Multiscale Modeling of Biological Networks

Kara Goodman, Austen Piers
Xuan Hong Dang, Hongyuan You, Sourav Medya, Kyoungmin Roh, Prof. Ambuj Singh
Department of Computer Science & Biomolecular Science and Engineering

Introduction

A genetic network consists of gene expression levels and the underlying PPI (protein-protein interaction) network. We identify a small number of subnetwork biomarkers that predict a phenotype. The machine learning algorithms MINDS (Mining Discriminative Subgraphs) and SNL (SubNetwork Spectral Learning) operate on the datasets from breast cancer patients, liver cancer patients, and strains of *Caenorhabditis elegans* to classify samples and search for subnetwork biomarkers.

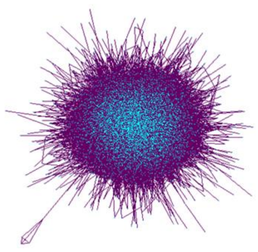
Research Topic

Our project investigated machine learning algorithms used to find the genetic source of a phenotype.

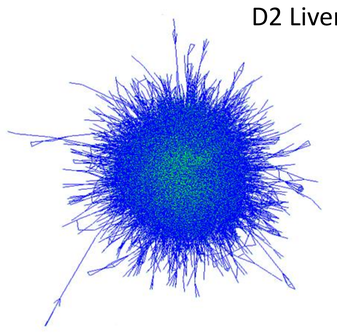
Accomplishments:

- Compared efficacy of MINDS and SNL algorithms for the purpose of subgraph discovery
- Visual interpretation of genetic network data
- Experiments on SNL results
- Contributions to code documentation

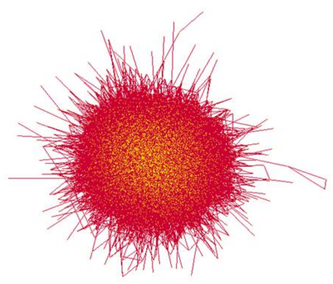
Biological Datasets



D2 Liver Metastasis



D1 Breast Cancer



D2 Liver Cancer Fully Connected

Network Structure

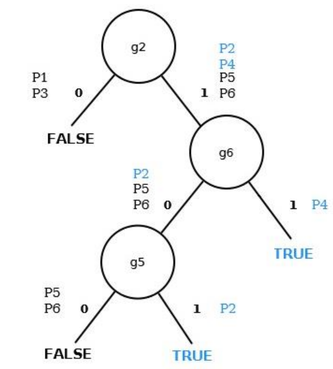
- **Node:** A single gene
- **Node Label:** Gene expression
- **Edge:** Interaction between two genes
- **Global State Network:** All possible labeled nodes and undirected, unweighted edges
- **Network Snapshot:** Patient or strain of worms
- **Network State:** Presence or lack of phenotype

SmallSample Network

Network Snapshots from SmallSample Network

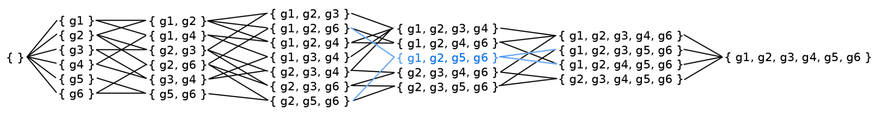
Method: MINDS

Mining Discriminative Subgraphs



NCDT on G = { g1, g2, g5, g6 } from SmallSample Network

MINDS uses Network-Constrained Decision Trees (NCDTs) as its classifier functions.
Discriminative potential of a subgraph is equivalent its NCDT accuracy.
Metropolis-Hastings algorithm is used to sample the search space and find more discriminative subgraphs.
The next subgraph is selected probabilistically from the best available paths on the edit map.



Edit Map for SmallSample Network
Featuring G = { g1, g2, g5, g6 }

Sampling Objectives:

- **Improve classification accuracy:** Add nodes with information gain in the networks misclassified by *G*
- **Don't converge to local optimums:** Delete nodes and occasional negligible node additions
- **Find the most compact subgraph:** Deletions considered more for subgraphs with high discriminative potential

Results:

- **69% - 83% average classification accuracy**
- **Inconclusive biological feature selection due to scale of solution set**

Method: SNL

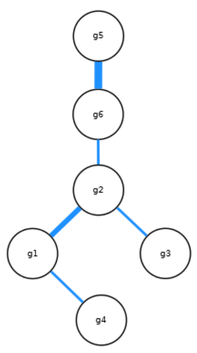
Subnetwork Spectral Learning

SNL accomplishes:

- Translation of high-dimensional data into low-dimensional data. A transformation matrix *U* for *n* nodes and *d* target dimensions is created under **network topology constraints**. Matrix *U* manipulates the original data so that snapshots are single points in *d*-dimensional space.
- Ranking of top most influential nodes (used to build discriminative subgraphs). The values of *U* correspond to node importance.

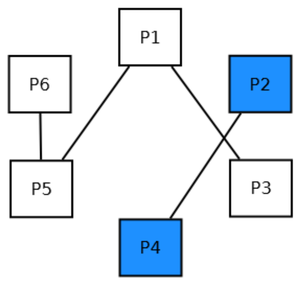
Network Topology Constraints

Retain strength of node interactions



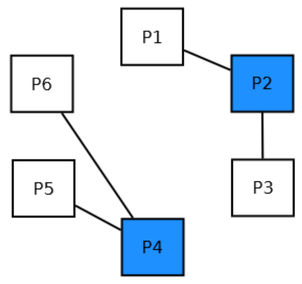
Network K on SmallSample

Minimize distance between similar snapshots with the same state



Metagraph *A+* on SmallSample

Maximize distance between similar snapshots with different states

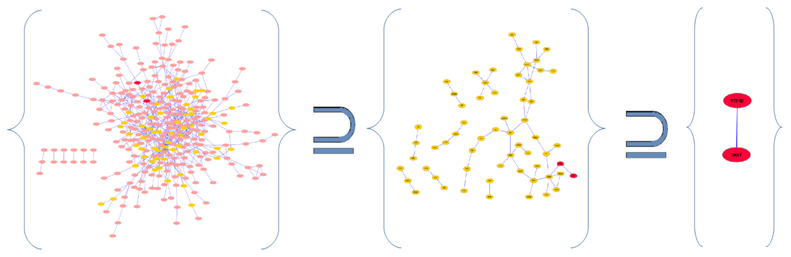


Metagraph *A-* on SmallSample

SNL Parameter Smoothing

D2 Liver Fully Connected Network:

Smoothing:	No Fold	One Fold	Two Fold
Features:	394	94	15
Accuracy:	92%	95%	83%
True Positive:	67%	89%	56%
True Negative:	95%	100%	95%



Top ranking selected genes

Breast Cancer	SEC24C,VPS28, PEG3,TNFRSF1A, CD40
C. elegans proliferation	***F48G7.8**, ***Y38E10A.3**, NHR-225, C29F9.14, C23G10.11
Liver Metastasis	***REG3A**, MMP10, MATN1, HAL,SLN

True Negative

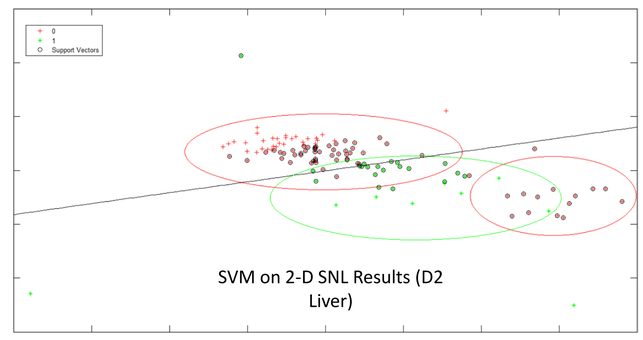
Smoothing Fold:	0	1	2	3
D1 Breast	80%	96%	100%	100%
D2 Liver Fully Connected	95%	100%	95%	100%
D2 Liver Real Values	95%	95%	100%	100%

True Positive

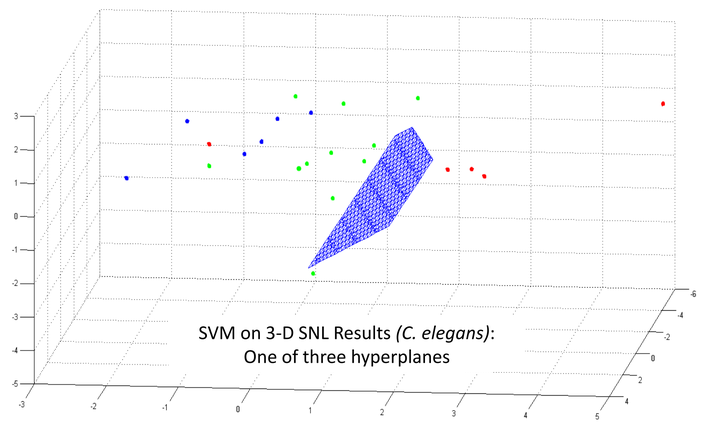
Smoothing Fold:	0	1	2	3
D1 Breast	56%	11%	0%	0%
D2 Liver Fully Connected	67%	89%	56%	0%
D2 Liver Real Values	56%	67%	0%	0%

SNL with SVM

By subspace learning, SNL allows one to visualize how the data is clustered and separated in the new, transformed and much lower dimensional space.



SVM on 2-D SNL Results (D2 Liver)



SVM on 3-D SNL Results *(C. elegans)*: One of three hyperplanes

Conclusion & Future Work

Conclusions

- MINDS and SNL+SVM both achieve high accuracy as classifiers
- MINDS is inconclusive in selecting significant biomarkers
- SNL has success in selecting significant biomarkers

Going Forward

- Confirmation of experimental results with further existing biological studies
- Mining the most influential node interactions from extensive MINDS output
- Additional research and experimentation with the complex *Caenorhabditis elegans* datasets

References
Xuan Hong Dang, Ambuj K. Singh, Petko Bogdanov, Hongyuan You, Bayyuan Hsu. "Discriminative Subnetworks with Regularized Spectral Learning for Global-state Network Data."
Minh Hoang, Ambuj K. Singh, Sayan Ranu. "Mining Discriminative Subgraphs from Global-state Networks." *KDD*. 509-517. 2013.
Dong Hyuk Ki, et al. "Whole Genome Analysis for Liver Metastasis Gene Signatures in Colorectal Cancer." *Int. J. Cancer*: 121. Wiley-Liss Inc, 2007.
Michael Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, Trey Ideker. "Cytoscape 2.8: new features for data integration and network visualization." Bioinformatics. 2011 February 1; 27(3): 431–432. December 2010.
Laura J. van 't Veer, et al. "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer." *Nature*: Vol. 415, 31 January 2002. 530-535. Macmillan Magazines Ltd, 2002.

This work was supported in part by the National Science Foundation through grant numbers IIS-0808772, IIS-1219254, and DGE-1258507.