Sharif University of Technology
Computer Engineering Department
Artificial Intelligence Group

Learning and Associating Phenotypic Behavior of Organisms using Biological Data

By:
Aslan Mehrabi

Supervisor:
Dr. Seyed Abolfazl Motahari
Dr. Hamid Beigy

My background

- Bachelor
  - Shiraz University
  - Major in Computer Engineering
  - Minor in Software

✓ *Algorithms & Data Structures*

✓ *Optimization Methods*

✓ *Discrete Math & Graph Theory*

- Master
  - Sharif University of Technology
  - Major in Computer Engineering
  - Minor in Artificial Intelligence

✓ *Bioinformatics*

✓ *Machine Learning*

- Data Scientist
  - Digikala Co.

✓ *(Big) Data Analysis*

✓ *Optimization Problems*

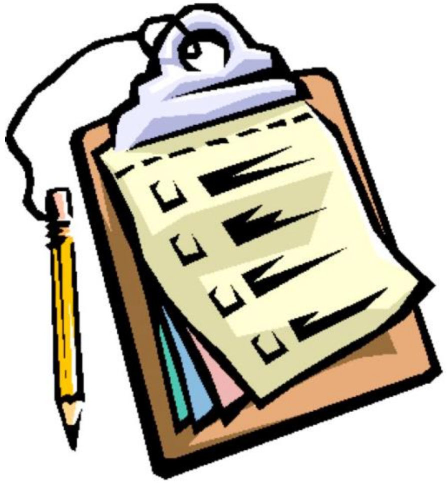
- Bioinformatician
  - Sharif Microarray Co.

✓ *Biological Data Analysis*

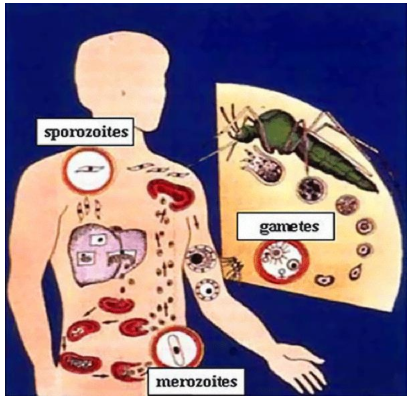
✓ *Microarray probe sequence design*

Outline

- Basic concepts
- Problem definition
- Previous studies
- Proposed method
- Results
- References



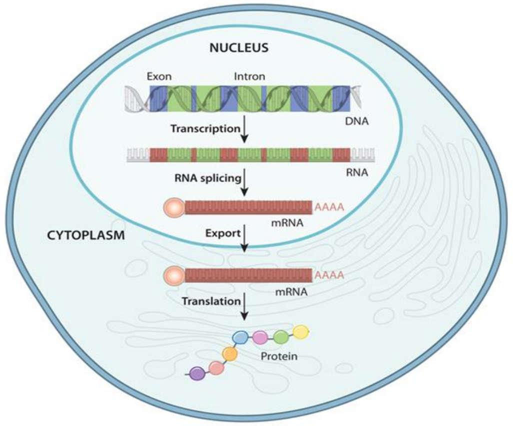
Phenotype association



Phenotype

Observable physical properties of an organism:

- Appearance
- Development
- Behavior

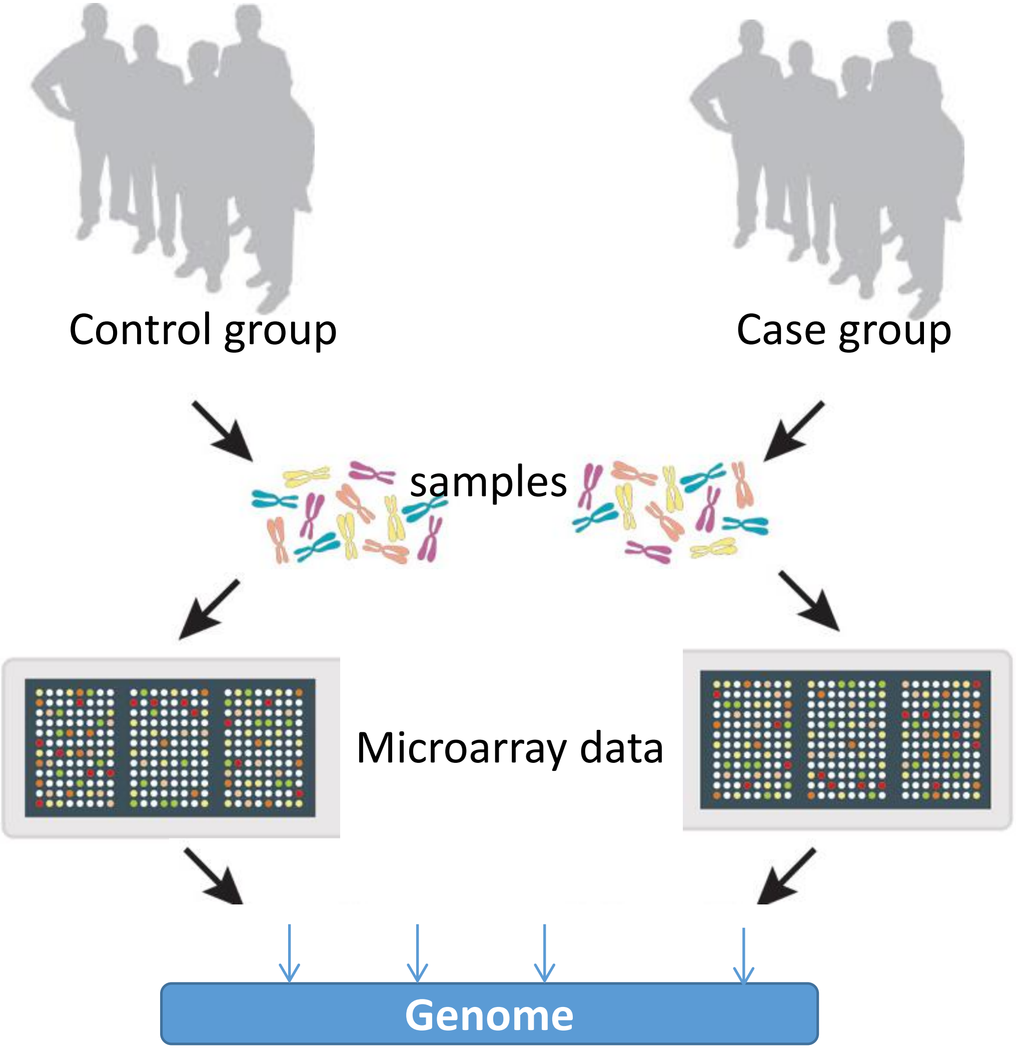


Genotype, Gene expression , ….

Phenotype association applications

- Better understanding of body defense mechanism
- Disease prediction and prevention
- Developing new medication methods

Case – control studies



- **Case group**
  - Samples having a special phenotype
- **Control group**
  - Normal samples
- **Microarray experiment**
  - Gene expression data of all samples
- Finding related genes to the phenotype
  - Difference expression patterns of case and control group

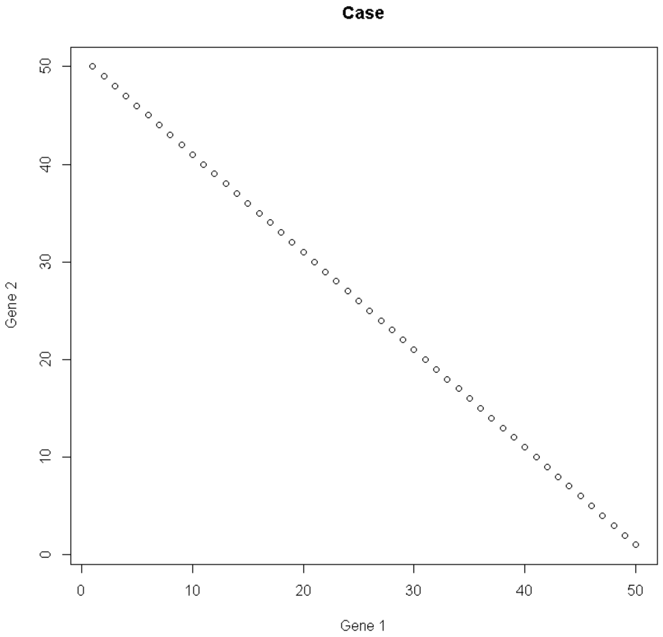
Difficulties of microarray data analysis

- Data accuracy
- Simultaneous effects of genes
- Low effect genes
- Insufficient amount of samples
    - ~ 100 samples vs ~ 40000 probes

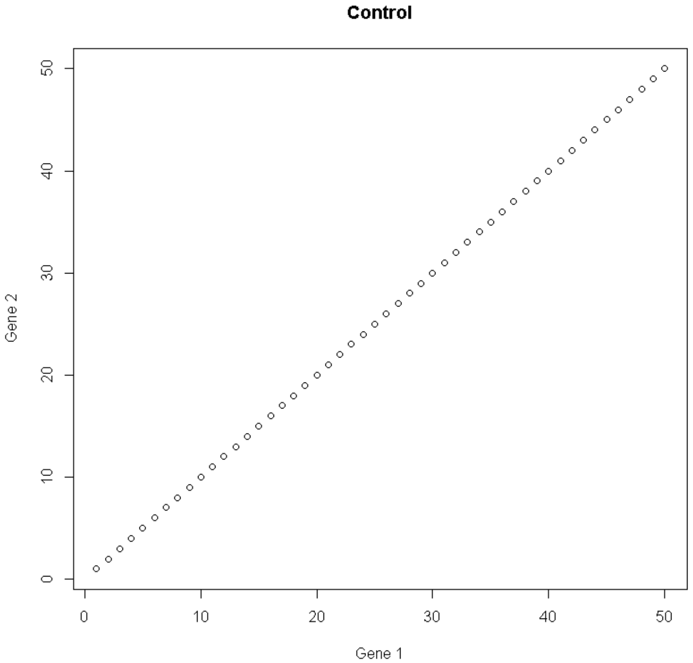
Previous Studies

	Method	Comments
1	Meta analysis	Integrating the results of former studies
2	Population stratification	Classification of the population
3	Network assistance analysis	Using the biological networks to analyze the microarray data
4	Gene set analysis	Focusing on some known gene sets
5	Dimension reduction	Reducing number of dimensions(genes or probes) to simplify the computations
6	Network clustering by microarray data	Clustering biological networks by microarray co-expression data and determine significant group of genes

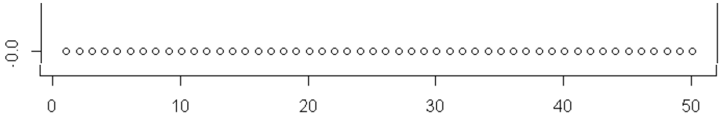
Considering the co-expression of group of genes



Expression of 2 genes
of case samples



Expression of 2 genes
of control samples



Expression of each gene (one dimension)

Challenging area

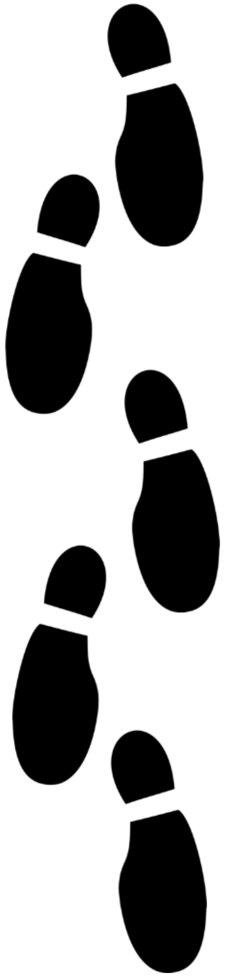
- Probes should not be eliminated based on their individual signals
- Sets of probes should be considered
- Considering all of the possible probe sets is impossible
- Additional source of data is needed to select potential effective sets of probes

Proposed method

- Integrating PPI with microarray experiment data to select sets of probes
    - Detecting the sets of probes which are supposed to be correlated
    - Reducing the problem space
    - Over-fitting prevention

- Considerations
    - The network is not specific to a special cell or tissue
    - Some of the protein-protein interactions are unknown

Steps of the proposed method



Mapping microarray probes to the PPI

Extracting effective sets of proteins form the PPI

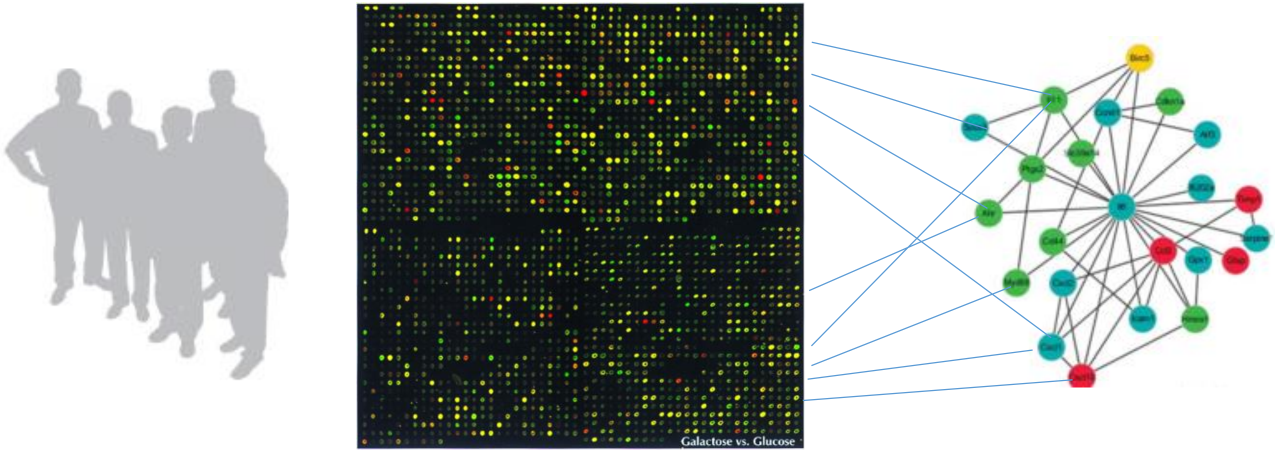
Validation of the extracted effective sets by the microarray experiment data

Comparing the chosen sets

Choosing final list of related genes to the phenotype

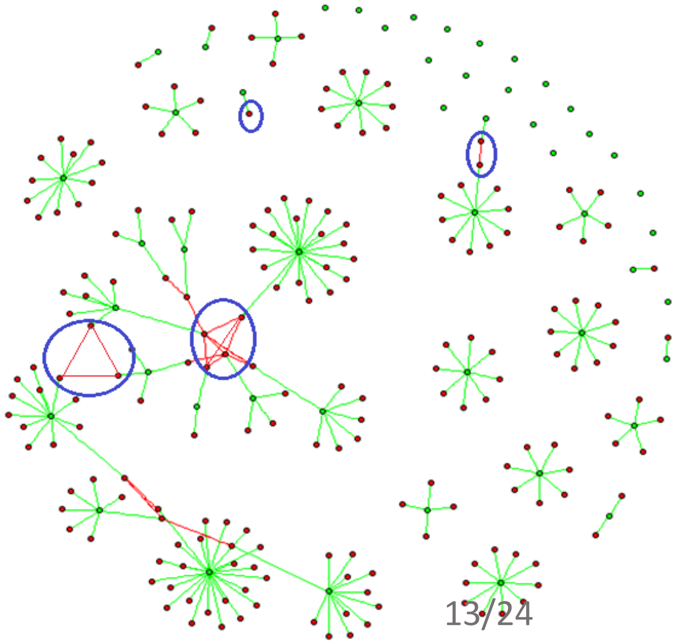
Mapping microarray probes to PPI

- Each probe should be mapped to its corresponding protein
- PPI is a weighted graph representing interactions of the proteins

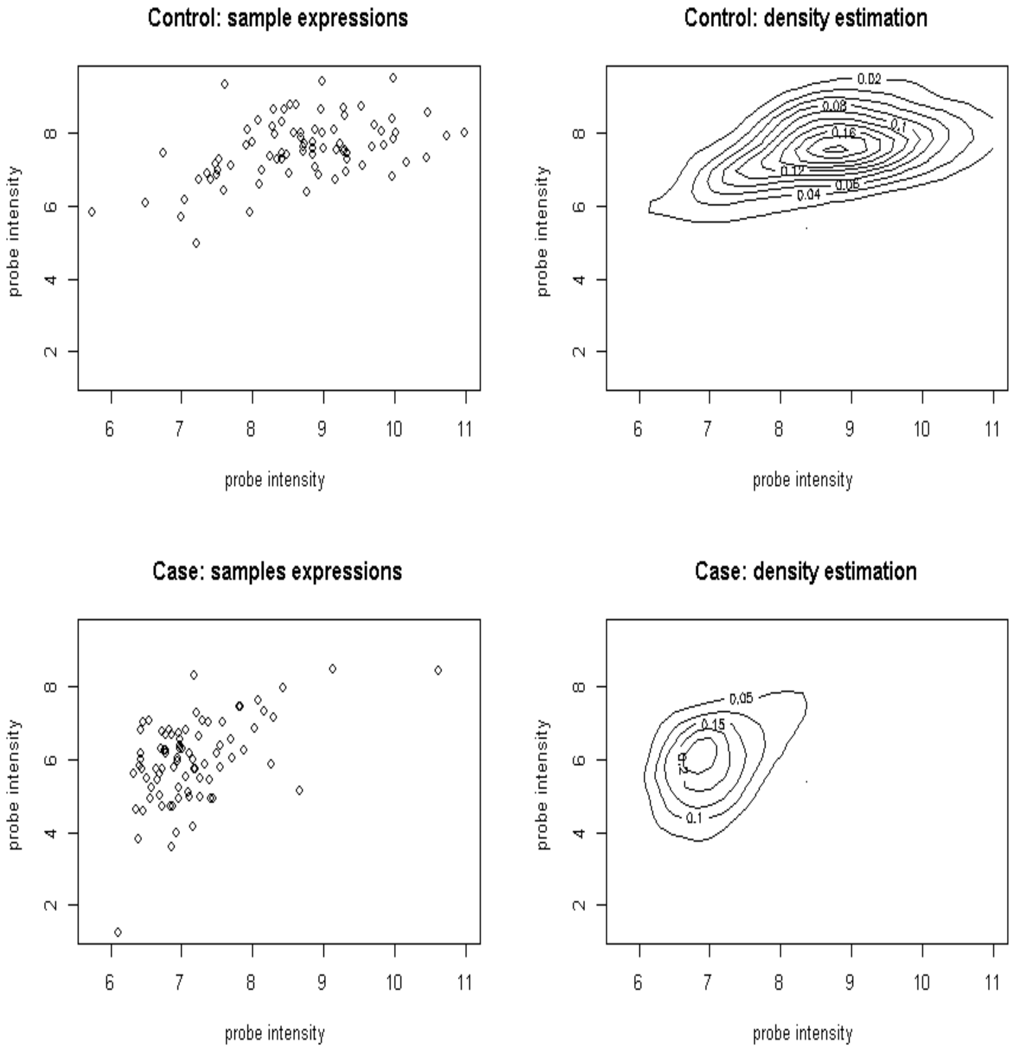


Extracting effective sets of proteins form the PPI

- Complete subgraphs with considerable weights represent effective correlations of proteins
- Protein sets with less than 5 members were selected
- Each probe is also considered as a set with one meml



Validation of extracted effective sets by microarray data

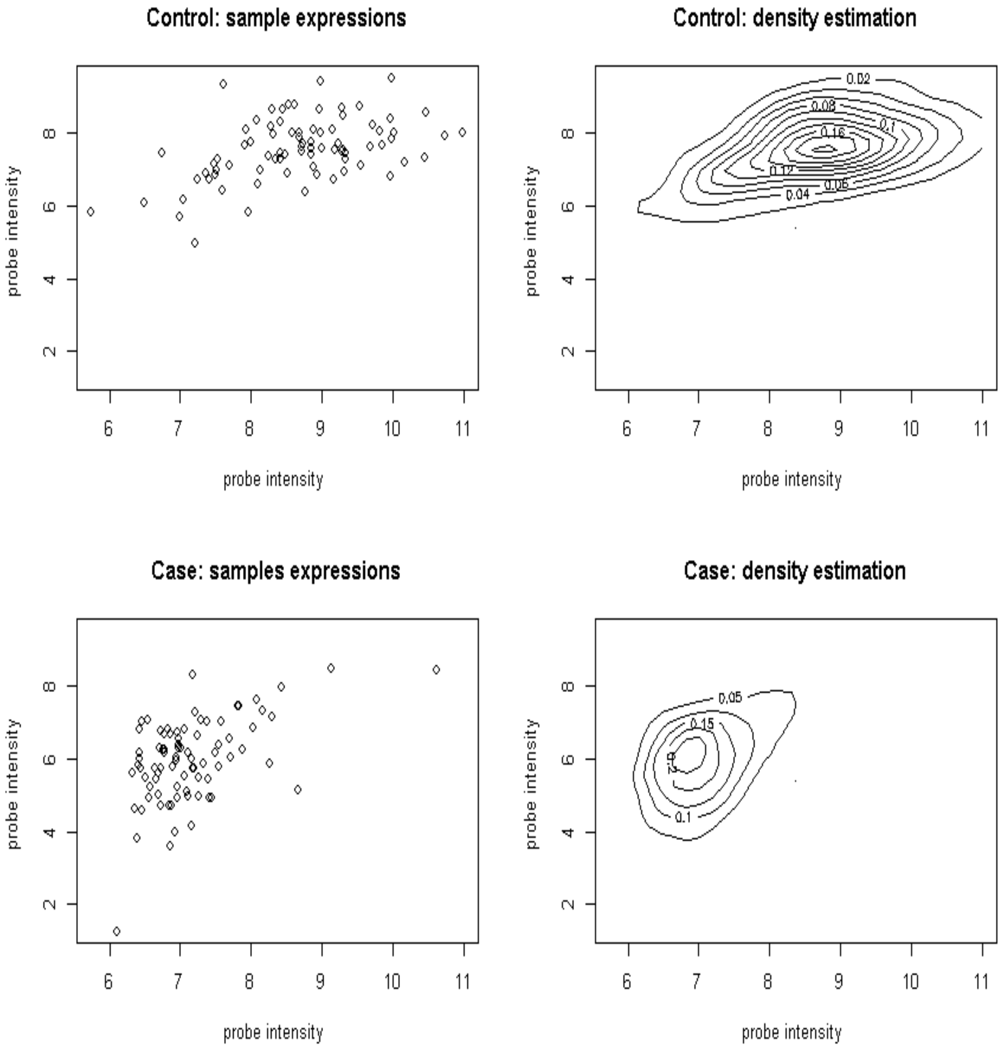


- Efficiency of selected sets of probes in separating case and control groups should be quantified by a **Distance Measure**

- **Probability density estimation** was calculated for samples of each group using Gaussian kernel based nonparametric method

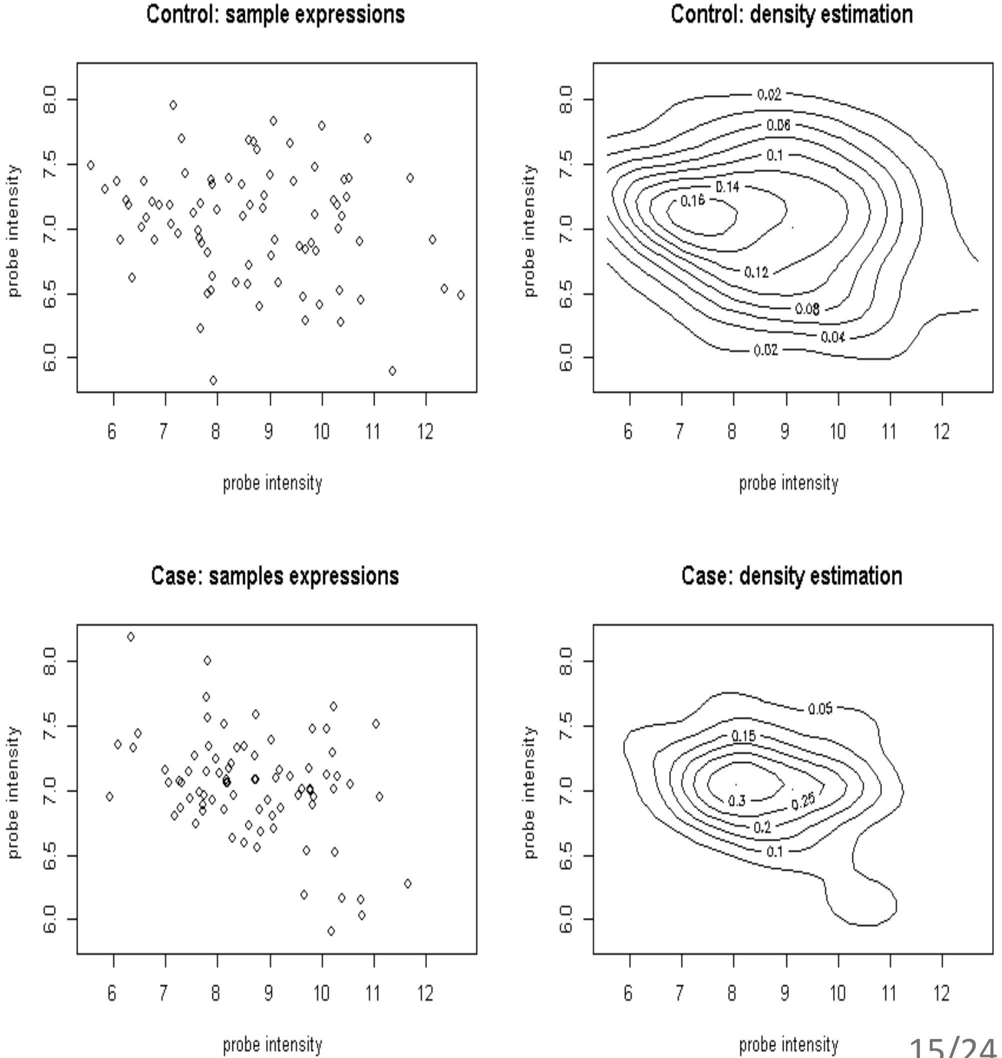
- **KL divergence** was used as the Distance Measure

Validation of extracted effective sets by microarray data

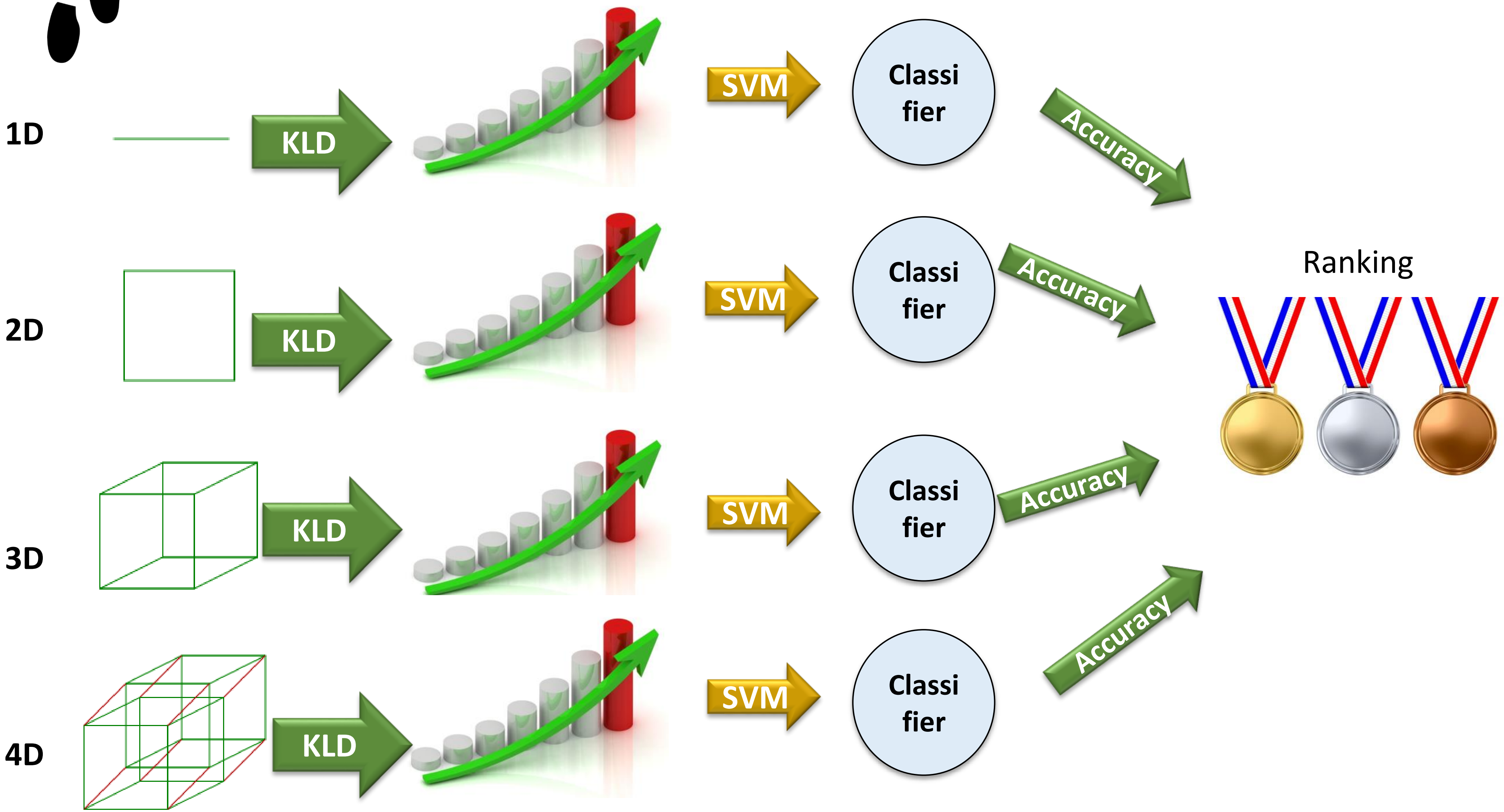


KLD

>



Comparing the chosen sets of different dimensions



1D KLD SVM Classi fier Accuracy

2D KLD SVM Classi fier Accuracy

3D KLD SVM Classi fier Accuracy

4D KLD SVM Classi fier Accuracy

Ranking

Datasets

- Agilent whole genome gene expression microarray
  - 41000 probes of length 60
- E-GEOD-54236 experiment
  - Hepatocellular carcinoma
  - 80 case samples and 81 control samples
- String PPI network
  - 86308 proteins
  - 8548002 interactions

Proof of concept

- Is PPI data integration beneficial in choosing effective sets of probes?
  - Comparison of KLD average based on threshold on interaction weight of PPI

Interaction weight Threshold	**998**	**900**	**400**	Random pair of probes (avg 1000 runs)
2D-KLD avg	**0.700**	**0.697**	**0.673**	**0.586**

Interaction weight Threshold	**901**	Random triangle of probes (avg 1000 runs)
3D-KLD avg	**1.453**	**1.035**

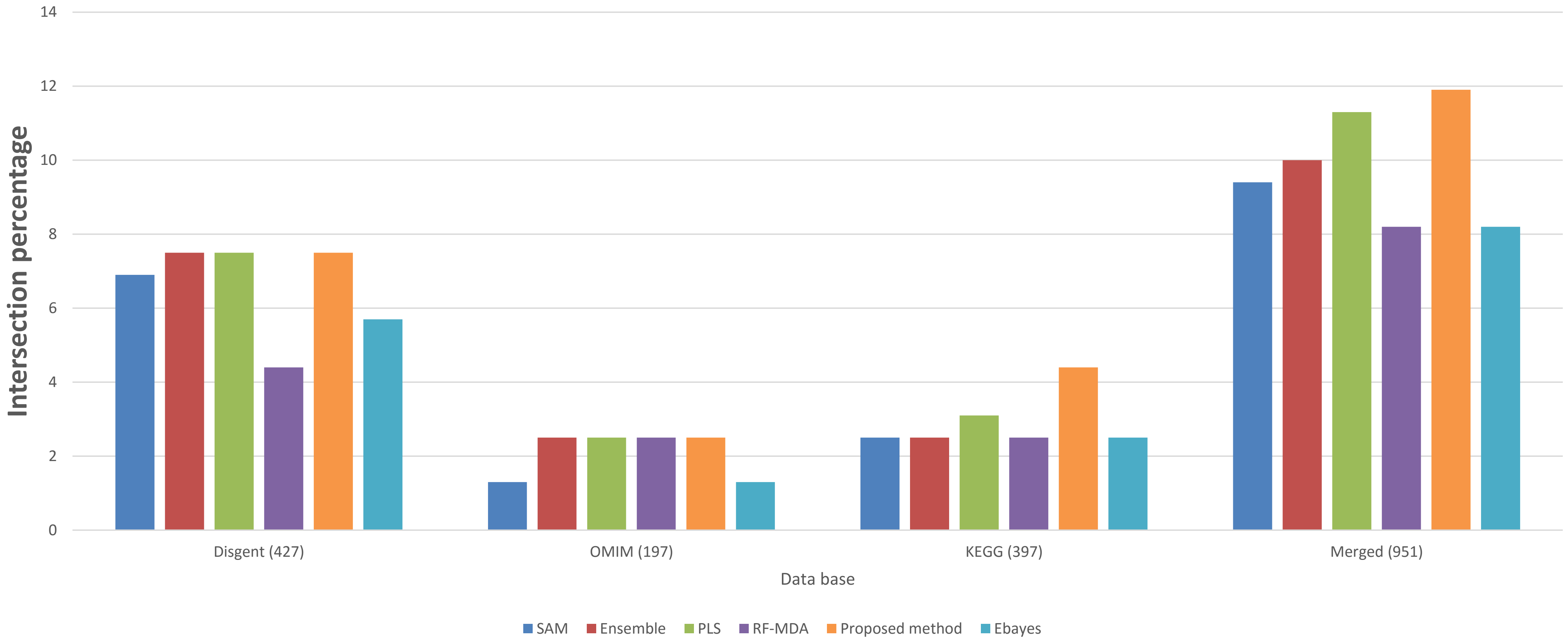
Choosing final sets of genes

- Number of chosen sets from each dimension is equal to the number of classifiers of that dimension with greater accuracy than a threshold (here 85%)

Size of set	# selected sets	# selected probes
1	2	2
2	70	76
3	58	68
4	42	51

Results

Comparing intersection of output genes of 6 methods with the reported genes of databases as effective in the disease



Results

- Most of methods omit a big portion of probes based on their individual weak signal.
- No probe were removed in preprocess here
  - Evaluation of its advantage
    - Probes which their p-value is not among lowest 10% was considered
    - Statistics of relevant considered probes to the phenotype detected by considering them as a group:

# intersected genes with merged DB	# unique probes	# patterns	Size of probe set
37	**396**	**443**	**2**
19	**215**	**277**	**3**
66	**76**	**253**	**4**

Conclusion

- Phenotype association in microarray analysis is a tough problem
- correlation of sets of genes with phenotype should be considered
- Integration of PPI data with microarray experiment were used to select related sets of genes with phenotype
- Statistical methods were used to select outstanding sets of probes
- Proposed method were able to detect specific related sets of genes with the phenotype

References

" Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease." Nature Reviews Genetics 12.1 (2011): 56-68.".

" Ellegren, Hans. "First gene on the avian W chromosome (CHD) provides a tag for universal sexing of non-ratite birds." Proceedings of the Royal Society of London B: Biological Sciences 263, no. 1377 (1996): 1635-1641.".

" Jenner, Lasse, Natalia Demeshkina, Gulnara Yusupova, and Marat Yusupov. "Structural rearrangements of the ribosome at the tRNA proofreading step." Nature structural & molecular biology 17, no. 99 (2010): 1072-1078.".

" Berk, Arnold, and S. Lawrence Zipursky. Molecular cell biology. Vol. 4. New York: WH Freeman, 2000.".

"O'Connell, Daniel J., Joshua WK Ho, Tadanori Mammoto, Annick Turbe-Doan, Joyce T. O'Connell, Psalm S. Haseley, Samuel Koo et al. "A Wnt-bmp feedback circuit controls intertissue signaling dynamics in tooth organogenesis." Science signaling 5, no. 206 (201".

"Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proceedin".

"] Jia, P., & Zhao, Z. (2014). Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. Human genetics, 133(2), 125-138.".

"Gennari, L., D. Merlotti, V. De Paola, A. Calabro, L. Becherini, G. Martini, and R. Nuti. "Estrogen receptor gene polymorphisms and the genetics of osteoporosis: a HuGE review." American journal of epidemiology 161, no. 4 (2005): 307-320.".

"] Richards, J. Brent, Fotini K. Kavvoura, Fernando Rivadeneira, Unnur Styrkarsdottir, Karol Estrada, Bjarni V. Halldorsson, Yi-Hsiang Hsu et al. "Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture.".

" Vermeeren, Veronique, and Luc Michiels. Evolution Towards the Implementation of Point-Of-Care Biosensors. INTECH Open Access Publisher, 2011.".

"] Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy et al. "Finding the missing heritability of complex diseases." Nature 461, no. 7265 (2009): 747-753.".

"] Frazer, Kelly A., Dennis G. Ballinger, David R. Cox, David A. Hinds, Laura L. Stuve, Richard A. Gibbs, John W. Belmont et al. "A second generation human haplotype map of over 3.1 million SNPs." Nature 449, no. 7164 (2007): 851-861.".

" Nooren, Irene MA, and Janet M. Thornton. "Diversity of protein–protein interactions." The EMBO journal 22, no. 14 (2003): 3486-3492.".

" Rual, Jean-François, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F. Berriz et al. "Towards a proteome-scale map of the human protein–protein interaction network." Nature 437, no. 7062 (2005): 1173-1178.".

"] Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy et al. "Finding the missing heritability of complex diseases." Nature 461, no. 7265 (2009): 747-753.".

"Glaab, Enrico, Jonathan M. Garibaldi, and Natalio Krasnogor. "ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization." BMC bioinformatics 10, no. 1 (2009): 358.".

"Leslie, R., O'Donnell, C. J., & Johnson, A. D. (2014). GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics, 30(12), i185-i194.".

"] Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap et al. "Genes mirror geography within Europe."Nature 456, no. 7218 (2008): 98-101.".

"] Jia, P., & Zhao, Z. (2014). Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. Human genetics, 133(2), 125-138.".

"Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy et al. "Finding the missing heritability of complex diseases." Nature 461, no. 7265 (2009): 747-753.".

"] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic acids research, 38(suppl 1), D355-D360.".

"Woo, Jung Hoon, Yishai Shimoni, Wan Seok Yang, Prem Subramaniam, Archana Iyer, Paola Nicoletti, María Rodríguez Martínez et al. "Elucidating Compound Mechanism of Action by Network Perturbation Analysis." Cell 162, no. 2 (2015): 441-451.".

Ehsan Pourabeda, Zahra-Sadat Shobbara, Aslan Mehrabi, Farzan Ghane Golmohamadia, "Reconstruction of drought responsive gene network in rice (Oryza sativa L)," in *5th Iranian conference on bioinformatics*, Tehran, (2014).

"Kang, U., Spiros Papadimitriou, Jimeng Sun, and Hanghang Tong. "Centralities in large networks: Algorithms and observations." In Society for Industrial and Applied Mathematics. Proceedings of the SIAM International Conference on Data Mining, p. 119. Socie".

Thanks for your attention