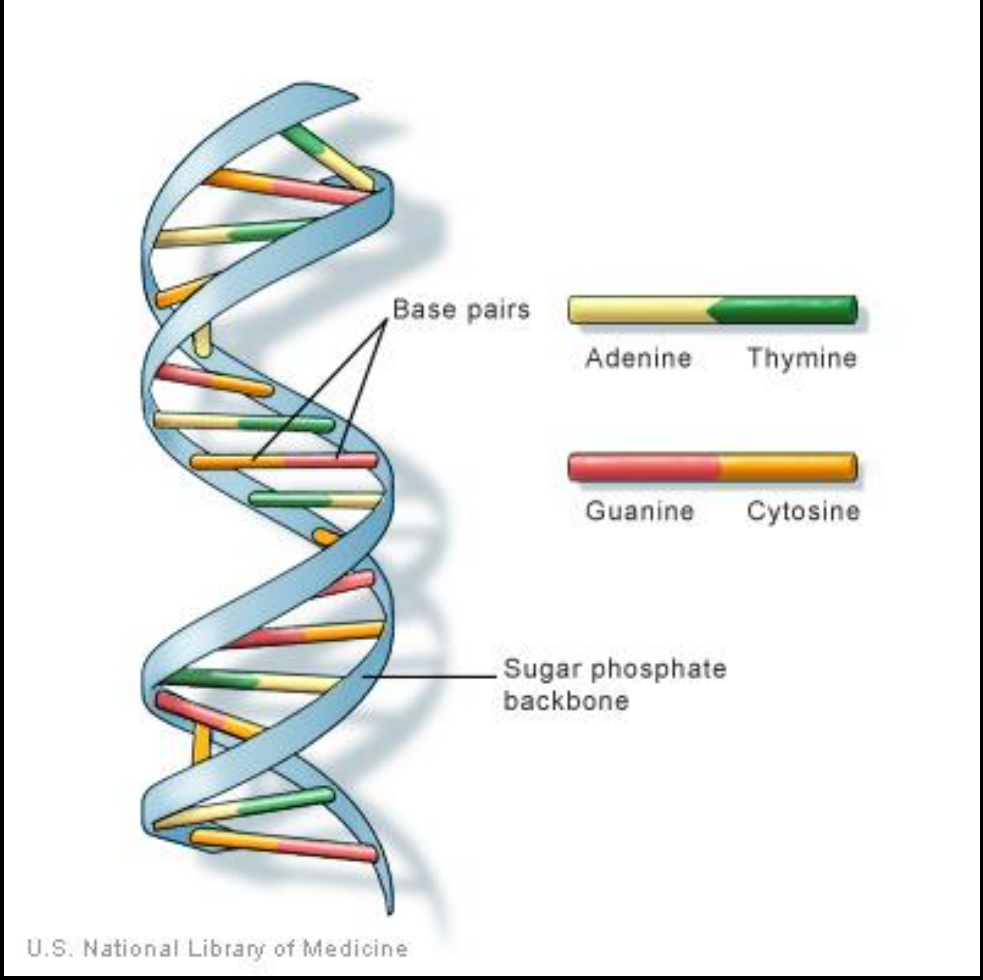
Learning Biological Regulatory Networks

Tanmay Batra
Umang Gupta
Vivek Mangal

DNA


Base pairs
Adenine
Thymine
Guanine
Cytosine
Sugar phosphate backbone
U.S. National Library of Medicine

Gene Regulatory Networks

Transcription Factor

Codon

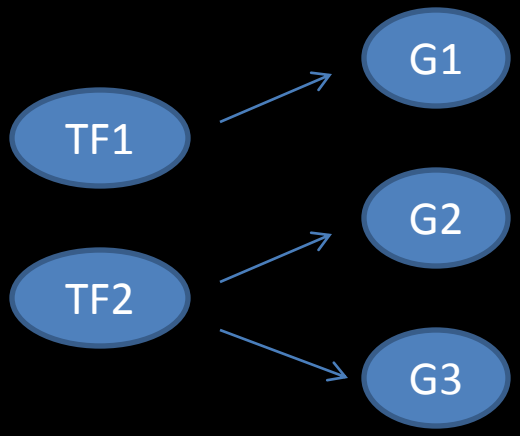
Gene

Amino acid chains

Protein

Gene expression data gives us protein levels at various conditions (or time intervals)

There are 20 amino acids associated with codons. 'Start' and 'Stop' commands are associated with some of the remaining codons.



$$\frac{d(G1)}{dt} = f(TF1, TF2, \ldots)$$

In addition to the genes (TFs), many environmental factors also affect protein levels.

Dataset used

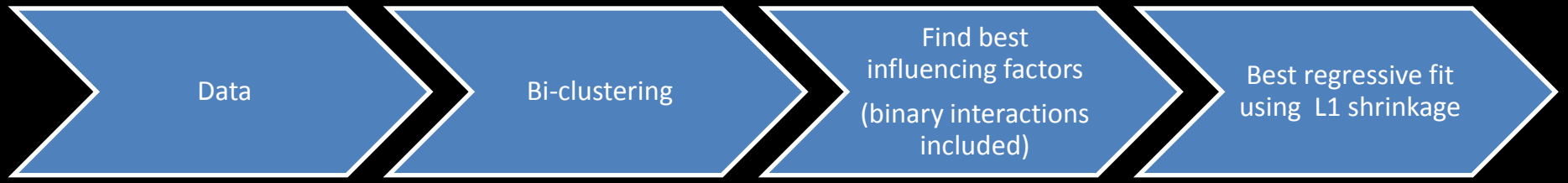
- The data contains the matrix of gene expression values for a network. Each row corresponds to a microarray chip, and each column to a gene. In other words, element (*i*, *j*) is the expression value of gene *j* in chip *i* of the compendium.
- Chip features contain meta information for each microarray chip for the network. The information is presented as a matrix, where rows correspond to chips and columns to descriptive features.
  Row *k* gives the features for row *k* of the file

- Chip features

	#Experiment	Perturbations	PerturbationLevels	Treatment	DeletedGenes	OverexpressedGenes	Time	Repeat
1	1	NA	NA	NA	NA	NA	NA	1
2	1	NA	NA	NA	NA	NA	NA	2
3	2	NA	NA	NA	NA	NA	NA	1
4	2	P1	0.5	NA	NA	NA	NA	1
5	2	P1	1.0	NA	NA	NA	NA	1
6	3	NA	NA	NA	NA	NA	0	1
7	3	NA	NA	NA	NA	NA	30	1
8	3	NA	NA	NA	NA	NA	60	1
9	3	NA	NA	NA	G5	NA	30	1
10	3	NA	NA	NA	G5	NA	60	1
11	4	NA	NA	NA	G5,G8	NA	NA	1
12	5	P2,P3	NA	NA	NA	G4	NA	1
13	5	P2,P3	NA	1	NA	G4	NA	1

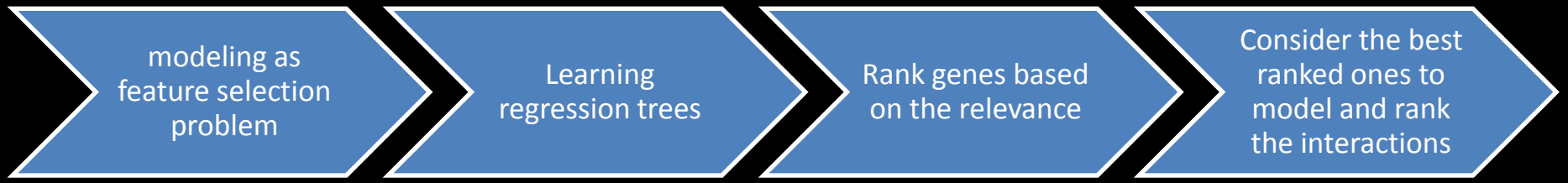
Algorithms

- Correlation (Pearson Correlation)
- Mutual Information
- Regression (differential equations approximated with difference equations)

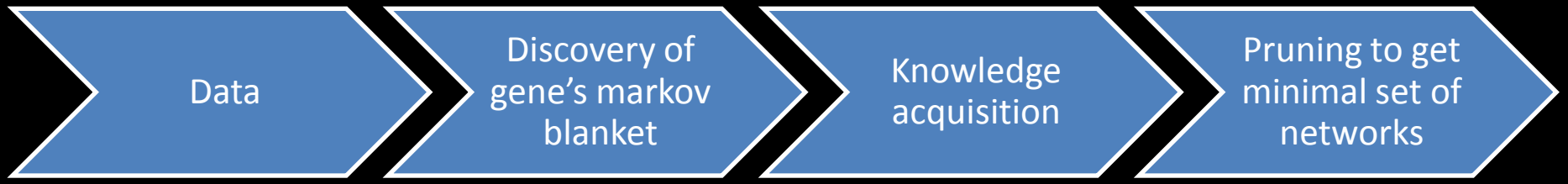


Algorithms

- Regression trees (random forests)



- Bayesian Networks (Markov blanket based model)



Results

RF based method

1	G109	G1406	0.206201
2	G48	G981	0.189174
3	G188	G938	0.181127
4	G95	G470	0.177215
5	G26	G741	0.174968
6	G49	G978	0.171681
7	G48	G1588	0.171568
8	G48	G600	0.168773
9	G158	G1434	0.168540
10	G158	G383	0.167735
11	G187	G737	0.167022
12	G48	G1398	0.165729
13	G95	G1224	0.165544
14	G84	G590	0.164282
15	G95	G1106	0.163339
16	G119	G809	0.161599
17	G84	G545	0.160625
18	G27	G57	0.158892
19	G191	G962	0.158335
20	G158	G1082	0.156724
21	G16	G687	0.156592
22	G191	G289	0.156495
23	G35	G227	0.155696
24	G48	G972	0.155225
25	G84	G427	0.154460
26	G158	G1097	0.154275
27	G126	G1050	0.154085
28	G194	G1405	0.153521
29	G48	G537	0.153375
30	G95	G1052	0.152795

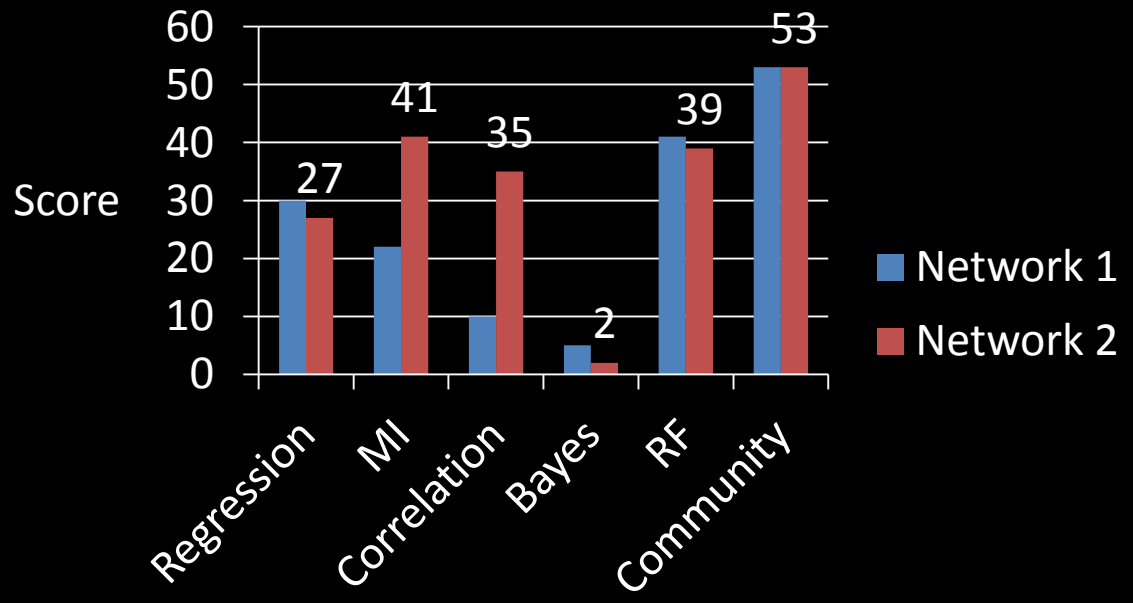
Results

- Firstly, we applied the methods individually on two networks. The performance of these methods was evaluated by using area under precision-recall curve.
- The results for one network :

	METHODS	Area under PR curve
1	Random Forests	35.41 sq units
2	Regression	25.30 sq. units
3	MI	15.81 sq. units
4	Correlation	8.08 sq. units
5	Bayes	6.62 sq. units

Results

- After that we used averaged prediction of various methods to get the best consensus network. (MI + RF gave the best results.)
- In general these “Community Networks” we far more accurate than the individual methods.



Analysis

- There is no category of inference methods that is inherently superior and that performance depends largely on

A) Data set

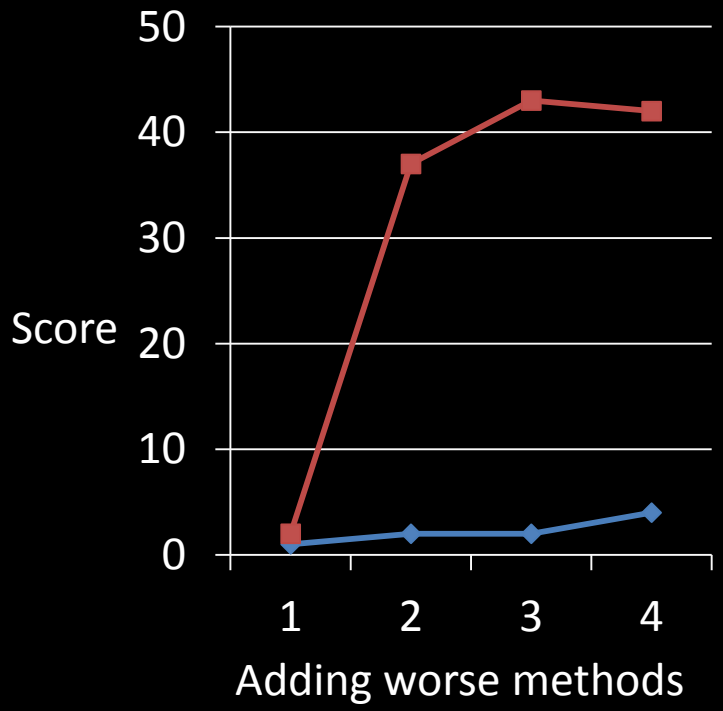
B) Specific implementation of methods

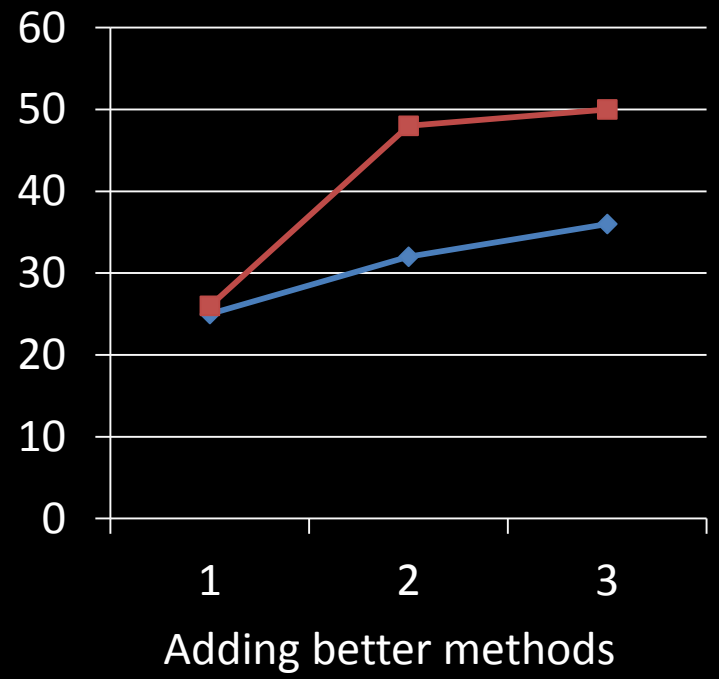
   (ex. Bootstrapping/re-sampling and L1 shrinkage gave different results)

- On an average, the community networks outperformed individual inference methods. The intuitive explanation is that the performance of individual methods does not generalize across networks (as we saw in previous analysis). Here different methods complement each other and limitations tend to be canceled out

Analysis

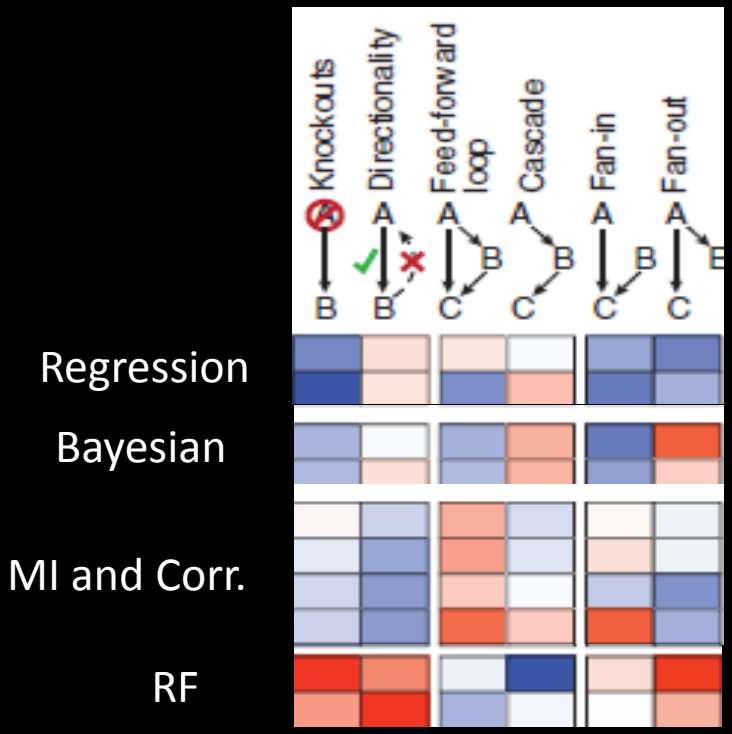
- One key feature of the community methods is shown below:





Analysis

- Let us see how method-specific biases influenced the recovery of different connectivity patterns.



Dark red: Max confidence
Dark blue: Least confidence

Analysis

- We can observe that feed-forward loops were recovered most reliably by mutual-information and correlation-based methods, whereas regression and Bayesian-network methods performed worse at this task.
- Linear cascades were more accurately predicted by regression and Bayesian-network methods. This shows that current methods experience a trade-off between performance on cascades and performance on feed-forward loops.
- The best community network (MI + RF) gave an accuracy of 40% on novel interactions which is in line of our estimate of 50% precision based on known networks. There was a large variations in case of individual methods (from 2% to 23%)

Thank You !!!