Malik Yousef 2017

Research Statement

My Goal:

Crowd-sourcing gene annotation and pathology using an algorithm guided community approach.

My general research interest encompasses the development of machine learning techniques to resolve problems in bioinformatics/computational biology and text classification. I am interested in applying these computational tools to the analysis of complex and varied biomedical data, in order to establish gene-based diagnostic tests and therapeutic strategies for improving public health, by helping to understand the genetic foundation of diseases. These discoveries will further enable the development of unifying global perspectives and principles in biology that can be applied to the advancement of medical research. My postdoctoral work at the Wistar Institute Cancer Center provided me with the opportunity to work with a number of research groups on clinical projects ranging from the development of cancer diagnostics and identifying potential targets to studies on infectious diseases including extensive collaboration with investigators working on HIV. During this time, I also developed a research program focused on identifying miRNAs and their targets.

Within bioinformatics/computational biology, my present research can be divided into two basic domains:

- MicroRNA gene and target identification. Studying miRNAs and their targets is an important area of research because of their role in gene expression regulation.
- Highly collaborative coding and non-coding gene expression analysis as they relate to biomedical studies. My extensive experiences collaborating with biomedical researchers at all levels put me in a unique position to work closely within a diverse groups of bench scientists from varied disciplines as well as computational researchers

The new project (big project) that I am planning to develop for gene expression analysis is called “Crowd-sourcing gene annotation and pathology using an algorithm guided community approach”. The aim of the project is allowing human involvement and the wisdom of the crowd that will enhance the understanding of the outcome of the bioinformatics tool. The user will then be exposed to a variety of opinion about his data based on other researcher’s involvement. Additionally the system will be based on different biological database that will be integrated in order to get the optimal list of significant information about genes for a specific disease.

The aim of my research in the microRNA field is to combine all my research in miRNA in one **integrated system** that will find the hidden message in the microRNA sequences and their targets.

My summary plan of research is the following:

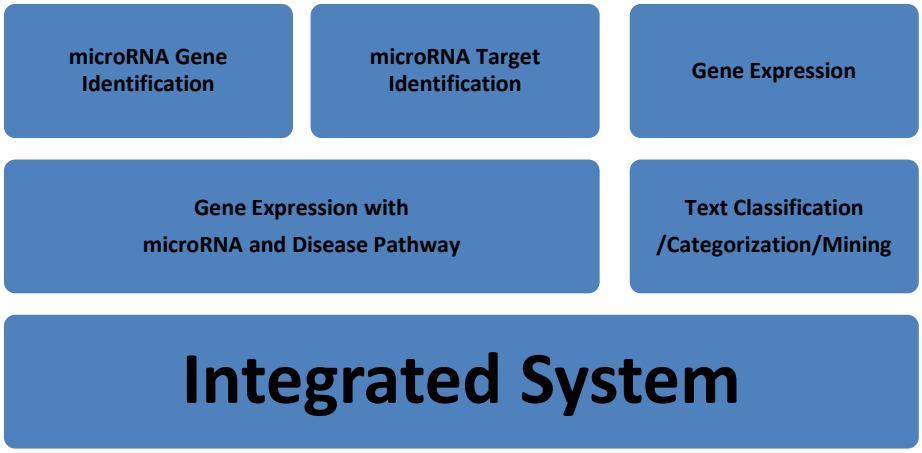
- In first two years to establish a research group with suitable infrastructure in collaboration with existing faculty that can benefit from my computational support and by applying for basic science grants to support my research and the development of new approaches to the analysis of big data sets. In support of these efforts, I will strive to continue to publish my studies in highly ranked journals and to present my data at important conferences.
- I will continue to develop my skills working on big data and text mining and data integration for miRNAs expression to be better understand the microRNA role in

gene expression and especially in cancer.

- I am also interested in applying these computational tools to the analysis of complex and varied biomedical data, in order to establish gene-based diagnostic tests and therapeutic strategies for improving public health. Our work on plant miRNAs will have interesting applications in understanding their roles in crop yields and adaptation to various environments.

Current and Past Research

The Diagram of my Research



One-Class for Text classification

The study of developing one-class for text classification was part of my Ph.D. thesis. The idea was develop a classifier that able to classify documents based on the positive examples only. Interestingly this study become the basis of other future studies that has about 980 citation. Examine the history of the citation based on Google scholar one can clearly see the importance of this study in the field. Although developed in the context of text mining I have expanded these studies and continue to use and develop using one-class classification in a variety of biological studies

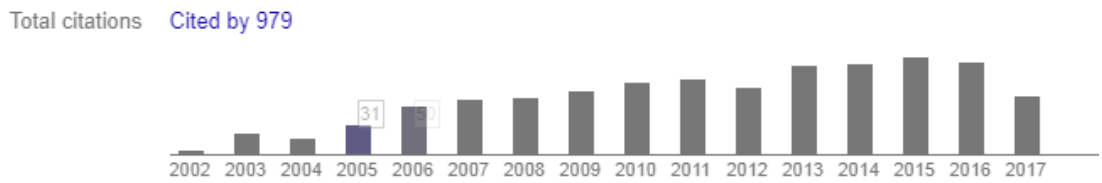


Figure 1. The citation over years of One-class SVMs for document classification LM Manevitz, M Yousef - Journal of Machine Learning Research, 2001, cited by 979 - Related articles.

Current and Past Research in Text classification

Past: applied on Reuters Data Set.
One-class SVMs for document classification[1]
One-class document classification via neural networks[2]
A web navigation system based on a neural network user-model trained with only positive web documents[3]
Current Study: Applied on PubMed for microRNA article categorization
We just started a new project for PubMed categorization of microRNA articles based on species using text classification approaches that just published[4].

MicroRNA Gene Identification

There has been recent evidence that miRNAs play a key role in regulating several biological processes, including early development; cell reproduction, differentiation and death; apoptosis; and fat metabolism. Expression studies further indicate that miRNAs are also involved in 1) cancers, including, but not limited to, chronic lymphocytic leukemia, colonic adenocarcinoma and Burkitt's lymphoma; 2) neurological development; and 3) viral diseases.

Furthermore, within the scientific community, it has been suggested that in higher eukaryotes, the role of miRNAs' in the regulation of gene expression may be as significant as the role of transcription factors.

MicroRNA Gene Target Predictions:

Recently, several miRNA target prediction techniques have been reported. Computational approaches for identifying miRNA targets mainly use sequence complementarity, thermodynamic stability calculations, and evolutionary conservation among species to determine the likelihood of a miRNA/mRNA duplex for target identification.

Current and Past Research in microRNA

Past Research- Predicting MicroRNA using Structural and Sequence Features

Two Class Approach		One-Class Approach	
MicroRNA Genes	Target Site	MicroRNA Genes	Target Site

Combining Multi-Species Genomic Data for MicroRNA Identification Using Naïve Bayes Classifier[5]	Naïve Bayes classifier for microRNA target gene identification. [6]	Learning from Positive Examples when the Negative Class is Undetermined-microRNA gene identification. [7]	One-Class Machine Learning for MicroRNA Target Detection.
	A Comparison Study Between One-Class and Two-Class Machine Learning for MicroRNA Target[8]		A Zero-Norm Feature Selection Method for Improving the Performance of the One-Class Machine Learning for MicroRNA Target Detection[9].
			Feature Selection for MicroRNA Target Prediction: A Comparison of One-Class Feature Selection Methodologies[10]

Current Research-Predicting MicroRNA using ***Sequence Motifs and k-mers***

Two Class Approach		**One-Class Approach**
MicroRNA Genes	**Target Site**	**MicroRNA Genes**
Accurate Plant MicroRNA Prediction can be Achieved using Sequence Motif Features [11]	Distinguishing between MicroRNA Targets from Diverse Species using Sequence Motifs and K-mers[12]	Sequence Motif-based One-Class Classifiers can Achieve Comparable Accuracy to Two-Class Learners for Plant MicroRNA Detection [13]
The impact of feature selection on one and two-class classification performance for plant microRNAs [14]	Species categorization based on 3'UTR microRNA target sites using sequence features. International Conference on Bioinformatics Models, Methods and Algorithms (Submitted)	Feature Selection has a Large Impact on One-Class Classification Accuracy for MicroRNAs in Plants [15]
MicroRNA Categorization using Sequence Motifs and k-mers[16]		
Ensemble Clustering Classification compete SVM and One-Class classifiers applied on plant microRNAs Data[17]		

Categorization of Species based on their MicroRNAs Employing Sequence Motifs, Information-Theoretic Sequence Feature Extraction, and k-mers(Submitted)		

High throughput Technology - Gene Expression

The field of cancer bioinformatics is stimulating a widespread synthesis of knowledge arising from the life and clinical sciences. The complexity of the questions being addressed requires experts from diverse backgrounds to engage in close and ongoing discourse and collaboration.

This part of my research is divided into two domains: using available methods to analyze complex microarray data sets and the development of new approaches to this analysis.

Developing Computational Methods	Data Analysis
Past Research	
VISDA: an open-source caBIG™ analytical tool for data clustering and beyond [18]	Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease [19]
Recursive Cluster Elimination applied on gene expression data[20]	Quantitative PCR on 5 Genes Reliably Identifies CTCL Patients with 5-99% Circulating Tumor Cells with 90% Accuracy[21]
Classification and biomarker identification using gene network modules and support vector machines [22]	Gene expression in monocytes from chronic HIV-1
Current Research	
Selecting Significant Clusters of Genes based on Ensemble Clustering using Recursive Cluster Elimination (RCE)[23]	

Big Projects: Crowd-sourcing gene annotation and pathology using an algorithm guided community approach

References

[1] L. M. Manevitz and M. Yousef, “One-Class SVMs for Document Classification.”

pp. 139–154, 2001.

[2] L. Manevitz and M. Yousef, “One-class document classifcation via neural networks,” *Neurocomputing*, vol. In Press,.

[3] L. M. Manevitz and M. Yousef, “A web navigation system based on a neural network user-model trained with only positive web documents,” *Web Intelligence & Agent Systems*, vol. 2. pp. 137–144, 2004.

[4] M. Yousef, D. Nigatu, and L. AbdAllah, “Automatic Categorization of PubMed microRNA Target Abstracts Based on Text Classification Techniques,” *J. Appl. Bioinforma. Comput. Biol.*, 2017.

[5] M. Yousef, M. Nebozhyn, H. Shatkay, S. Kanterakis, L. C. Showe, and M. K. Showe, “Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier,” *Bioinformatics*, vol. 22, no. 11, pp. 1325–1334, 2006.

[6] M. Yousef, S. Jung, A. V Kossenkov, L. C. Showe, and M. K. Showe, “Naive Bayes for microRNA target predictions machine learning for microRNA targets,” vol. 23, no. 22. pp. 2987–2992, 2007.

[7] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, “Learning from positive examples when the negative class is undetermined--microRNA gene identification.,” *Algorithms Mol. Biol.*, vol. 3, p. 2, Jan. 2008.

[8] M. Yousef, N. Najami, and W. Khalifa, “A Comparison Study Between One-Class and Two-Class Machine Learning for MicroRNA Target Detection,” *J. Biomed. Sci. Eng.*, 2010.

[9] M. Yousef and W. Khalifa, “A Zero-Norm Feature Selection Method for Improving the Performance of the One-Class Machine Learning for MicroRNA Target Detection,” *5th Int. Symp. Heal. Informatics Bioinforma.*, pp. 45–50, 2010.

[10] M. Yousef, J. Allmer, and W. Khalifa, “Feature Selection for MicroRNA Target Prediction - Comparison of One-Class Feature Selection Methodologies,” in *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies*, 2016, pp. 216–225.

[11] M. Yousef, J. Allmer, and W. Khalifa, “Accurate Plant MicroRNA Prediction Can Be Achieved Using Sequence Motif Features,” *J. Intell. Learn. Syst. Appl.*, vol. 8, no. 1, pp. 9–22, 2016.

[12] M. Yousef, W. Khalifa, \.I. E Acar, and J. Allmer, “Distinguishing Between MicroRNA Targets From Diverse Species Using Sequence Motifs And K-Mers, Proceedings of BIOSTEC 2017, 10th International Joint Conference on Biomedical Engineering Systems and Technologies,” *Porto.*, vol. 3, 2017.

[13] M. Yousef, J. Allmer, and W. Khalifa, “Sequence Motif-Based One-Class Classifiers Can Achieve Comparable Accuracy to Two-Class Learners for Plant microRNA Detection,” *J. Biomed. Sci. Eng.*, vol. 8, no. 10, pp. 684–694, 2015.

[14] W. Khalifa, M. Yousef, M. D. Sacar Demirci, and J. Allmer, “The impact of feature selection on one and two-class classification performance for plant microRNAs.,” *PeerJ*, vol. 4, p. e2135, 2016.

[15] M. Yousef, M. D. Saçar Demirci, W. Khalifa, and J. Allmer, “Feature Selection Has a Large Impact on One-Class Classification Accuracy for MicroRNAs in Plants,” *Adv. Bioinformatics*, vol. 2016, pp. 1–6, 2016.

[16] and J. A. Malik Yousef, Waleed Khalifa, İlhan Erkin Acar, “MicroRNA

Categorization using Sequence Motifs and k-mers," *Submitted*, 2016.

[17] M. Yousef, W. Khalifa, and L. AbedAllah, "Ensemble Clustering Classification compete SVM and One-Class classifiers applied on plant microRNAs Data.," *J. Integr. Bioinform.*, vol. 13, no. 5, p. 304, Dec. 2016.

[18] J. Wang *et al.*, "VISDA: An open-source caBIGTM analytical tool for data clustering and beyond," *Bioinformatics*, p. btm290, 2007.

[19] M. K. Showe *et al.*, "Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease," *Cancer Res.*, vol. 69, no. 24, pp. 9202–9210, 2009.

[20] M. Yousef, S. Jung, L. Showe, and M. Showe, "Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data," vol. 8, no. 1. p. 144, 2007.

[21] M. Nebozhyn *et al.*, "Quantitative PCR on 5 genes reliably identifies CTCL patients with 5% to 99% circulating tumor cells with 90% accuracy," *Blood*, vol. 107, no. 8, pp. 3189–3196, 2006.

[22] M. Yousef, M. Ketany, L. Manevitz, L. C. Showe, and M. K. Showe, "Classification and biomarker identification using gene network modules and support vector machines.," *BMC Bioinformatics*, vol. 10, p. 337, 2009.

[23] L. AbdAllah, W. Khalifa, L. C. Showe, and M. Yousef, "Selection of Significant Clusters of Genes based on Ensemble Clustering and Recursive Cluster Elimination (RCE)," *J. Proteomics Bioinform.*, vol. 10, no. 8, pp. 186–192, 2017.