University of Cincinnati

Date: 9/29/2010

I, Sandhya Shahdeo, hereby submit this original work as part of the requirements for the degree of Master of Science in Computer Science.

It is entitled:
MiR-Drug Relationships: Mining and discovering bi-domain dense subclusters using greedy randomized algorithm

Student's name: Sandhya Shahdeo

This work and its defense approved by:

Committee chair: Yizong Cheng, PhD
Committee member: Anil Jegga, DVM, MRes
Committee member: Raj Bhatnagar, PhD
MiR-Drug Relationships: Mining and discovering bi-domain dense subclusters using greedy randomized algorithm

A thesis submitted to the

Graduate School
of the University of Cincinnati
in partial fulfillment of the
requirements for the degree of

Masters in Science

in the Department of Computer Science
of the College of Engineering and Applied Science

by

Sandhya Shahdeo

B.E. Birla Institute of Technology, Mesra
May 2005

Committee Chair: Dr. Yizong Cheng, Ph.D
Abstract

Gene regulation, primarily achieved at the transcriptional level, is central to the normal development and functioning of all organisms. It therefore represents an obvious target for therapeutic drugs which could act either by stimulating or inhibiting specific gene transcription to elicit desired effects. Likewise, microRNAs, the recently discovered small non-coding RNAs, negatively regulate target genes post-transcriptionally through degradation or suppression of protein translation. Since both these regulators, endogenous and exogenous, affect biological and chemical pathways targeting gene regulation, the question is whether they can be used alternatively or combinatorially. Thus, hypothesizing that microRNAs— the endogenous regulators (mostly suppressors) of genes, could be used as alternatives to drugs or as combinatorials (with drugs) to fine tune the drug-response or mitigate potential side-effects, we used existing techniques like clustering, self organizing maps and computation of nearness in n dimensional space, to analyze the correlation between drugs and microRNAs. An algorithm to incorporate the heterogeneity of data by introducing a per data-species ratio, using a randomized greedy approach, was conceived. This algorithm finds the largest square matrices complying with the drug/microRNA ratio and density threshold by randomly selecting seed matrices and then systematically growing the best ones available and converges over multiple runs.
Acknowledgement

I thank Dr. Yizong Cheng, my advisor for all the support rendered to me. Support in terms of independence in choice of area of research, guidance in the work, critical analysis of work and encouragement in all endeavors. This thesis would be incomplete without his guidance and help. I am indebted to Dr. Anil Jegga for his invaluable help and meticulous supervision of my work right from the choice of the thesis topic and conceptualization to the regular brainstorming on every issue, every problem during the course of its completion. He provided unflinching guidance and support and endured with me through every failure and success of the thesis.

I extend my gratitude to Dr. Ali Minai, whose insightful courses gave me innovative ideas pertinent to my problem at hand and also for answering all my questions with utmost patience.

I am obliged to Chunsheng (Victor) Fang a friend, ex-classmate and guide who walked me through the intricacies of MATLAB and made several valuable suggestions especially when I was losing my way.

I thank my committee members Dr. Cheng, Dr. Jegga and Dr. Bhatnagar for going through the entire thesis, making valuable suggestions and refining the quality of work.

Finally, I acknowledge my friends and family for being the strength behind me and guiding me through the roller coaster ride of research work.
Contents

1 Overview ...................................................................................................................................... 1

2 Techniques Used ...................................................................................................................... 3

2.1 Data Resources and Construction of Base Matrix .......................................................... 3

3 Scoring ....................................................................................................................................... 12

4 Drug MicroRNA Distance ..................................................................................................... 15

5 Clustering ................................................................................................................................ 17

5.1 K-means clustering ......................................................................................................... 17

6 Self Organizing Map ............................................................................................................. 20

7 Integrating the results ....................................................................................................... 27

8 Network Creation .............................................................................................................. 30

8.1 EAGLE Clustering Algorithm ....................................................................................... 30

8.2 FAG Clustering Algorithm ............................................................................................ 33

9 Greedy Randomized Algorithm ............................................................................................. 36

9.1 Dharan Nair paper .......................................................................................................... 36

9.2 Non-incremental algorithm ............................................................................................ 37

9.3 Incremental algorithm ................................................................................................... 40

9.3.1 Rectangle matrix generation ...................................................................................... 46

9.3.2 Square matrix generation ........................................................................................... 47
9.4 Limitations .............................................................................................................................................. 51

10 Results for validated data ........................................................................................................................... 52

11 Conclusion .................................................................................................................................................. 60

12 Future work .................................................................................................................................................. 61

13 References .................................................................................................................................................. 62
List of Figures

1. Determination of indirect relations ................................................................. 8
2. Cell values for matrix based on gene targets ...................................................... 10
3. Cell values for matrix based on associated pathways ........................................ 10
4. Drug/MicroRNA vs gene/pathway matrix .......................................................... 11
5. Heat map for matrix for anti-cancer drugs and microRNAs ............................... 11
6. Scoring for drug microRNA pairs based on gene targets and pathways ............ 13
7. Heat map for drug-microRNA pair scoring based on common targets and pathways for anti-Parkinson drugs and associated microRNAs ......................................................... 14
8. Heat map for drug-microRNA pair scoring based on common targets and pathways for anti-Malarial drugs and associated microRNAs ................................................................. 14
9. Heat map showing the distance between drug microRNA pairs for anti-Malarial drugs and associated microRNAs ........................................................................................................ 16
10. Heat map showing the distance between drug microRNA pairs for anti-Parkinson drugs and associated microRNAs ........................................................................................................ 16
11. Mean silhouette values calculated from k = 5 to k = 30 for anti-Parkinson drugs and related microRNAs. Value peaks at k = 22 ................................................................. 18
12. K-means plot for anti-Parkinson disease drugs and associated microRNAs for optimal k = 22 .............................................................................................................................. 19
13. Silhouette plot for K-means clusters for optimal k = 22, anti-Parkinson disease drugs and microRNAs. Mean silhouette = 0.3119 ................................................................. 19
14. An overview of a self organizing map for disease specific drugs and related microRNAs showing topological neighborhood .................................................................23

15. Part of SOM feature map showing neighborhood of anti-Parkinson drug Bromocriptine 25

16. Part of SOM feature map showing neighborhood of anti-Malarial drug Docetaxel ....... 26

17. A functional module derived by the EAGLE algorithm for anti-Parkinson drug

   Gabapentine and associated miRNA hsa-miR-196a .................................................. 33

18. A functional module derived by the FAG-EC algorithm for anti-Parkinson drug

   Gabapentine and associated microRNA hsa-miR-653 ............................................. 35

19. Comparison between incremental and non incremental algorithm results .................... 40

20. Matrix Compaction for matrix for anti-Parkinson drug and associated miRNAs.

   Matrix has 99.5% data within x = 35. The matrix is cut at x = 35 .............................. 42

21. Dense rectangular submatrix for anti-Malarial drugs and related microRNA .......... 46

22. Dense rectangular submatrix for anti-Cancer drugs and related microRNA .......... 47

23. Dense 7X7 submatrix for anti-Parkinson disease drugs and associated microRNAs..... 49

24. Dense 4X4 submatrix for anti-Parkinson drugs and related microRNAs ................. 49

25. Dense 24X24 submatrix for anti-Cancer drugs and related microRNAs .................. 50

26. High scoring drug-microRNA pairs for validated gene targets and pathways ........... 52

27. Heat map showing the distance between drug microRNA pairs for the validated data set .............................................................................................................. 53

28. Mean silhouette values calculated from k = 0 to k = 10 for validated dataset ........... 52

29. K-means clustering results for the drugs and microRNAs for the validated data set for optimal k = 8 ............................................................................................................... 54

30. Silhouette plot for K-means clusters for optimal k = 8. Mean silhouette = 0.4976 ....... 54
31. The drug microRNA pairs found in the intersection of K means clustering, high scoring
and less distance sets are found to be in each other’s proximity in the self organizing map
........................................................................................................................................... 56

32. A 10X 5 dense submatrix found by greedy randomized multi-domain algorithm for the
validated data set............................................................................................................... 57

33. A 8X8 dense submatrix found by greedy randomized multi-domain algorithm for the
validated data set............................................................................................................... 58
List of Tables

1. Data resources, description and references ................................................................. 4
2. Attribute code .............................................................................................................. 22
3. Symbol Code ............................................................................................................. 23
4. Integrated results for anti-Parkinson drugs and associated microRNAs .................... 27
5. Integrated results for anti-Malarial drugs and associated microRNAs .......................... 28
6. Intersection of results for the validated data set ......................................................... 55
7. Comprehensive relations of intersection results ......................................................... 59
1 Overview

Specific changes in the gene transcriptional machinery are fundamental to gene regulation and thereby the normal developmental and physiological pathways and the perturbed states, resulting in disease. Most of drugs act by modulating gene regulation via transcription factors and transcriptional mechanisms [1]. MicroRNAs (miRNAs or miRs), the recently discovered short (~22bp), non-coding RNA molecules, act typically as negative regulators of gene expression post-transcriptionally by inhibiting mRNA translation or promoting mRNA degradation. Extensively studied first in model organisms, such as C. elegans (nematode worm) and Drosophila (fruit-fly), where they have been shown to control developmental mechanisms, recent studies in higher organisms including humans have revealed their critical role in development and several diseases including cancer and drug response. Currently, there are about ~1000 miRNAs in the human genome which could regulate more than 50% of the known protein-coding genes (~25000). One of the interesting features of miRNA-based gene regulation is the ability to regulate multiple functionally related genes. This property is in contrast to the mode of actions of most drugs, which typically act on specific targets. It is probably this feature of miRNAs which enables them to regulate complex metabolic and regulatory pathways avoiding potentially redundant mechanisms. Since microRNAs typically act as inhibitors of gene expression, adding specific miRNA-mimetic decreases the expression of the target mRNAs. Conversely, inhibiting miRNAs (using antagomirs) leads to activation of gene expression by blocking the target gene suppression activity. Majority of the drugs also function in a similar manner by either activating or suppressing gene expression. This led us to speculate that if a microRNA and drug have similar gene targets and affect similar biological pathways, can they
be used as an alternative for one another or as a combinatorial? In other words, the question we are addressing is: Can microRNAs, the endogenous regulators (mostly suppressors) of genes, be used as alternatives to drugs or as combinatorials (with drugs) to fine tune the drug-response and/or mitigate potential side-effects arising because of off-target effects? In support of our hypothesis, several recent studies have shown that microRNAs indeed play a role in determining drug sensitivity/resistance. For instance, Yang et al., (2008) have reported that in ovarian cancer, mir-214 induces cell survival and resistance to the drug Cisplatin by targeting the gene PTEN[2]. Another facet of miRNA-based regulation in drug response relates to the effects of polymorphisms in target genes leading to either preventing the miRNA-gene binding or creating spurious sites [3]. Nevertheless, the therapeutic potential of microRNAs is still mostly unexplored. Thus, in the current study, using computational approaches, we integrate drug and microRNA target gene datasets and use clustering and mining techniques to identify patterns or motifs comprising at least one drug, one microRNA, and one target gene or pathway.
2 Techniques Used

The techniques used for mining similarity between the drugs and microRNAs were k-means clustering, self organizing map[17], pair wise distance between drugs and microRNAs in the n-dimensional space, EAGLE algorithm [18] (available as a plug-in in Cytoscape[25,26]) and FAG-EC algorithm [19] (Cytoscape plug-in). A greedy randomized algorithm incorporating the heterogeneity of data to be clustered was conceived, designed and applied to the data set to identify modules comprising at least each of the components.

2.1 Data Resources and Construction of Base Matrix

The different databases used for extracting the drug target genes, microRNA target genes, pathways, and protein interactions are enlisted in table (1) along with a brief description and references. Databases DrugBank [4,5], Stitch [6,7] and CTD [8,9] contain formatted representation of drug-gene targets. While the drug targets from DrugBank are experimentally validated, drug target genes from the latter two databases are either predicted (text-mined) or are based on differential gene expression data in human and model organisms. TargetScan[10,11,12] and TarBase [13] are databases of microRNAs and their targets. TarBase contains compiled lists of experimentally validated miRNA targets from human and other model organisms. TarBase, broadly classifies all targets as either ‘translationally repressed targets’ or ‘downregulated/cleaved targets’, and provides limited information about ‘single site sufficiency’, ‘direct support’ and ‘indirect support’ for the miRNA–target interactions[14]. TargetScan
contains predicted gene targets of microRNAs based on the presence of conserved 8mer and 7mer sites that match the seed region of each miRNA. Additionally, to support species-specific miRNA target genes, it also contains gene targets that have non-conserved microRNA binding sites [10]. We use this information for scoring (section 3) the drug and microRNA pairs based on common gene targets, where a derived relation having a validated target (e.g., from DrugBank and/or TarBase) is scored higher than a predicted target. The KEGG (Kyoto Encyclopedia of Genes and Genomes) database contains pathways (372 reference pathways) from a wide variety of organisms (>700) including human. Each of the pathways is hyperlinked to metabolite and protein/enzyme information, where available [4,5]. The biological pathways and the genes that are involved in those pathways are used for the purpose of linking drugs and microRNAs to the pathways by exploiting the transitive relation. In other words, if drug A targets gene G, and gene G is involved in pathway P1, P2 …Pn, then drug A is related to pathways P1, P2…Pn. Likewise, for microRNAs and their target genes.

Table 1: Data resources, description and references

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
<th>Reference/URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DrugBank</td>
<td>Bioinformatics and cheminformatics resource that combines drug data with drug target information. Currently, it has ~4800 drug entries including &gt;1,350 FDA-approved small molecule drugs, 123 FDA-approved biotech (protein/peptide) drugs, 71 nutraceuticals and &gt;3,243 experimental drugs. Each DrugCard</td>
<td><a href="http://www.drugbank.ca">http://www.drugbank.ca</a></td>
</tr>
</tbody>
</table>
An entry contains several data fields with information ranging from drug/chemical data to drug target or protein data [4,5].

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stitch</td>
<td>A compilation of known and predicted interactions of drugs/chemicals and proteins extracted from experiments, other specialized databases and mined from the literature [6,7].</td>
<td><a href="http://stitch.embl.de">http://stitch.embl.de</a></td>
</tr>
<tr>
<td>TargetScan</td>
<td>Database of biological targets of miRNAs predicted based on the presence of conserved 8mer and 7mer sites that match the seed region of each miRNA. The miRNA-target predictions are ranked based on the predicted efficacy of targeting as calculated using the context scores of the sites [10,11,12].</td>
<td><a href="http://www.targetscan.org">http://www.targetscan.org</a></td>
</tr>
<tr>
<td>TarBase</td>
<td>Compilation of manually curated and experimentally validated microRNA targets in several animal species including humans. Current version includes more than 1300 experimentally supported targets and each target site is described by the miRNA that binds it, the gene in which it occurs, the nature of the</td>
<td><a href="http://diana.cslab.ece.ntua.gr/">http://diana.cslab.ece.ntua.gr/</a></td>
</tr>
</tbody>
</table>
experiments that were conducted to test it, the sufficiency of the site to induce translational repression and/or cleavage, and the paper from which all these data were extracted [13].

<table>
<thead>
<tr>
<th>Biocarta</th>
<th>Catalog of both classical pathways and suggestions of new pathways formed by encapsulation of resources proving information for over 120,000 multi-specie genes [22].</th>
<th><a href="http://www.biocarta.com/">http://www.biocarta.com/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>NCI- PID</td>
<td>Pathway Interaction Database (PID) of information about known biomolecular interactions and key cellular processes assembled into signaling pathways provided by The National Cancer Institute (NCI) in collaboration with Nature Publishing Group[23].</td>
<td><a href="http://pid.nci.nih.gov/">http://pid.nci.nih.gov/</a></td>
</tr>
</tbody>
</table>
Since genes typically function by interacting with other genes, a drug targeting a specific gene might indirectly affect another gene that interacts with the target gene. For instance, if drug D acts by targeting gene G1 (direct effect) and if G1 is known to interact with gene G2, then the drug D might affect the interaction of G1 and G2 and hence G2 and its related interactions (indirect effect). Likewise, if gene G1 is involved in a pathway P and gene G2 (which interacts with G1) is suppressed/promoted (e.g., by a drug or a miRNA), it could impact pathway P.

To elucidate this further, let us consider the following hypothetical example:

\[ A \longrightarrow B \]
\[ B \longrightarrow X, Y, Z, \ldots \]
\[ C \longrightarrow Y \]
\[ A \longrightarrow Z \]

Here A is a drug, B is a pathway which involves genes X, Y, Z, etc. and drug A acts by modulating pathway B. C is a microRNA which suppresses gene Y and Z is a known target of drug A. Although the drug and miR target two different genes, these genes belong to the same pathway and there is a likelihood of having an indirect effect on the system by manipulating any individual or all of the players in this network of individual relationships. Indeed, this could be the way physiological systems function because genes and proteins typically work together. Since, it is difficult to experimentally validate all such relationships it would be useful, if these indirect or inferred relationships can be captured and scored (e.g., drug and miR targeting a common gene will be scored high). In figure 1 drug1 targets gene3, whereas microRNA1 targets gene2. Since drug1 affects pathway1 and gene1, gene2 and gene3 all come from pathway1, there is an indirect relation between the drug and the microRNA.
Figure 1: Determination of indirect relations

Based on the above hypothesis we then formulated processes which could start off with target genes for drugs and microRNAs, and pathways they are involved in, followed with filtering out the intricacies of their individual relationships and relate them to one another.

The DrugBank database is used to download drug-targets and also drug category (e.g., anti-Psychotic, anti-Parkinson, anti-Cancer, anti-Malarial, etc.) By cross referencing the gene-gene interactions, drug-gene, and gene-pathways associations, drug-gene-pathways connections are established and the one-to-many relation between drugs and pathways are deduced. Similar approach was repeated on the miRNA target gene data sets to extract microRNA-gene-pathways connections.

Matrices of drugs/microRNAs versus genes/pathways are created to depict their relation. The data used is a matrix of drugs and microRNAs (both treated as one entity) as the rows of the matrix with the attribute columns as genes and pathways (both treated as attribute types). If a drug/miR target is a gene, the value of Matrix[drug/microRNA][gene] is set to 1. If the
drug/microRNA targets a gene G1 which interacts with gene G2 then the value of Matrix[drug/microRNA][G2] is set to 0.8. If the drug/microRNA targets a gene G1 which targets another gene G2 and G2 targets gene G3, then the value Matrix[drug/microRNA][G3] is set to 0.6. Here the subscripts are the indices of the particular drug, microRNA or gene in the rows and columns of the matrix. The value differentiation is done to add more weightage to direct gene targets as against indirect ones. The same concept is used for pathways. If a drug is involved in pathway P1, then the value of Matrix[drug/microRNA][P1] is set to 1. If the drug/microRNA targets a gene G1 which interacts with another gene G2 which is a part of a pathway P1 then the value Matrix[drug/microRNA][P1] is set to 0.8. This matrix is used as the base for the different mechanisms that are used to score, rank and group drugs and microRNAs together.
Matrix[drug/microRNA][gene1] = 1

Matrix[drug/microRNA][gene2] = 0.8

Matrix[drug/microRNA][gene3] = 0.6

Figure 2: Cell values for matrix based on gene targets

Matrix[drug/microRNA][pathway1] = 1

Matrix[drug/microRNA][pathway1] = 1

Matrix[drug/microRNA][pathway1] = 0.8

Matrix[drug/microRNA][pathway1] = 0.6

Figure 3: Cell values for matrix based on associated pathways
<table>
<thead>
<tr>
<th></th>
<th>gene1</th>
<th>gene2</th>
<th>pathway1</th>
<th>pathway2</th>
</tr>
</thead>
<tbody>
<tr>
<td>drug1</td>
<td>1 (drug1 → gene1)</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>drug2</td>
<td>0.8 (drug2 → geneX — gene1)</td>
<td>1 (drug2 → gene1)</td>
<td>1 (drug2 → pathway1)</td>
<td>0</td>
</tr>
<tr>
<td>microRNA1</td>
<td>0.8</td>
<td>0.6 (microRNA1 → geneX — geneY — gene2)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>microRNA2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1 (microRNA → pathway)</td>
</tr>
</tbody>
</table>

**Figure 4: Drug/MicroRNA vs gene/pathway matrix**

![Heat map for matrix for anti-Cancer drugs and microRNAs](image)

**Figure 5: Heat map for matrix for anti-Cancer drugs and microRNAs**
3 Scoring

A simple mechanism to score drug microRNA similarity relationship was formulated.

1) If a drug D and microRNA M have a common gene target their similarity is scored at 90.

2) If a drug D has a gene target G1 which interacts with gene G2 which is a target of microRNA M, then the pair is scored at 80.

3) If a drug D has a gene target G1 which interacts with gene G2 which interacts with gene G3 which is a target of microRNA M, and all three genes G1, G2 and G3 belong to the same pathway then they are scored at 75.

4) If a drug D has a gene target G1 which interacts with gene G2 which interacts with gene G3 which is a target of microRNA M, they are scored at 70.

5) If a drug D has a gene target G1 and gene G2 is a target of microRNA M and genes G1 and G2 belong to the same pathway, they are scored at 60.

6) For all validated targets satisfying conditions from 1) to 5), add 5 to the score each for validated drug target and validated microRNA target.

The above scoring can be used for ranking drug microRNA pairs in their similarity. For the purpose of this study, the drugs were sampled per category – anti-Cancer, anti-Parkinson, anti-Malarial. For a given category of drugs, the high ranking pairs can be observed and can be use as an index into measuring the usability of microRNAs as alternatives to the drugs.
Figure 6: Scoring for drug microRNA pairs based on gene targets and pathways
Figure 7: Heat map for drug-microRNA pair scoring based on common targets and pathways for anti-Parkinson drugs and associated microRNAs

Figure 8: Heat map for drug-microRNA pair scoring based on common targets and pathways for anti-Malarial drugs and associated microRNAs
4 Drug MicroRNA Distance

Another parameter that can be used as an index to evaluate drug-microRNA pairs is the distance between the drugs and microRNAs in the n dimensional space. The distance between each drug-microRNA pair in the n dimensional space is calculated. Each row in the matrix composed in section 2, represents each drug /microRNA in the n dimensional space. The Euclidean distance between each drug-microRNA pair is calculated and the pairs are ranked in increasing order of distance. The distance used is the Euclidean distance. The distances are ranked in descending order. The microRNAs which are closer to certain drugs in the n dimensional space can be hypothesized to be the latter’s alternative by virtue of being closer to it in the n dimensional space showing similarity in nature and behavior. Drug and microRNA pairs which have minimum distance in the n dimensional space, as against the other pairs farther away are such pairs. The distance was calculated for all drug- microRNA pairs for the different categories of drugs, and some of the significant results were recorded. Sulfadoxine , which is an anti-Malarial drug and microRNAs has-miR-191, msa-miR-210 and hsa-miR-299-3p have zero distance in the n dimensional space if the comparing criteria is gene targets and pathways. These pairs can be taken up for experimentation. Each drug can also go through and rank its distance from each of the microRNAs . The candidates for each drug can be selected from the best ranked microRNAs based on closest distance.
Figure 9: Heat map showing the distance between drug microRNA pairs for anti-Malarial drugs and associated microRNAs

Figure 10: Heat map showing the distance between drug microRNA pairs for anti-Parkinson drugs and associated microRNAs
5 Clustering

For adopting a clustering approach on the given data set, the matrix generated in section 2 was subjected to clustering to organize drugs and microRNAs into “groups” or clusters based on their targets and pathways. Each row in our matrix represents a drug or microRNA in the n dimensional space while the target genes and pathways are the attributes. We first used K-means clustering to cluster the drugs and microRNAs.

5.1 K-means clustering

The drug – microRNA matrix constructed earlier is used for K-means clustering. The drugs and microRNAs are treated as homogenous entities for this purpose. A value of k is decided upon and the data set is fragmented into k clusters in which each entity in the cluster is closest to the mean of its own cluster as against the means of the other clusters.

An issue faced in K means clustering is the determination of the value of k. In order to find out an optimal value of k the silhouette coefficient can be used. Silhouette coefficients are a measure of how different the clusters are from one another - a dissimilarity measure for the clusters. K means clustering is run on the matrix for a range of values of k. The silhouette coefficients [16,24] are calculated for each drug/microRNA in each of the clusters for a particular value of k. This value of silhouette coefficient is averaged out for all the drugs/microRNAs in their respective clusters, after a run for a particular value of k. Silhouette coefficients which have values closer to 1 reflect the better disjoint of the clusters. This is repeated for a range of k and
the value of k which gives the best (closest to +1) value of average silhouette is considered as the optimal number of clusters. The corresponding value of k is considered. This entire process is repeated for several iterations, since each time K-means clustering is done on the data set, the clusters are not identical. The value of k is then averaged out.

The clustering can be done with a number of measures. Parameters such as cityblock distance, cosine distance’, ‘correlation’ were tried out but the silhouette values were negative or very low for these measures. The measure ‘cityblock distance’ gave -ve silhouette values, ‘cosine’ gave -ve values, ‘correlation’ gave small +ve values and ‘Euclidean distance’ gave +ve values of silhouette coefficient closer to 1. The Euclidean distance measure was used both for the purpose of K-means clustering as well as for the calculation of silhouette in this work. Matlab’s inbuilt library function ‘kmeans’ was used for this purpose.

![Figure 11: Mean silhouette values calculated from k = 5 to k = 30 for anti-Parkinson drugs and related microRNAs. Value peaks at k = 22.](image-url)
Figure 11: K-means plot for anti-Parkinson disease drugs and associated microRNAs for optimal k = 22

Figure 13: Silhouette plot for K-means clusters for optimal k = 22, anti-Parkinson disease drugs and microRNAs. Mean silhouette = 0.3119

The clusters are analyzed and drugs and microRNAs which fall into the same cluster are filtered out for further observation. These results will be used later.
6 Self Organizing Map

Self organizing maps are a way for the low dimensional visualization of higher dimensional data using a special type of artificial neural network which trains and maps data by competitive learning. The neurons are placed at the nodes of a lattice that is usually a one or a two dimensional lattice. During the course of training the neurons become selectively tuned to the various input patterns (stimuli) or classes of input patterns. The winning neurons align themselves in such a way that “a meaningful coordinate system for different input features is created in the lattice (Kohonen. 1990a)” “A self organizing map is therefore characterized by the formation of a topographic map of the input patterns in which the spatial locations (that is coordinates) of the neurons in the lattice are indicative of intrinsic statistical features contained in the input patterns hence the name ‘Self Organizing map’.”[17]

Algorithm[17]

1) Initialization : Random values for the initial weight vectors $W_j(0)$ are chosen. The values of $W_j(0)$ should be different for $j = 1, 2, \ldots, l$, where $l$ is the number of neurons in the lattice. Keeping the weights small should be beneficial.

Another way of initializing the algorithms is to randomly select the weight vectors $\{W_j(0)\}_{j=1}^{l}$ from the available set of input vectors $\{X_i\}_{i=1}^{N}$.

2) Sampling : A sample $x$ from the input space is drawn with a certain probability. The vector $x$ represents the activation pattern that is applied to the lattice. The dimension of vector $x$ is equal to $m$ where $m$ is the dimension of the weight vector.
3) **Similarity matching:** The best matching (winning) neuron \( i(x) \) at time step \( n \) is found by calculating the Euclidean distance between the input vector and the weight vector.

\[
I(x) = \arg \min_j \| x(n) - w_j \|; j = 1,2,\ldots,l
\]

4) **Adaption:** The synaptic weight vectors of all neurons is adjusted by using the update formula

\[
w_j(n+1) = w_j(n) + \eta(n) h_{j,i(x)}(n)(x(n) - w_j(n))
\]

Where \( \eta(n) \) is the learning rate parameter and \( h_{j,i(x)}(n) \) is the neighborhood function centered around the winning neuron \( i(x) \). Both \( \eta(n) \) and \( h_{j,i(x)}(n) \) are chosen as non constant dynamically variable functions to yield better results.

5) **Continuation:** Steps 2 through 4 are continued until no noticeable changes in the feature map are observed.

One of the methods of visualizing a self-organizing map is representing the output layer as a 2D lattice of neurons with each neuron having a class label depending on how each test pattern excites a particular neuron in a self-organizing network [15]. The representation of data set for SOM (Self Organizing Map) implementation has been done in such a way that each data set pertains to the total number of drugs plus the total number of microRNAs. Each column of the table is a schematic description of the drug/microRNA. 0 means “no relation”, 1 implies direct relation. The intermediate relations have been represented using the following criteria - 0.8 means indirect relation and 0.6 means an additional degree of indirection. The construction of this part of the matrix is the same as that discussed in section 2.

Each drug/microRNA is made of a symbol code and an attribute code. The symbol code is a column vector with a positive value at the \( i^{th} \) location where \( i \) is the \( i^{th} \) drug /miR and the rest are
0s. To make sure that the attribute code has greater weightage, the positive value of the attribute has been given as 0.2 [17].

The attribute code is of length equal to the total number of genes plus the total number of pathways. The data is a matrix of drugs and microRNAs (both treated as single entity) as the rows and genes and pathways as columns. If a drug/miR target is a gene, the value of that row would be 1, if they target a gene which interacts with that gene then the value is taken as 0.8. If they target a gene which targets another gene which targets this gene, then the value is 0.6. This is to give more weightage to direct target genes as against indirect ones. The same concept is used for pathways. If a drug is involved in pathway, then the value of that cell becomes 1, if it deals with a gene which interacts with another gene which is a part of a pathway then the value 0.8.

**Table 2: Attribute code**

<table>
<thead>
<tr>
<th></th>
<th>gene1</th>
<th>gene2</th>
<th>gene3</th>
<th>gene4</th>
<th>………</th>
<th>pathway1</th>
<th>pathway2</th>
<th>pathway3</th>
<th>………</th>
</tr>
</thead>
<tbody>
<tr>
<td>drug1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.6</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>drug2</td>
<td>0</td>
<td>0</td>
<td>0.8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>………</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>microRNA1</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>0.8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>microRNA2</td>
<td>0</td>
<td>1</td>
<td>0.6</td>
<td>0.8</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>………</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 3: Symbol Code

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>drug1</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>drug2</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>........</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>microRNA1 = drugi</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>microRNA2 = drugi+1</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>........</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Finally the data is normalized to unit length. The weights of the lattice are randomized. The patterns of data are presented to the lattice and their weights are modified according to the SOM algorithm. This is repeated for about 2000 iterations after which the feature map reaches a steady state. Now only the symbol code of the patterns are presented to the neuron and the neuron which gets the most excited by a particular pattern is given the name of that pattern. This is repeated for each of the patterns.
Figure 14: An overview of a self organizing map for disease specific drugs and related microRNAs showing topological neighborhood
Figure 15: Part of SOM feature map showing neighborhood of anti-Parkinson drug

Bromocriptine
What is noticed in the results is that all the drugs retreat to one end of the map and the microRNAs to the other side. This was anticipated since drugs used to treat a particular disease would be similar and microRNAs would be similar. Some results like the proximity of Bromocriptine to hsa-miR-154, hsa-miR-339-5p, hsa-miR-708, hsa-miR-491-5p do show up.

One issue is the demarcation of similar clusters. The feature maps do tend to geographically distribute the proximity of drugs and microRNAs according to their correlation to each other, but the boundaries of these clusters cannot be ascertained. The results can be used as a basis of experimentation to endorse the relation.
7 Integrating the results

The results obtained from scoring, and drug-microRNA distance measure ranking in the n-dimensional space and K-means clustering are intersected to find the drugs and microRNA pairs which made it to the subset of all three. The drug – microRNA pair which were scored high (>=70), were ranked 1st through 5th closest to each other and were also a part of the same cluster were filtered out.

These results were cross checked with the self organizing feature map by seeing the proximity of the result drug and microRNA in the map.

Some of the results in the form of drug microRNA pairs which were an intersection of the better ranked results in all the techniques are as follows. The first two columns are drug and microRNA, the third column is distance of drug from the microRNA and the fourth column is the score.

Table 4: Integrated results for anti-Parkinson drugs and associated microRNAs

<table>
<thead>
<tr>
<th>Drug</th>
<th>MicroRNA</th>
<th>Distance</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orphenadrine</td>
<td>hsa-miR-191</td>
<td>2.828427</td>
<td>85</td>
</tr>
<tr>
<td>Memantine</td>
<td>hsa-miR-210</td>
<td>2.645751</td>
<td>75</td>
</tr>
<tr>
<td>Orphenadrine</td>
<td>hsa-miR-210</td>
<td>2.645751</td>
<td>75</td>
</tr>
<tr>
<td>Bromocriptine</td>
<td>hsa-miR-339-5p</td>
<td>2.645751</td>
<td>85</td>
</tr>
<tr>
<td>Bromocriptine</td>
<td>hsa-miR-324-5p</td>
<td>3</td>
<td>75</td>
</tr>
<tr>
<td>Bromocriptine</td>
<td>hsa-miR-184</td>
<td>3.316625</td>
<td>70</td>
</tr>
<tr>
<td>Drug</td>
<td>MicroRNA</td>
<td>Distance</td>
<td>Score</td>
</tr>
<tr>
<td>------------------</td>
<td>------------</td>
<td>----------</td>
<td>-------</td>
</tr>
<tr>
<td>Chloroquine</td>
<td>hsa-miR-205</td>
<td>5.09902</td>
<td>85</td>
</tr>
<tr>
<td>Quinidine</td>
<td>hsa-miR-485-5p</td>
<td>3.162278</td>
<td>85</td>
</tr>
<tr>
<td>Quinidine</td>
<td>hsa-miR-33a</td>
<td>3.872983</td>
<td>85</td>
</tr>
<tr>
<td>Quinidine</td>
<td>hsa-miR-33b</td>
<td>3.872983</td>
<td>85</td>
</tr>
<tr>
<td>Quinidine</td>
<td>hsa-miR-122</td>
<td>3</td>
<td>85</td>
</tr>
<tr>
<td>Primaquine</td>
<td>hsa-miR-504</td>
<td>3.316625</td>
<td>85</td>
</tr>
<tr>
<td>Mefloquine</td>
<td>hsa-miR-485-5p</td>
<td>2.44949</td>
<td>80</td>
</tr>
<tr>
<td>Quinine</td>
<td>hsa-miR-485-5p</td>
<td>2.645751</td>
<td>80</td>
</tr>
<tr>
<td>Amodiaquine</td>
<td>hsa-miR-485-5p</td>
<td>2.645751</td>
<td>80</td>
</tr>
<tr>
<td>Halofantrine</td>
<td>hsa-miR-33a</td>
<td>3</td>
<td>80</td>
</tr>
<tr>
<td>Halofantrine</td>
<td>hsa-miR-33b</td>
<td>3</td>
<td>80</td>
</tr>
<tr>
<td>Dapsone</td>
<td>hsa-miR-338-3p</td>
<td>2.236068</td>
<td>80</td>
</tr>
<tr>
<td>Trimethoprim</td>
<td>hsa-miR-338-3p</td>
<td>2.236068</td>
<td>80</td>
</tr>
<tr>
<td>Pyrimethamine</td>
<td>hsa-miR-338-3p</td>
<td>2.236068</td>
<td>80</td>
</tr>
<tr>
<td>Proguanil</td>
<td>hsa-miR-338-3p</td>
<td>2.236068</td>
<td>80</td>
</tr>
<tr>
<td>Primaquine</td>
<td>hsa-miR-486-5p</td>
<td>1.414214</td>
<td>80</td>
</tr>
<tr>
<td>Mefloquine</td>
<td>hsa-miR-486-5p</td>
<td>1.414214</td>
<td>80</td>
</tr>
<tr>
<td>Quinine</td>
<td>hsa-miR-486-5p</td>
<td>1.732051</td>
<td>80</td>
</tr>
<tr>
<td>Amodiaquine</td>
<td>hsa-miR-486-5p</td>
<td>1.732051</td>
<td>80</td>
</tr>
<tr>
<td>Dapsone</td>
<td>hsa-miR-379</td>
<td>2.236068</td>
<td>70</td>
</tr>
<tr>
<td>Trimethoprim</td>
<td>hsa-miR-379</td>
<td>2.236068</td>
<td>70</td>
</tr>
<tr>
<td>Pyrimethamine</td>
<td>hsa-miR-379</td>
<td>2.236068</td>
<td>70</td>
</tr>
<tr>
<td>Proguanil</td>
<td>hsa-miR-379</td>
<td>2.236068</td>
<td>70</td>
</tr>
<tr>
<td>Dapsone</td>
<td>hsa-miR-421</td>
<td>2.44949</td>
<td>70</td>
</tr>
<tr>
<td>Trimethoprim</td>
<td>hsa-miR-421</td>
<td>2.44949</td>
<td>70</td>
</tr>
<tr>
<td>Pyrimethamine</td>
<td>hsa-miR-421</td>
<td>2.44949</td>
<td>70</td>
</tr>
<tr>
<td>Proguanil</td>
<td>hsa-miR-421</td>
<td>2.44949</td>
<td>70</td>
</tr>
<tr>
<td>Primaquine</td>
<td>hsa-miR-542-3p</td>
<td>1.414214</td>
<td>80</td>
</tr>
<tr>
<td>Mefloquine</td>
<td>hsa-miR-542-3p</td>
<td>1.414214</td>
<td>80</td>
</tr>
<tr>
<td>Quinine</td>
<td>hsa-miR-542-3p</td>
<td>1.732051</td>
<td>80</td>
</tr>
<tr>
<td>Amodiaquine</td>
<td>hsa-miR-542-3p</td>
<td>1.732051</td>
<td>80</td>
</tr>
<tr>
<td>Quinacrine</td>
<td>hsa-miR-873</td>
<td>4.242641</td>
<td>75</td>
</tr>
<tr>
<td>Docetaxel</td>
<td>hsa-miR-182</td>
<td>3.655133</td>
<td>95</td>
</tr>
<tr>
<td>Docetaxel</td>
<td>hsa-miR-204</td>
<td>4.4</td>
<td>95</td>
</tr>
<tr>
<td>Docetaxel</td>
<td>hsa-miR-211</td>
<td>4.4</td>
<td>95</td>
</tr>
<tr>
<td>Docetaxel</td>
<td>hsa-miR-132</td>
<td>4.04475</td>
<td>85</td>
</tr>
<tr>
<td>Docetaxel</td>
<td>hsa-miR-212</td>
<td>4.04475</td>
<td>85</td>
</tr>
<tr>
<td>Docetaxel</td>
<td>hsa-miR-34a</td>
<td>3.37046</td>
<td>100</td>
</tr>
<tr>
<td>Docetaxel</td>
<td>hsa-miR-34c-5p</td>
<td>3.37046</td>
<td>95</td>
</tr>
<tr>
<td>Docetaxel</td>
<td>hsa-miR-449a</td>
<td>3.37046</td>
<td>95</td>
</tr>
<tr>
<td>Drug</td>
<td>miRNA</td>
<td>Score</td>
<td>Fold</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>Docetaxel</td>
<td>hsa-miR-449b</td>
<td>3.37046</td>
<td>95</td>
</tr>
<tr>
<td>Docetaxel</td>
<td>hsa-miR-365</td>
<td>1.16619</td>
<td>95</td>
</tr>
<tr>
<td>Docetaxel</td>
<td>hsa-miR-433</td>
<td>3.37046</td>
<td>85</td>
</tr>
</tbody>
</table>

When the SOM was cross checked, pairs like anti-Parkinson drug Bromocriptine and hsa-miR-339-5p were actually found to be neighbors (figure 15) and anti-Malarial drug Docetaxal and hsa-miR-34c-5p, hsa-miR-449a, hsa-miR-449b, hsa-miR-443, hsa-miR-182, hsa-miR-365 and hsa-miR-212 were found to be neighbors (figure 16) in their respective SOM feature maps.
8 Network Creation

The resulting subset of drugs and microRNAs are then used with the corresponding data about their respective gene targets and pathways and fed to Cytoscape [25, 26] for network creation. Cytoscape is an open source software tool used for the purpose of network analysis, integration and visualization. This gives us network diagrams of drugs and microRNAs and how each of them is linked to the various genes and pathways. This can be used to analyze the connections between these nodes. The two clusterviz or cluster visualization algorithms used for network analysis are discussed below.

8.1 EAGLE Clustering Algorithm

Shen, Cheng, Cai and Hu came up with the algorithm AgglomerativE hierarchicAl clusterinG based on maximal clique. It is an algorithm which “detects both the overlapping and hierarchical properties of complex community structures “[18]. The maximal cliques in the network are found out. To avoid the subordinate maximal cliques which are small by size a threshold k is set and all maximal cliques less than that size are neglected. In real world networks most subordinate maximal cliques are between 3-6 in size.[18]

Their algorithm has two stages.

A) Dendrogram generation

B) Selection of appropriate cuts which breaks the dendrogram into communities.
The maximal cliques in the network are discovered and the subordinate maximal cliques are neglected. The communities are initialized with the cliques remaining after eliminating the subordinate cliques inclusive of the subordinate vertices comprising of the sole vertex. The similarity between the communities is computed. Following this the pair of communities with the maximum similarity is selected and combined to form a new one. The similarity between the new community and the other communities is then calculated. This is repeated until only one community remains.

In the algorithm Shen, Cheng, Cai and Hu defined the similarity measure between two communities \( C_1 \) and \( C_2 \) as

\[
M = \frac{1}{2m} \sum_{v \in C_1, w \in C_2, v \neq w} [A_{vw} - K_v K_w / 2m]
\]

The cut for the dendrogram is determined by using an extensive measure which is defined to judge the quality of the cover.

\[
EQ = \left( \frac{i}{2m} \right) \sum_{v \in C_i, w \in C_j} \left[ \frac{1}{O_v O_w} \left[ A_{vw} - K_v K_w / 2m \right] \right]
\]

In the above equations \( A_{vw} \) is the element of the adjacency matrix, \( m \) is the total number of edges in the network and \( K_v \) is the degree of vertex \( v \).

The above clusterviz algorithm which is a plugin in Cytoscape was used for the purpose of this thesis to discover communities in the networks that are formed with the following relationships

A) drug → gene

B) gene —— gene

C) microRNA → gene

D) pathway —— gene
The functional modules derived by EAGLE using the data set were studied. The functional complexes that were identified were biased towards similar species. Either they had a dominance of microRNA and their respective gene targets or pathways or drug and their respective gene target or pathways. So essentially it was biased towards the raw, unprocessed data from Drugbank or Tarbase or gene-pathway data. It is mostly the microRNAs that exhibit themselves with their respective dependencies, in such complexes because of the quantitative predominance of microRNAs in the data set. This does not fulfill our criteria of coming up with conclusive complexes having both drugs and microRNAs. This is because of the heterogeneous nature of the data we wish to cluster.
Figure 17: A functional module derived by the EAGLE algorithm for anti-Parkinson drug Gabapentine and associated microRNA hsa-miR-196a

8.2 FAG Clustering Algorithm

The cluster visualization algorithm in Cytoscape – FAG-EC clustering algorithm conceived and implemented by Li, Wang and Chen was also applied on the data set. FAG expands to -- A Fast Agglomerate Algorithm for Mining Functional Modules in Protein Interaction Networks [19]. The algorithm is based on the definitions of edge clustering coefficients and modules.
The input to FAG-EC is PPI (Protein-Protein Interactions). An undirected simple graph \( G \) \((V,E)\) is created first, followed by the calculation of the clustering coefficients of all the edges in the graph \( G \). The formula used for the calculation is given below [19].

If \( N_i \) and \( N_j \) are considered to be two sets of neighbors of node \( i \) and node \( j \), the clustering coefficient of edge \((i,j)\) was defined as

\[
C_{i,j} = \frac{| N_i \cap N_j | }{ \min (k_i , k_j ) }
\]

where \( k_i \) and \( k_j \) are the degrees of node \( i \) and node \( j \).

Based on the edge clustering coefficients, all the edges are enqueued into a queue \( S_q \) from the highest one to the lowest one. The higher clustering coefficient of an edge indicated higher probability of the edge to be inside a module. Initially all the nodes in the graph \( G \) are initialized as a singleton subgraph, and are labeled as mergable. By gradually adding edges in the queue \( S_q \) to subgraphs, the algorithm FAG-EC finally assembles all the singleton subgraphs into modules[19].

The functional modules derived by FAG-EC using the current data set were studied. The functional complexes that were identified were found to be biased towards similar species. Either they had a dominance of microRNA – microRNA links and their respective gene targets or pathways or drug – gene, drug – pathways links. It is mostly the microRNAs that exhibit themselves with their respective dependencies, in such complexes because of the quantitative predominance of microRNAs in the data set. This did not fulfill our criteria of coming up with conclusive complexes having both drugs and microRNAs.

This is because of the heterogeneous nature of the data we wish to cluster.
Figure 18: A functional module derived by the FAG-EC algorithm for anti-Parkinson drug Gabapentine and associated microRNA hsa-miR-653
9 Greedy Randomized Algorithm

The primary issue with the other clustering techniques is the separate grouping of all the drugs and their dependencies together and microRNAs and their dependencies together. This masks the very inter specie relation that we seek to find. In order to integrate that aspect in a mining technique, an algorithm was conceived and implemented to find out the densest largest submatrices having a predefined ratio of drugs and microRNAs. The approach adopted was that of a randomized greedy approach found to be previously used by Dharan and Nair [20]. The approach is to randomly generate small submatrices and then select the best ones at each step and then grow them. This approach was augmented with method for inclusion of multi-domain data points instead of single domain data.

9.1 Dharan Nair paper

The algorithm approach taken by Dharan and Nair was the Greedy Randomized Adaptive Search procedure [20]. Using this approach they generated high quality seeds of submatrices using the linear K-means clustering, and from here on incrementally grew the submatrices by randomly adding rows and columns to the existing submatrices, subject to the lower threshold of mean squared residue score, HScore, proposed by Cheng and Church[21].

The limitations of this algorithm with respect to our data set is the determination of an initial value of k for the seed elements, which was encountered as one of the pitfalls when k means clustering was implemented in section 5. The selection of k has not been discussed by the paper.
Another issue is that the rows are columns in Dharan and Nair’s dataset were homogeneous in nature. The rows were all genes and the columns were all expression levels under experimental conditions. If similarity between heterogeneous data is to be found then this algorithm would give results similar to the results that the EAGLE clustering and FAG clustering were giving- the data from the same species are predominant in the clusters and more often than not the clusters themselves represent only one kind of data.

Building from the above approach and including parameters to suit the data and problem at hand, the additional inclusion in this algorithm is the forced selection of row candidates, firstly in the seed data and next in the iteration steps, such that the ratio of the row elements in the final matrix fulfills a certain threshold criteria. This would enable the visualization of similarity between non homogeneous data, as in our case where we could hypothesize the inter usability of one kind of drug with another microRNA.

9.2 Non-incremental algorithm

The greedy randomized algorithm which will be discussed shortly is compared with the absolute randomized approach. In this approach, \( n \times n \) submatrices are generated randomly with no previous reference. The submatrices whose sum values cross the threshold are retained. Here threshold is the lower limit of (sum of submatrix elements)/(total size of submatrix). With each iteration the submatrix crossing the threshold, with sum greater than the sum of the submatrix with the minimum sum in the SolutionSet, replaces the submatrix with the minimum sum. This is continued till, for a full cycle of maxIteration which was taken as 5000000000 iterations, not a single sub matrix is generated which crosses the threshold value.
Algorithm

1) Generate randomly a \( n \times n \) submatrix \( \text{submat} \) from the matrix \( \text{Matrix} \) (section 2).

2) \( \text{sum} = \text{sum} (\text{submat}) \)

3) If \( \text{sum} < \text{threshold} \)
   
   Throw away \( \text{submat} \)
   
   Continue

   Else

   Test \( \text{SolutionSet} \) for redundancy

   If no redundancy is found

   \( \text{SolutionSet} = \text{SolutionSet} \cup \text{submat} \)

   \( \text{crossthresh} = \text{crossthresh} + 1 \)

   Else

   Continue

   Endif

   Endif

4) Repeat steps 1 through 3, \( \text{maxIteration} \) no of times

5) If \( \text{crossthresh} > 0 \)

   \( n = n+1 \) // growing the size of the square submatrix

   Goto step 1

   Else
Final Solution Set = Solution Set
End
Endif

This algorithm is based on absolute random sub matrix generation. Since every iteration generates fresh matrices as against the algorithm in section 9.3 which good seed values from the previous iteration and incrementally growing them, hence the name ‘Non-incremental Algorithm’. Even though for some runs this algorithm found good matrices with high densities, but there is no assurance that it always will, since at every step it searches for a n \( \times \) n matrix from scratch and then sums up the elements, so the probability of the sum of elements not crossing the threshold is high. The stopping criteria is also not very satisfactory since because of generation of random sub matrices from scratch at each point, the probability that crossthresh would be zero, signifying that not a single of the generated sub matrices crossed the threshold over maxIteration iterations, is also high. So the program could terminate before the larger sub matrices are found. If this criterion is relaxed then the program could run for indefinite number of iterations. This algorithm was run on the disease specific data sets. The results for the anti-Parkinson data set are given in figure 19. It is seen that the non incremental randomized algorithm, comes up with smaller dense submatrices. For anti-Parkinson disease drugs and associated microRNAs, the algorithm comes up with 9-4 \( \times \) 4, 6-5 \( \times \) 5, 10-6 \( \times \) 6 and 5-7 \( \times \) 7 dense submatrices which cross the threshold of 0.8 and comply to the drug/microRNA ratio of 0.3 whereas for the same parameters, the incremental greedy randomized algorithm comes up with 30 14 \( \times \) 14 submatrices. For the same parameters but with the threshold increased to 0.95, this algorithm comes up with 40-4 \( \times \) 4 submatrices whereas the incremental version comes up with 12-4 \( \times \) 4, 11-5 \( \times \) 5, 6-6 \( \times \) 6 and 1-7 \( \times \) 7 submatrices.
The dense submatrices detected for anti-Parkinson disease, drug/microRNA ratio = 0.3, threshold = 0.80 using absolutely randomized non incremental algorithm which comes up with fresh submatrices every iteration.

The dense submatrices detected for anti-Parkinson disease, drug/microRNA ratio = 0.3, threshold = 0.80 using greedy randomized incremental algorithm growing seed values from the previous iteration.

The dense submatrices detected for anti-Parkinson disease, drug/microRNA ratio = 0.3, threshold = 0.95 using absolutely randomized incremental algorithm which uses high quality seeds from previous iteration.

Figure 19: Comparison between incremental and non incremental algorithm results

9.3 Incremental Algorithm

The incremental version of the algorithm has advantages over the non incremental version because at each step the already ‘density threshold criteria’ fulfilling seed sub matrices from the
previous step are grown. The probability of finding a dense \( n \times n \) matrix crossing the threshold by adding a random row and column each to an existing \((n-1) \times (n-1)\) sub matrix which already fulfils the threshold criteria is higher than if each of the \( n \) rows and columns of the sub matrix are randomly generated.

In order to reduce the search space the original matrix is compacted to try to get most of the non-zero part of the matrix to one side and then truncate off the part of the matrix which has density less than a minimum percentage.

**Algorithm MatrixCompaction**

1) While no rearrangement of matrix takes place anymore

2) Arrange each row in descending order

3) Arrange each column in descending order

4) Goto step 1

This converges in about 3-5 iterations and we are left with a compacted matrix with the dense areas scooted to one side and the large spaces of 0 values towards the other side. The matrix is then clipped off from the column side to eliminate the empty space. When the matrix is sorted to reflect arrangement like in figure 20, then the values in each column starting from the right most column and gradually moving towards the left, are summed up. Since the arrangement is such that the sparse portion of the matrix is shifted towards the right, the initial (right most) columns sum up to very low values. The moment the column sums up to values > 0.5% of the entire matrix bits, then all columns prior to this (0 to this column number) are incorporated in the new
matrix leaving out the columns with <0.5% out of the data. The rows cannot be clipped off because the rows have the items which essentially need to be clustered. Those should not be eliminated. This algorithm completes in constant time.

Figure 20: Matrix Compaction for matrix for anti-Parkinson drug and associated microRNAs. Matrix has 99.5% data within x = 35. The matrix is cut at x = 35

Algorithm Generate Seed Set

1) Call MatrixCompaction
2) Initialize an empty SeedSet
3) Generate 2 X 2 sub matrix randomly such that sum(elements of submatrix) > threshold
4) SeedSet = SeedSet U submatrix
5) Repeat steps 2 through 3 a million times, each time replacing a submatrix from SeedSet such that sum(new submatrix) > min {sum { SeedSet}}
The SeedSet is initialized to an empty set. 2 X 2 matrices are generated such that the sum of the values in the submatrix is greater than a threshold percentage of the whole matrix. This is repeated for a million iterations and the best submatrices are stored as the initial Seed set. Now we have our seed set ready

**Algorithm IncrementalGreedyRandomized**

1) Call GenerateSeedSet
2) Initializing GrowingSet = SeedSet
3) submat_min = submatrix with minimum sum in GrowingSet and sum_min = sum(submat_min)
4) For each submatrix submat € SeedSet, add a randomly generated row and column to submat such that \( \sum \text{ (rowtype) } / \sum \text{ (all rows)} > \text{threshold ratio} \)
5) If \( \sum \text{ (submat growth) } < \text{threshold new} \)
   - Discard submat
   - Continue
   Else
   - If submat_growth € GrowingSet && sum(submat_growth) > sum_min
   - GrowingSet = GrowingSet - submat_min U submat_growth
   - num_crossthresh++;
   Endif
6) Repeat steps 3 through 5 for a large number of iterations
7) If num_crossthresh == 0
GrowingSet is the result

Finish

Endif

8) SeedSet = GrowingSet
9) AvoidLocalMinima
10) Repeat steps 2 through 8 for a large number of iterations

Now that the initial SeedSet is there, the algorithm is applied to this set of initial high yielding matrices. GrowingSet is the set of matrices which are grown from the SeedSet. The GrowingSet is initialized to the SeedSet. The submatrix submat_{min} with the minimum sum is marked in GrowingSet.

For each submatrix in the SeedSet, a row and column is added, and for the new submatrix submat_{grow} the sum is calculated. If this sum is smaller than the threshold for that matrix size, then the submatrix is thrown away. If sum is greater than the threshold, it is compared with the submatrix submat_{min} with the minimum sum in GrowingSet. If the sum is smaller than the sum of the submat_{min}, then the newly generated submatrix is thrown away. If not, then the entire GrowingSet is scanned for redundancy. If there is no redundancy, then submat_{min} is replaced by submat_{grow}. The value of crossthreshold which keeps track of how many submatrices are growing enough to replace submatrices from the current growing set is incremented. If for an entire loop having a large number of iterations, not a single growing submat_{grow} is good enough to replace an submatrix in GrowingSet, then the algorithm is terminated.
Since we start off with randomized submatrices, even though they are the best at that point of time due to the greedy approach undertaken, this might cause the problem of being stuck in local minima. This means that submatrices which were not good enough initially, but could later yield good submatrices, were thrown out in the initial stages. As a solution to this AvoidLocalMinima is called.

**Algorithm AvoidLocalMinima**

1) Find the submatrix with the minimum sum (subs) in the SeedSet
2) Compute the row (each type) and column dimensions
3) Generate submatrices of the dimension row X column randomly
4) If sum(new submatrix) > sum(Original submatrix) and new submatrix \( \sim \in \) SeedSet, Substitute submatrix with new submatrix
5) Repeat steps 1 through 4 for a predetermined number of iterations

SeedSet is now replaced by GrowingSet. The submatrix in the new SeedSet with the minimum sum submat \( \text{min} \) is found out. A submatrix with dimensions same as that of submat \( \text{min} \) is randomly generated from the Original Matrix. If the sum of the values of this submatrix is greater than submat \( \text{min} \), then submat \( \text{min} \) is replaced by this newly generated submatrix. The new submat \( \text{min} \) is again found out. The above process is repeated. Only a percentage – 33% in this case of SeedSet is allowed to be replaced in this phase.
9.3.1 Rectangle matrix generation

The algorithm was also implemented to generate rectangular submatrices instead of square ones. This was done by modifying the algorithm to select only one row or one column at a time while growing the seed submatrices from the previous iteration, instead of mandatory adding of both a row and column while growing a submatrix. What was found was that for some runs (1/5) the size distribution for the matrices for some runs was such that the breadth to length ratio was $\geq 0.8$, but for the remaining runs (4/5) the resulting matrices are elongated rectangular matrices having a high frequency of drugs/microRNAs and few common genes targets/pathways. This means that a few genes are common targets of most of the drugs/microRNAs.

![Dense rectangular submatrix for anti-Malarial drugs and related microRNA](image)

**Figure 21: Dense rectangular submatrix for anti-Malarial drugs and related microRNA**
Figure 22: Dense rectangular submatrix for anti-Cancer drugs and related microRNA

9.3.2 Square matrix generation

The algorithm was implemented to generate square sub matrices. This was done by mandatory selection of both a row and a column for each iteration while growing the seed submatrices from the previous iteration. It is seen that the algorithm comes up with the densest square sub matrices which satisfy drug/microRNA ratio and threshold condition. The algorithm is run several times
to check if the results differ each time due to the randomized approach. The several runs actually converge to a more or less consistent result set covering the same drugs and microRNAs. The run of the algorithm on data dealing with anti-Parkinson disease, converged to $45 \times 5$ dense submatrices, $16 \times 6$ submatrix and $254 \times 4$ submatrices for 3 different runs, where the drug-microRNA ratio was 0.3, the density percentage was 0.95. For density 0.8 and drug-microRNA ratio 0.3 for anti-Parkinson disease, the biggest dense square submatrices the algorithm came up with were $14 \times 14$ submatrices for 3 consecutive runs.

These submatrices having high density are clusters of drugs and microRNAs which are highly similar in nature over a range of gene targets and biological processes. So the drugs and microRNAs can be regarded to have high degree of correlation to each other. These can further be studied in collaboration with the intersection results of scoring, K-means clustering, self-organizing feature map and drug distance. This intersection set would contain drug microRNA pairs which satisfy high scoring, less distance in the n dimensional space, same cluster membership by K-means and proximity in SOM. This would mean that these drugs and microRNA are paired together using various techniques, so they show marked correlation with each other and can be used for experimentation to test as a drug alternative or combinatorial.
Figure 23: Dense 7 X 7 submatrix for anti-Parkinson disease drugs and associated microRNAs

Figure 24: Dense 4 X 4 submatrix for anti-Parkinson drugs and related microRNAs
Figure 25: Dense 24 X 24 submatrix for anti-Cancer drugs and related microRNAs
9.4 Limitations

a) Due to the randomization approach, the exploration and discovery of all the dense submatrices cannot be guaranteed.

b) A value for the number of clusters desired, needs to be decided upon. If this value is small, lots of dense submatrices will be ignored. If very large then the computation time consumption will be large and if as many dense submatrices are not available in the dataset the extra computation is a waste of time.

c) The local minima avoidance is not guaranteed to work due to randomization
10 Results for validated data

The algorithms were run for validated data from DrugBank, Tarbase and NCI pathways. The results derived from integrative mining of high quality, validated data sets can be used as positive controls and any patterns discovered should be interesting. The data set was scored, clustered using k-means clustering, the distance between the drug-microRNA pairs in the n-dimensional space calculated, and mapped in the lower dimensions of the self organizing map.

Figure 26: High scoring drug-microRNA pairs for validated gene targets and pathways
Figure 27: Heat map showing the distance between drug microRNA pairs for the validated data set

Figure 28: Mean silhouette values calculated from $k = 0$ to $k = 10$ for validated dataset
Figure 29: K-means clustering results for the drugs and microRNAs for the validated data set for optimal k = 8

Figure 30: Silhouette plot for K-means clusters for optimal k = 8. Mean silhouette = 0.4976
Table 6: Intersection of results for the validated data set

<table>
<thead>
<tr>
<th>Drug</th>
<th>MicroRNA</th>
<th>Distance in n dimensional space</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-miR-124</td>
<td>Perphenazine</td>
<td>3.162278</td>
<td>90</td>
</tr>
<tr>
<td>hsa-miR-124</td>
<td>Pimozide</td>
<td>3.000000</td>
<td>90</td>
</tr>
<tr>
<td>hsa-miR-124</td>
<td>Fluphenazine</td>
<td>3.000000</td>
<td>90</td>
</tr>
</tbody>
</table>
Figure 31: The drug microRNA pairs found in the intersection of K means clustering, high scoring and less distance sets are found to be in each other’s proximity in the self organizing map
| hsa-miR-7  | 1 | 1 | 1 | 1 | 1 | 1 |
| Pimozide  | 1 | 1 | 1 | 1 | 1 | 1 |
| hsa-miR-29a | 1 | 1 | 1 | 1 | 1 | 1 |
| hsa-miR-16 | 1 | 1 | 1 | 1 | 1 | 1 |
| Perphenazine | 1 | 1 | 1 | 1 | 1 | 1 |
| hsa-miR-15a | 1 | 1 | 1 | 1 | 1 | 1 |
| hsa-miR-29b | 1 | 1 | 1 | 1 | 1 | 1 |
| Fluphenazine | 1 | 1 | 1 | 1 | 1 | 1 |
| hsa-miR-17 | 1 | 0 | 1 | 1 | 0 | 0 |
| hsa-miR-124 | 1 | 1 | 1 | 1 | 1 | 1 |

**Figure 32:** A 10 X 5 dense submatrix found by greedy randomized multi-domain algorithm for the validated data set

Figure 32 shows a dense submatrix obtained by running the incremental greedy randomized algorithm on the validated data set for rectangular submatrices, with drug/miRNA ratio = 0.3 and threshold 0.95
<table>
<thead>
<tr>
<th>hsa-miR-15a</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fluphenazine</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>hsa-miR-29a</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>hsa-miR-29b</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Pimozide</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>hsa-miR-29c</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>hsa-miR-124</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Perphenazine</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Figure 33: A 8 X 8 dense submatrix found by greedy randomized multi-domain algorithm for the validated data set**

Figure 33 shows one of the most dense submatrix obtained by running the incremental greedy randomized algorithm for square submatrices on the validated data set, with drug/microRNA ratio = 0.3 and threshold 0.95.
<table>
<thead>
<tr>
<th>Drug</th>
<th>Drug’s target gene</th>
<th>Interacting gene/common pathway</th>
<th>MicroRNA’s target gene</th>
<th>MicroRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluphenazine</td>
<td>CALM1</td>
<td>4137 (MAPT microtubule-associated protein tau)</td>
<td>IQGAP1</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Fluphenazine</td>
<td>DRD2</td>
<td>84687 (PPP1R9B protein phosphatase 1, regulatory (inhibitor) subunit 9B [ Homo sapiens ])</td>
<td>GNAI3</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Fluphenazine</td>
<td>CALM1</td>
<td>ErbB1 downstream signaling</td>
<td>IQGAP1</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Fluphenazine</td>
<td>CALM1</td>
<td>Lissencephaly gene (LIS1) in neuronal migration and development</td>
<td>IQGAP1</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Fluphenazine</td>
<td>CALM1</td>
<td>Signaling events mediated by VEGFR1 and VEGFR2</td>
<td>IQGAP1</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Pimozide</td>
<td>CALM1</td>
<td>4137 (MAPT microtubule-associated protein tau)</td>
<td>IQGAP1</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Pimozide</td>
<td>DRD2</td>
<td>84687 (PPP1R9B protein phosphatase 1, regulatory (inhibitor) subunit 9B [ Homo sapiens ])</td>
<td>GNAI3</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Pimozide</td>
<td>CALM1</td>
<td>ErbB1 downstream signaling</td>
<td>IQGAP1</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Pimozide</td>
<td>CALM1</td>
<td>Lissencephaly gene (LIS1) in neuronal migration and development</td>
<td>IQGAP1</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Pimozide</td>
<td>CALM1</td>
<td>Signaling events mediated by VEGFR1 and VEGFR2</td>
<td>IQGAP1</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Perphenazine</td>
<td>CALM1</td>
<td>4137 (MAPT microtubule-associated protein tau)</td>
<td>IQGAP1</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Perphenazine</td>
<td>DRD2</td>
<td>84687 (PPP1R9B protein phosphatase 1, regulatory (inhibitor) subunit 9B [ Homo sapiens ])</td>
<td>GNAI3</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Perphenazine</td>
<td>CALM1</td>
<td>ErbB1 downstream signaling</td>
<td>IQGAP1</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Perphenazine</td>
<td>CALM1</td>
<td>Lissencephaly gene (LIS1) in neuronal migration and development</td>
<td>IQGAP1</td>
<td>hsa-miR-124</td>
</tr>
<tr>
<td>Perphenazine</td>
<td>CALM1</td>
<td>Signaling events mediated by VEGFR1 and VEGFR2</td>
<td>IQGAP1</td>
<td>hsa-miR-124</td>
</tr>
</tbody>
</table>
11 Conclusion

Consolidating the results obtained from the validated dataset, it is seen that the microRNA hsa-miR-124 is scored 90 in its pairing with Fluphenazine, Pimozide and Perphenazine. The microRNA hsa-miR-124 is at a distance of 3.162278 from Perphenazine, 3.0 from Pimozide and 3.0 from Fluphenazine in the n-dimensional space. The three are found in the same cluster amongst the clusters formed by K-means clustering. They are also found in each other’s vicinity in the self organizing feature map. These 3 drugs and the microRNA are also part of the most dense submatrices found out by the greedy randomized incremental algorithm. The other drugs included in the submatrices can be studied with the different genes/pathways that they are related to. Using these results as substantial evidence of similarity in nature and behavior, these pairs can be taken up for experimentation to see if the microRNA has-miR-124 can be used as the drug combinatorial with or alternative to Fluphenazine, Pimozide and Perphenazine. This eliminates random sampling from a large permutation of drug microRNA pairs. Similar drug microRNA pairs can be discovered by using predicted targets too and then taken up for experimentation according to the various drug categories. If the empirical results are positive, the new knowledge could be of invaluable importance in the domain of medicine.
12 Future Work

1. The process can be refined by incorporating other clustering mechanisms like fuzzy clustering mechanism, support vector machines etc.

2. The data can introduce other pathways like chemical pathways.

3. Other databases can be used and their results can be compared.

4. Other distance metrics can be used with the data and permuted and the results can be compared (A little of this was done already).

5. The randomized portion of the algorithm can be replaced with exhaustive search and the results can be compared.

13 References


