Creating an Automated Prediction Model for Chemotherapy Treatment Regimens through Molecular Feature Selection

Divya Vatsa

under the direction of
Prof. Gil Alterovitz
Biomedical Cybernetics Laboratory
Harvard Medical School

Research Science Institute
August 2, 2016

Abstract

Cancer is one of the leading causes of death around the world. Chemotherapy regimens are one of the most common treatment options, in which a combination of drugs is administered to patients to combat the tumor cells. However, it is impossible to doctors to test out every combination of drugs to identify the most effective regimens. The objective of this study was to create a model for computationally predicting the most effective drug combinations for regimens. We achieved this by analyzing molecular drug attributes and overlaying them on a historical outcomes network for Chronic Myeloid Leukemia. Ultimately, we were able to determine specific molecular properties that are likely to influence the efficacy of a drug in chemotherapy regimens (polar surface area and acidic properties) and properties that appear to have very little impact on determining a drug's efficacy (water solubility and polarizability). This model allows us to understand drug properties that are directly involved in efficacy such that we can target drugs based on specific physicochemical traits.

Summary

While cancer continues to affect millions of people around the world, the treatment options are not as effective as they should be. This study specifically looks at chemotherapy regimens, the third most common treatment option with an estimated 15% tumor recurrance rate. Chemotherapy regimens consist of a combination of drugs that are administered to patients to stop the cancerous cells from spreading in the body. However, it is impossible for doctors to identify the most effective treatments, because there are an unlimited number of drug combinations. Therefore, this study attempted to solve this problem computationally by understanding the molecular features of drugs and how they relate to the efficacy of drugs when used in chemotherapy regimens. We compared the molecular analysis to an outcomes analysis of Chronic Myeloid Leukemia and looked for correlations between the molecular and outcomes data. Through our model, we were able to determine specific molecular properties that are most likely to influence the efficacy of a drug in chemotherapy regimens and properties that appear to have very little impact on determining a drug's efficacy. Understanding what properties of drugs are directly involved in efficacy is critical to identifying which combination of drugs will be most effective as chemotherapy regimens. Ultimately, we hope to use this model to improve the treatment options for cancer patients around the world.

1 Introduction

Cancer continues to be one of the most elusive diseases in regards to treatment, impacting millions of people of all races and ages. Despite the growing prevalence of cancer around the world, there remains substantial room for improving treatment options. This study specifically analyzes the effectiveness of chemotherapy, the third most common cancer treatment, with the aim of improving the prescription process for chemotherapy regimens. Chemotherapy is a systemic treatment that uses various combinations of drugs to target and destroy tumors throughout the body [4]. Using a mixture of drugs is especially important because tumors can arise from a variety of genetic mutations; therefore, the various drugs are designed to combat the various potential mutations [5]. Although such an approach should theoretically successfully account for the heterogeneity of cancer mutations, the variable efficacy and significant number of tumor reccurences (reported as roughly 15% across all cancers) highlights a problem with the current model [6]. This issue arises due to the heterogeniety of cancer compounded with the impracticality of testing every variation of drug regimens [1]. As a result, the current system of chemotherapy regimens is still very rudimentary. Based on the doctors′ understanding of the regimens and the patient′s conditions, prescriptions are needlessly variable and do not guarantee patients the best possible treatment.

Previous research has attempted to aid in this problem through outcomes analyses of historical regimen data [8]. The common method of statistical evaluation has been traditional meta-analyses [9]. Meta-analyses are correlational models that are harnessed to compare the efficacy of clinical trials for various cancer regimens. However, the traditional meta-analyses have many limitations that prevent the scientific community from gathering insights on the efficacy of different regimens [10]. The traditional models can only compare two trials at one time, which is tedious and does not provide a high-level understanding of how all the trials interact with each other. Additionally, only direct comparisons can be made between two trials [12]. This is problematic because methodological differences are most likely present between two regimen trials [13]. As a result, traditional models can only work with very limited data from the trials and are not an accurate method of determining efficacy of regimens [14].

Recent studies have started to use a more robust method called network meta-analysis. Network meta-analyses are a state-of-the-art approach that circumvent the limitations of the traditional models [16]. The network model can evaluate multiple trials at once because it is not restricted by direct comparisons [17]. Instead, the newer analysis uses both direct and indirect data from the trials [18]. Indirect data includes the heterogeneous variables in studies such as population, outcomes, timing, and setting study [19]. Furthermore, while the direct data must be taken from a single trial, indirect data can be taken from randomized controlled trials for each type of regimen. Therefore, more data is retained in the comparisons, allowing for a more comprehensive analysis. The network meta-analysis ultimately creates a network visualization of the efficacy of each drug used in the regimens and how each drug is connected to other drugs. The networks consist of clusters and edges that connect clusters together [2]. For the purposes of creating networks for regimens, the clusters represent individual drugs while the edges depict the synergy between the drugs. Through this design, it becomes apparent which combination of drugs will have the most positive effect and which combination of drugs will have the most negative effect. Thus, the current approach is most effective for creating a outcomes analysis of various regimens, and has gained popularity in understanding effective treatments for various cancers such as esophageal, small cell lung, and prostate cancer [3]. As a result, the network meta-analysis has improved our understanding of which combination of drugs is likely to be most effective.

However, the goal of this research is to create a new model that goes a step beyond the network meta-analysis. Due to the heterogeneity of cancer mutations, an outcomes analysis is not enough to predict how to treat each variation of cancer. Ultimately, to create the most comprehensive model to automate predictions for the optimal regimen, we plan to overlay molecular data onto the outcomes analysis. Through this approach, we can account for all the possible variations that affect the success of regimens. This present study focuses on the first step of overlaying molecular data on an outcomes network, by which the molecular network can capture the mechanisms that actually cause certain drugs to be more effective than others. Pharmaceutical drugs are complex molecules that have many unique physicochemical attributes, which determine a drug′s efficacy [7]. The general research community breaks down the main attributes of drugs into four gold-standard models: the Rule of Five (RO5), Ghose Filter, Vebers Rule, and MDDR-like Rule [11]. These four models include various pharmacological properties that most accurately characterize the majority of drugs, such as water solubility, physical charge, rotatable bonds, and polarizability [15]. By analyzing the drug attributes outlined by the four models, we can get a comprehensive idea of what molecular properties are directly responsible for inhibiting cancer growth.

The outcomes analysis used in this study is a published network meta-analysis approach for chronic myeloid leukemia (CML) regimens [20]. Taking data from the past 40 years of regimens for CML, factors such as area of inhibition, amount of time free of cancer post therapy, and populations sizes for treatments were considered to determine which regimens were most effective [Figure 1].

Ultimately, this current study aims to devise a new model that will expand on the network meta-analysis and provide insights on the molecular mechanisms that make certain drugs more effective as chemotherapy treatments than others. The significance of this approach is that it will aid doctors and CML patients by systematizing the prescription process for chemotherapy regimens and removing the human biases that currently exists in the practice. Additionally, the approach can benefit and progress rational drug design, because the focus on the molecular mechanisms highlights specific drug properties for both known and unknown drugs that have high correlations to efficacy. Lastly, this correlational model for CML and its

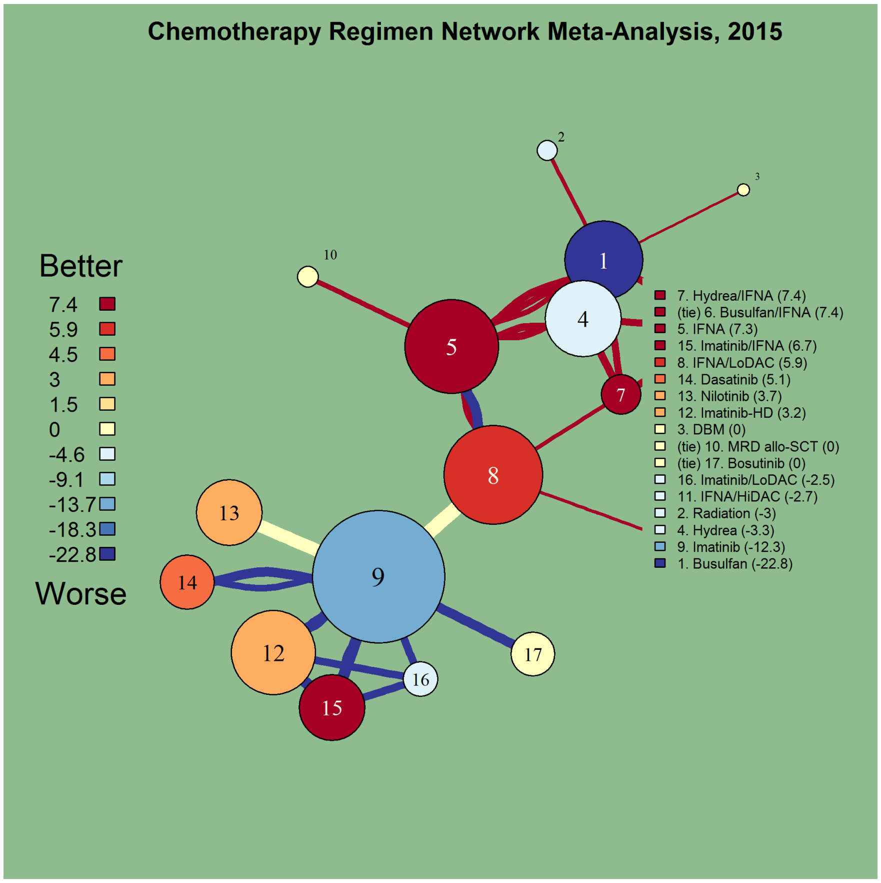


Figure 1: CML Outcomes Network. Clusters represent individual drugs, and edges represent the direction of interaction between each type of drug. Red edges and clusters show greater efficacy. Blue edges and clusters show lesser efficacy.

corresponding regimens can be expanded to improve the efficacy of chemotherapy regimens across multiple cancers.

2 Methods

2.1 Data Collection

We collected a list of drugs used in the regimens studied in the CML study as well as the National Cancer Institutes list of approved drugs for CML. A drug database called DrugBank was accessed to get information about the attributes outlined in the RO5, Ghose

Filter, Vebers Rule, and MDDR-like Rule models. Ultimately, 14 attributes or features were recorded for 25 drugs [Tables 1 and 2].

Table 1: List of Attributes Recorded

Molecular Weight	Water Solubility
Partition Coefficient (logP)	Logarithm Solubility (logS)
Strongest acidic (pK_a)	Strongest basic (pK_a)
Physical Charge	Hydrogen Acceptor Count
Hydrogen Donor Count	Polar Surface Area
Rotatable Bonds	Refractivity
Polarizability	Number of Rings

Table 2: List of the Drugs Used

Busulfan	Hydrea	Asparaginase
Ponatinib Hydrochloride	Blinatumomab	Mercaptopurine
Imanitib	Vincristine	Low and High Dose Imanitib
Nilotinib	Dasatinib	Dibromomannital
Bosutinib	Interferon-α	Prednisone
Methotrexate	Nelarabine	Daunorubicin
High Dose ara-C	Pegaspargase	Vincristine Sulfate
Cyclophosphamide	Clofarabine	Doxorubicin

2.2 Clustering Physiochemical Data through K-Means Algorithm

The data for each drug was then clustered for each specific attribute. Rather than cluster all features together like conventional standards, individual cluster graphs per attribute are necessary to find correlations between a single attribute and the outcomes analysis. Through R, we used the k-means clustering unsupervised machine learning algorithm. K-means clustering separates data points into "k" specified clusters that are centered around the closest "k" cluster means. We specified the function to group the data into six clusters per attribute. We selected six clusters because that appeared to maximize the variance of the data points while still creating significant groupings. For each attribute, we generated visual cluster graphs.

2.3 Correlation of Cluster Graphs to Outcomes Analysis

The drugs in the outcomes network we separated into three categories, namely positive correlation, weak correlation, and negative correlation drugs. For each of the six clusters per feature, we recorded the number of drugs in the cluster that belonged to each of the three categories [Figure 2]. Then, the distribution of the data points was determined by dividing the number of data points per category by the number of clusters they were separated into for each attribute′s cluster graph. This distribution will be refered to as the closeness variable throughout this paper. The closeness variables were normalized so that values from the three different correlation categories could be compared.

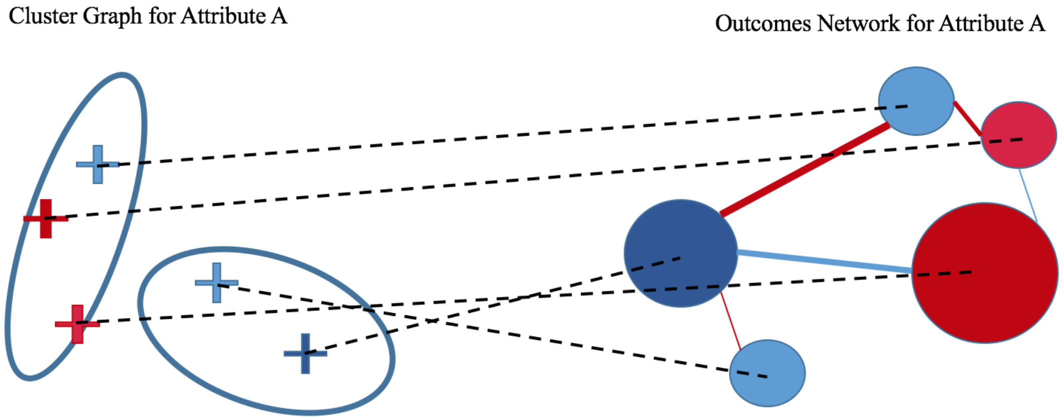

Cluster Graph for Attribute A
Outcomes Network for Attribute A

Figure 2: Correlating Cluster Graphs to Outcomes Analysis

3 Results

3.1 Closeness Scores for Drug Attributes

Ultimately, only 19 drugs were analyzed from the original 25. DrugBank did not have the necessary information for six drugs, so they could not be used in the study. Further, one of the drugs that was common to the positively correlated category in the outcomes network (interferon-α) was one of the six drugs that was removed. As a result, the positively correlated drugs were not represented in the clusters, and we could not produce any closeness variables for that data. After analyzing the data, four attributes had the most interesting results: water solubility, most acidic, polar surface area, polarizability.

Table 3: Closeness Variable Scores for Drug Attributes

Attribute	Weakly Correlated Group	Negatively Correlated Group
Molecular Weight	1.714	1
Water Solubility	3.429	1.5
LogP	1.714	1
LogS	1.714	1
Strongest acidic (pK$_a$)	0.857	1.5
Strongest basic (pK$_a$)	1.714	1.5
Physical Charge	1.714	1.5
Hydrogen Acceptor Count	1.143	1
Hydrogen Donor Count	1.143	1
Polar Surface Area	1.714	3
Rotatable Bonds	1.714	1.5
Refractivity	1.714	1
Polarizability	3.429	1
Number of Rings	1.714	1.5

3.2 Correlation between Drug Attributes and Original Outcomes Network

The water solubility and polarizability attributes had the largest closeness variable score ($closeness = 3.429$) for the weakly correlated drugs category [Figures 3 and 4]. The z-scores were calculated for both attributes ($z = 2.18$). This indicates that because the closeness variables for the two attributes are more than two standard deviation points away from the mean, they are outliers in the data and, therefore, significant.

The polar surface area attribute follows the same trend as the previous two features.

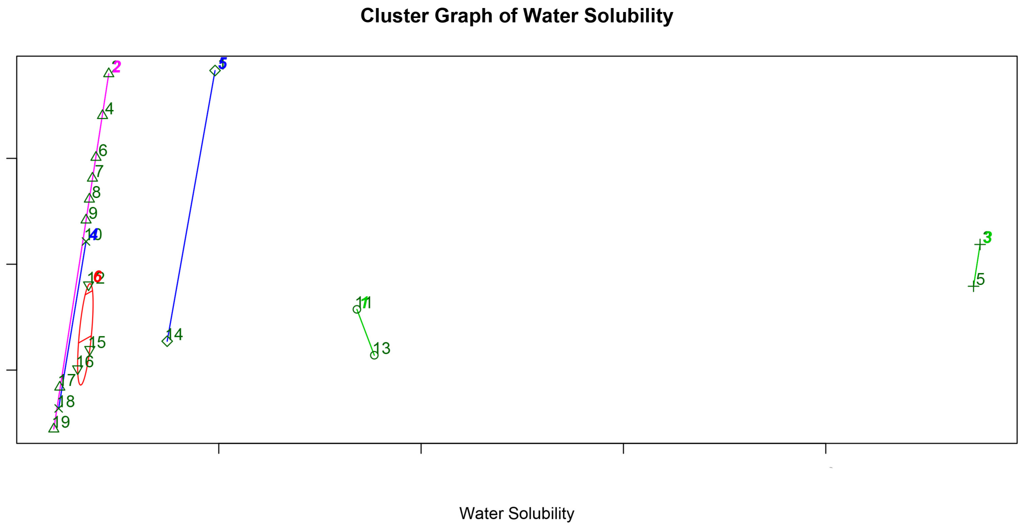


Figure 3: Cluster Graph of Water Solubility.

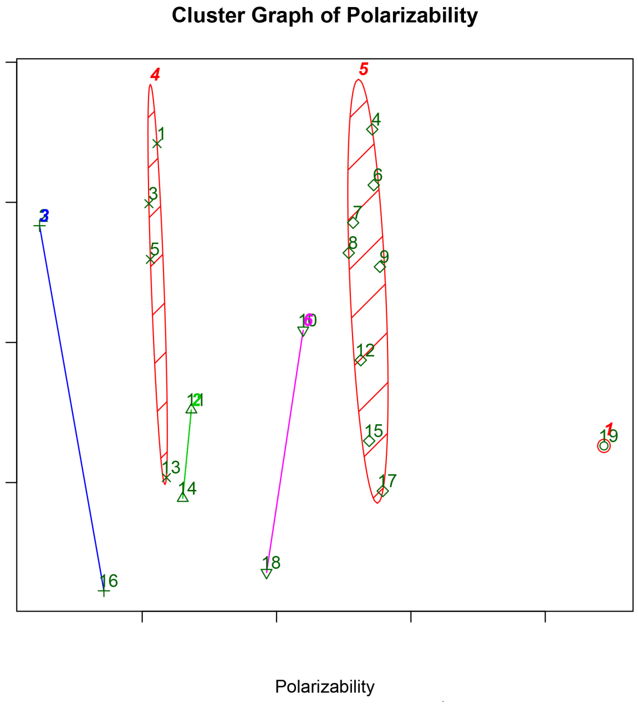


Figure 4: Cluster Graph of Polarizability.

However, the polar surface area had the highest closeness variable score ($closeness = 3$) among the negatively correlated category [Figure 5]. The z-score of $z = 3.07$ indicates that the score for polar surface area is far removed from the mean score for the negatively correlated drugs.

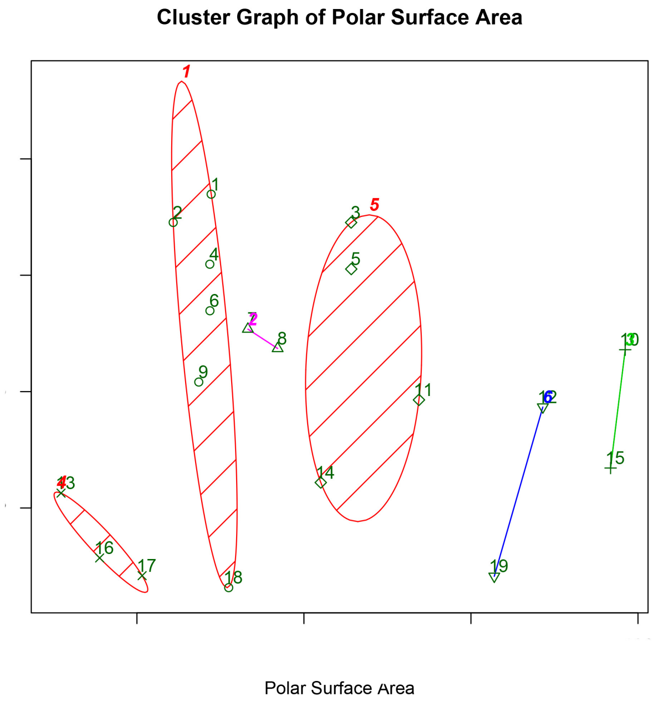


Figure 5: Cluster Graph of Polar Surface Area.

Another interesting result was found for the pK$_a$ (most acidic) and polar surface area features regarding the weakly correlated and negatively correlated group interaction. For all of the other attributes, the normalized closeness variable was larger for the weakly correlated category than for the negatively correlated group. For the pK$_a$ and polar surface area features, it was actually the opposite case. We ran a t-test to determine if the group differences are significant between the two categories ($t = 1.88, p = 0.07$). Although the differences between the weakly and negatively correlated groups are not significant ($p > 0.05$), the findings are still notable.

4 Discussion

The present study is among the first to understand the molecular mechanisms responsible for the efficacy of chemotherapeutic drugs. The most striking finding is that even with a limited number of drugs being analyzed, the model can already isolate certain attributes that have strong correlations to the outcomes analysis from historical data. Additionally, this research provides an understanding of interactions between drugs in the strongly and weakly correlated groups from the outcomes network. The current analysis points to the polar surface area and pK_a (most acidic) attributes as most directly responsible for determining the efficacy of a drug.

4.1 Within Group Findings

From the negatively correlated group, it was determined that the polar surface area attribute significantly had the highest closeness score. A high closeness score indicates that a large number of negatively correlated drugs were grouped in the least number of clusters. In the case of polar surface area, three of the drugs from the negatively correlated group were grouped in the same cluster. A high closeness score signifies a stronger correlation to the outcomes network. As a result, it can inferred that polar surface area is strongly correlated to the outcomes network for the negatively correlated group of drugs. The clustering data provides further insights on the actual quantitative properties that the drugs share for a specific attribute. The three drugs from the negatively correlated category were all grouped in the first cluster, which had a cluster mean of 208.3. Therefore, the model indicates to us that drugs with polar surfaces areas similar to 208.3 are likely to follow the trend of the negatively correlated drugs from the outcomes model and be ineffective in chemotherapy regimens.

From the weakly correlated group, water solubility and polarizability had the highest closeness variables scores. Therefore, these two features are most strongly correlated to the outcomes network. However, the weakly correlated category of drugs indicates that those drugs have little impact when used in chemotherapy regimens. Because water solubility and polarizability are significant for the weakly correlated category of drugs, we can determine that the two features are least likely to have an impact on the efficacy of chemotherapy regimens.

4.2 Between Group Findings

Between the weakly correlated and negatively correlated groups, we noticed a trend that the weak correlation had a higher closeness variable score than the negtive correlation category. However, pK$_a$ (most acidic) and polar surface area showed an opposite result; the negative correlation group had a higher closeness score. Although the group differences are not significant, this finding is still interesting because the data was normalized so between group comparisons can still be made. For the majority of the attributes, a higher closeness score for the weakly correlated group indicates that those features are not as directly responsible for drug efficacy. For pK$_a$ (most acidic) and polar surface area, however, the negatively correlated group had a higher closeness score. Therefore, those two attributes can be interpreted as having a greater role in determining the efficacy of a drug.

5 Conclusion

The main goal of this study was to create a new model for predicting the most effective drug combinations for chemotherapy regimens. We aimed to achieve this by creating cluster models for various molecular drug attributes and overlaying them on a historical outcomes network for CML. Through our model, we were able to determine specific mechanisms such as polar surface area and pK$_a$ (most acidic) that are most likely to influence the efficacy of a drug in chemotherapy regimens. We also found attributes such as water solubility and polarizability that appear to have very little impact on determining a drug′s effects in a regimen. The significance of this model of molecular attribute selection is profound because it overcomes the physical limitations of testing every single possibility of regimen variations in a wet lab. Instead, this model allows us to understand what properties of drugs are directly involved in efficacy such that we can target drugs based on their physicochemical similarity to the isolated attributes. The future directions of this research is to expand the model to other complex cancers and overlay the genomic data of patients. Thereby, we would understand how various genetic mutations interact with molecular drug properties and predict which combination of drugs would be most effective for specific mutations. With this knowledge, we can advance chemotherapy treatments and improve the lives of the millions of individuals who suffer from cancer around the world.

6 Acknowledgments

I would like to extend my sincere thanks to my mentor, Professor Gil Alterovitz, for giving me the wonderful opportunity of working in his lab and providing assistance and support for this research study. I would also like to thank my tutor, Connor Duffy, and the TAs for providing guidance and feedback on this paper. I would like to acknowledge the Center for Excellence in Education, the Research Science Institute, and MIT for giving me this unique opportunity to expand my interest in scientific research. Lastly, I recognize Mr. John Yochelson, President at Building Engineering and Science Talent; Dr. Melissa Flagg; Dr. Robin Staffin, Director for Basic Research at Office of the Assistant Secretary of Defense for Research and Engineering (ASD(A & E)); Dr. and Mrs. Rajiv Saxena; Mr. Abijith Krishnan; and Dr. Vandana L. Madhavan, without whom this experience would not have been possible.

References

[1] A. Mandal. History of chemotherapy. 2014.

[2] G. A. Jeremy Warner, Peter Yang. Automated sythesis and visualization of a chemotherapy treatment regimen network. 2013.

[3] Y. Yamanishi et al. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Oxford Journals*, 2010.

[4] D. Singh. Defining desirable natural product derived anticancer drug space: optimization of molecular physicochemical properties and admet attributes. 2016.

[5] J. DR. Meta-analysis: Weighing the evidence. *Stat Med*, 1995.

[6] L. A et al. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Ann Intern Med*, 2009.

[7] A. Arora and E. M. Scholar. Role of tyrosine kinase inhibitors in cancer therapy. *The Journal of Pharmacology*, 2005.

[8] Cancer treatment and survivorship facts and figures. *American Cancer Society*, 2015.

[9] J. W. Goldwein and C. Vachani. Chemotherapy: The basics. *OncoLink*, 2016.

[10] D. NR et al. Hyperthermia and radiotherapy with or without chemotherapy in locally advanced cervical cancer: A systematic review with conventional and network meta-analyses. *Int J Hyperthermia*, 2016.

[11] J. W. Goldwein and C. Vachani. Chemotherapy: The basics. *OncoLink*, 2016.

[12] S. LM et al. Mortality, cardiovascular risk, and androgen deprivation therapy for prostate cancer: A systematic review with direct and network meta-analyses of randomized controlled trials and observational studies. *Medicine (Baltimore)*, 2016.

[13] Z. Y et al. Optimized selection of three major egfr-tkis in advanced egfr-positive non-small cell lung cancer: a network meta-analysis. *Oncotarget*, 2016.

[14] M. Talpaz et al. Re-emergence of interferon-α in the treatment of chronic myeloid leukemia. *Leukemia*, 2013.

[15] S. E. Schaeffer. Graph clustering. *Elsevier*, 2007.

[16] S. G et al. Evaluating the quality of evidence from a network meta-analysis. *PLoS*, 2014.

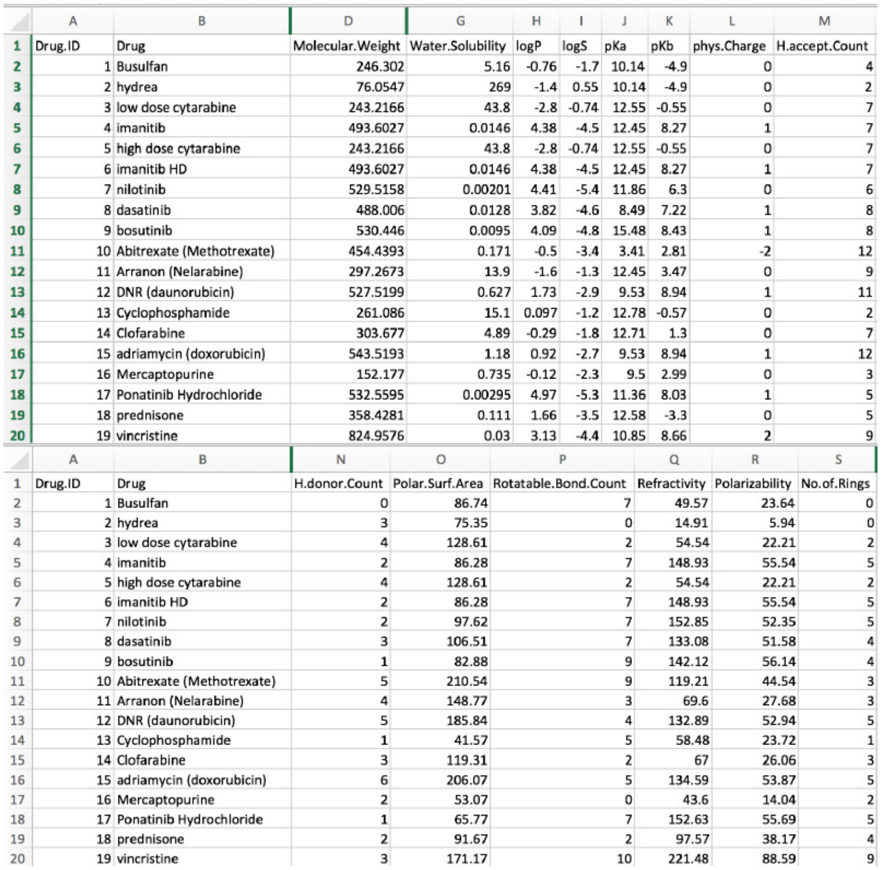
[17] B. M et al. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol*, 1999.

[18] W. V. D. Noortgate and P. Onghena. Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *SAGE Journals*, 1999.

[19] H. TC et al. Systematic review and network meta-analysis: neoadjuvant chemoradiotherapy for locoregional esophageal cancer. *Japanese Journal of Clinical Oncology*, 2015.

[20] S. I. Berger and R. Iyengar. Network analyses in systems pharmacology. *Bioinformatics*, 2009.

Appendix A Figures from Methods and Results



	A	B	D	G	H	I	J	K	L	M
1	Drug.ID	Drug	Molecular.Weight	Water.Solubility	logP	logS	pKa	pKb	phys.Charge	H.accept.Count
2	1	Busulfan	246.302	5.16	-0.76	-1.7	10.14	-4.9	0	4
3	2	hydrea	76.0547	269	-1.4	0.55	10.14	-4.9	0	2
4	3	low dose cytarabine	243.2166	43.8	-2.8	-0.74	12.55	-0.55	0	7
5	4	imanitib	493.6027	0.0146	4.38	-4.5	12.45	8.27	1	7
6	5	high dose cytarabine	243.2166	43.8	-2.8	-0.74	12.55	-0.55	0	7
7	6	imanitib HD	493.6027	0.0146	4.38	-4.5	12.45	8.27	1	7
8	7	nilotinib	529.5158	0.00201	4.41	-5.4	11.86	6.3	0	6
9	8	dasatinib	488.006	0.0128	3.82	-4.6	8.49	7.22	1	8
10	9	bosutinib	530.446	0.0095	4.09	-4.8	15.48	8.43	1	8
11	10	Abitrexate (Methotrexate)	454.4393	0.171	-0.5	-3.4	3.41	2.81	-2	12
12	11	Arranon (Nelarabine)	297.2673	13.9	-1.6	-1.3	12.45	3.47	0	9
13	12	DNR (daunorubicin)	527.5199	0.627	1.73	-2.9	9.53	8.94	1	11
14	13	Cyclophosphamide	261.086	15.1	0.097	-1.2	12.78	-0.57	0	2
15	14	Clofarabine	303.677	4.89	-0.29	-1.8	12.71	1.3	0	7
16	15	adriamycin (doxorubicin)	543.5193	1.18	0.92	-2.7	9.53	8.94	1	12
17	16	Mercaptopurine	152.177	0.735	-0.12	-2.3	9.5	2.99	0	3
18	17	Ponatinib Hydrochloride	532.5595	0.00295	4.97	-5.3	11.36	8.03	1	5
19	18	prednisone	358.4281	0.111	1.66	-3.5	12.58	-3.3	0	5
20	19	vincristine	824.9576	0.03	3.13	-4.4	10.85	8.66	2	9

	A	B	N	O	P	Q	R	S
1	Drug.ID	Drug	H.donor.Count	Polar.Surf.Area	Rotatable.Bond.Count	Refractivity	Polarizability	No.of.Rings
2	1	Busulfan	0	86.74	7	49.57	23.64	0
3	2	hydrea	3	75.35	0	14.91	5.94	0
4	3	low dose cytarabine	4	128.61	2	54.54	22.21	2
5	4	imanitib	2	86.28	7	148.93	55.54	5
6	5	high dose cytarabine	4	128.61	2	54.54	22.21	2
7	6	imanitib HD	2	86.28	7	148.93	55.54	5
8	7	nilotinib	2	97.62	7	152.85	52.35	5
9	8	dasatinib	3	106.51	7	133.08	51.58	4
10	9	bosutinib	1	82.88	9	142.12	56.14	4
11	10	Abitrexate (Methotrexate)	5	210.54	9	119.21	44.54	3
12	11	Arranon (Nelarabine)	4	148.77	3	69.6	27.68	3
13	12	DNR (daunorubicin)	5	185.84	4	132.89	52.94	5
14	13	Cyclophosphamide	1	41.57	5	58.48	23.72	1
15	14	Clofarabine	3	119.31	2	67	26.06	3
16	15	adriamycin (doxorubicin)	6	206.07	5	134.59	53.87	5
17	16	Mercaptopurine	2	53.07	0	43.6	14.04	2
18	17	Ponatinib Hydrochloride	1	65.77	7	152.63	55.69	5
19	18	prednisone	2	91.67	2	97.57	38.17	4
20	19	vincristine	3	171.17	10	221.48	88.59	9

Figure 6: Data Mining with DrugBank.

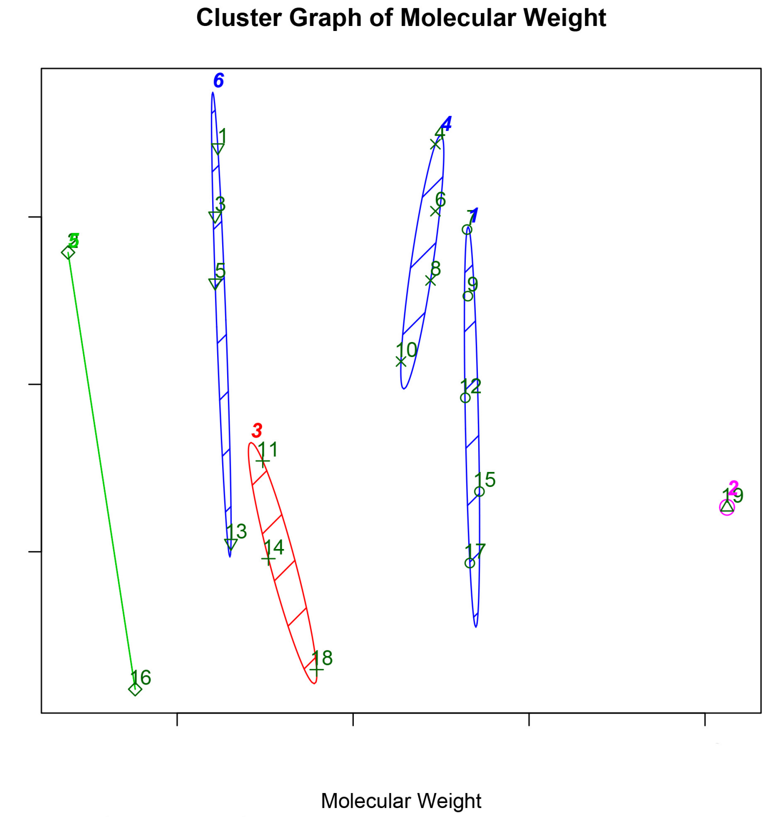


Figure 7: Molecular Weight Cluster Graph

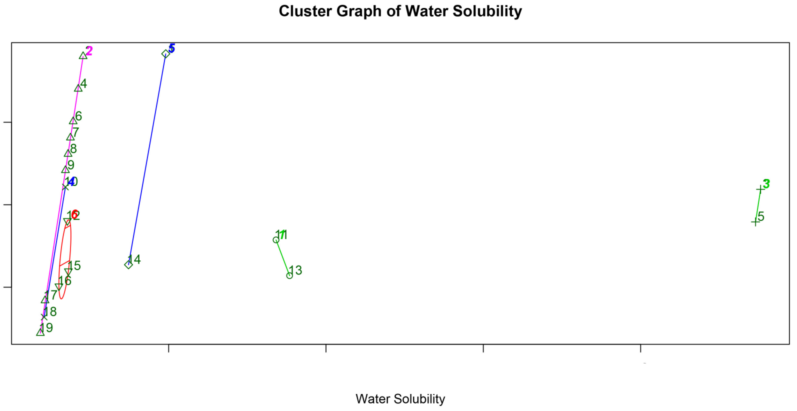


Figure 8: Water Solubility Cluster Graph

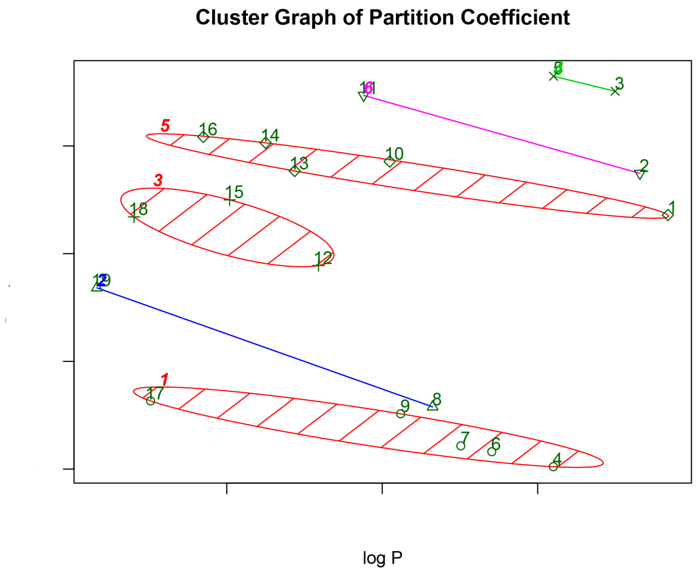


Figure 9: Partition Coefficient Cluster Graph

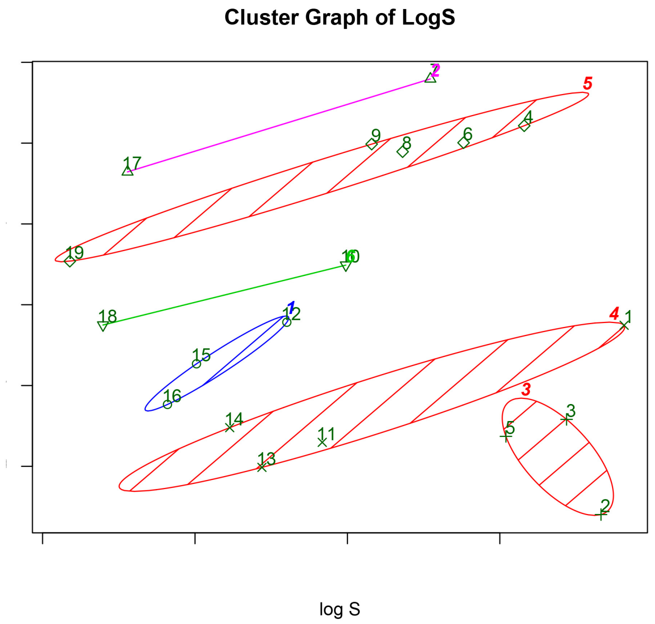


Figure 10: Logarithmic Solubility Graph

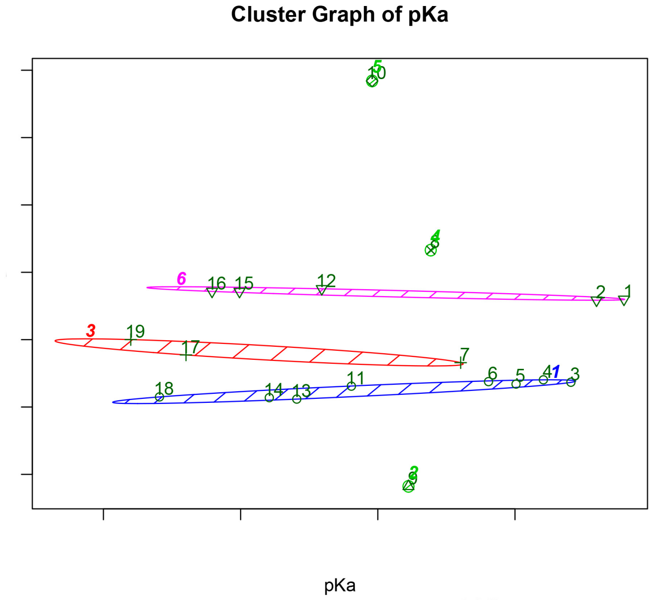


Figure 11: Most Acidic Cluster Graph

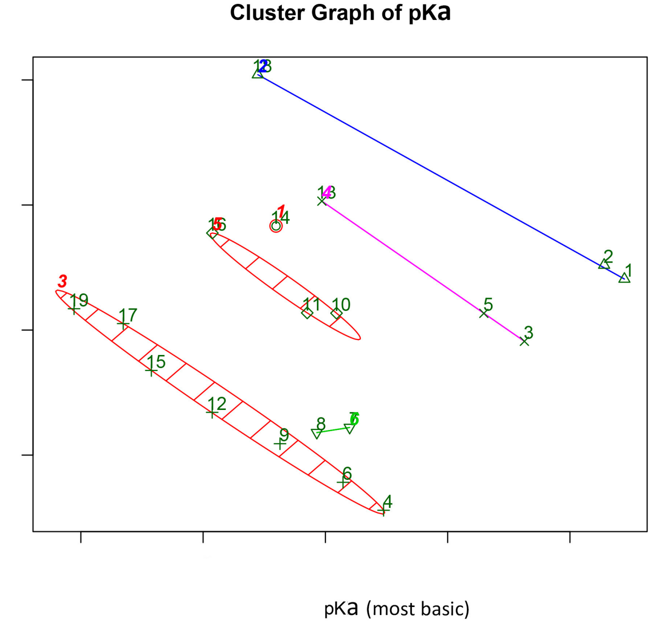


Figure 12: Most Basic Cluster Graph

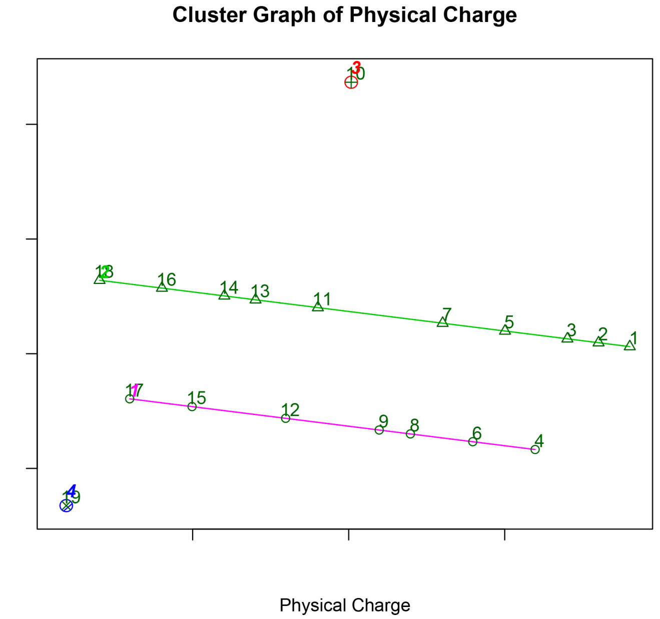


Figure 13: Physical Charge Cluster Graph

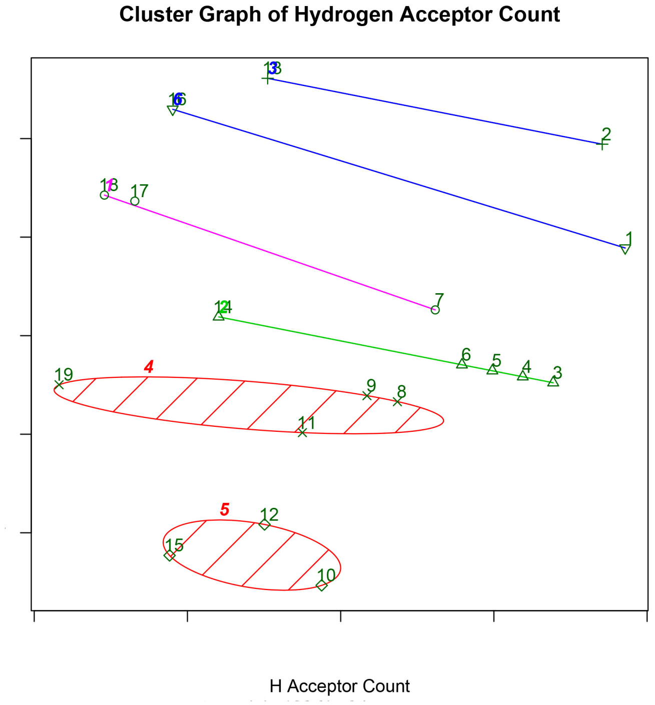


Figure 14: Hydrogen Acceptor Count Cluster Graph

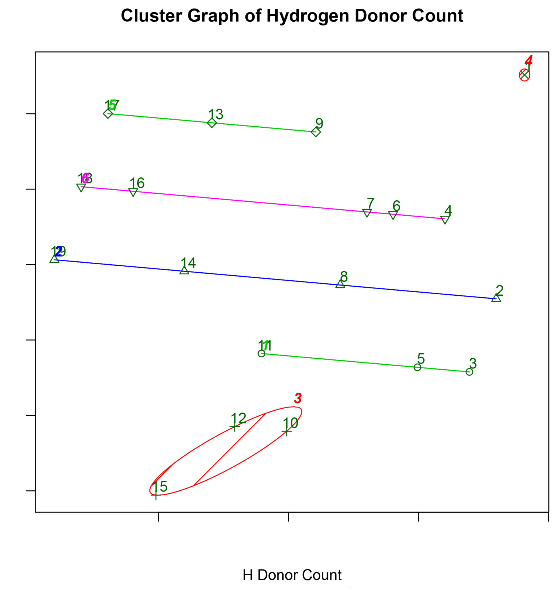


Figure 15: Hydrogen Donor Count Cluster Graph

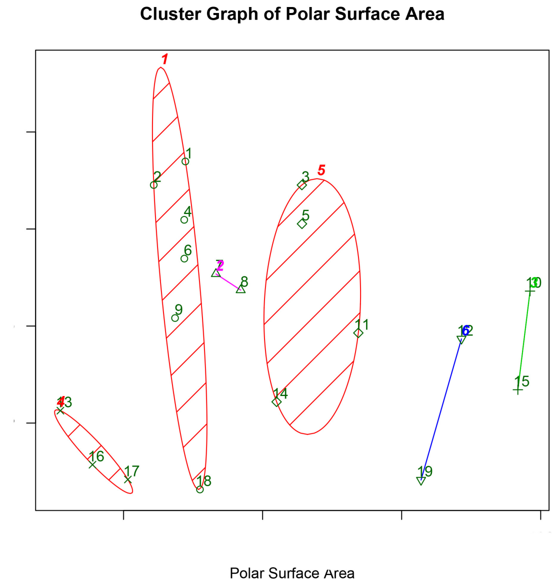


Figure 16: Polar Surface Area Cluster Graph

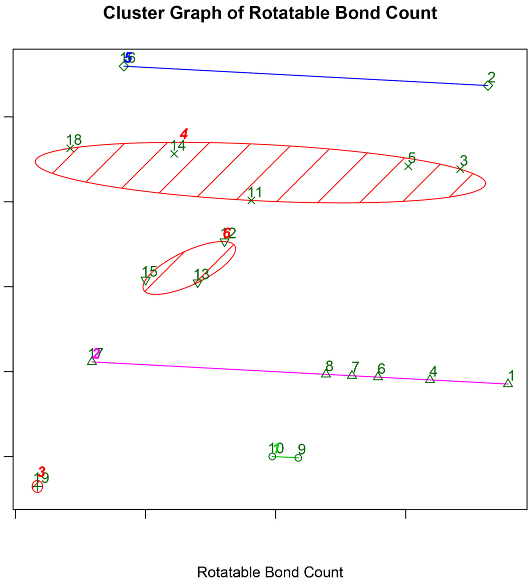


Figure 17: Rotatable Bond Count Cluster Graph

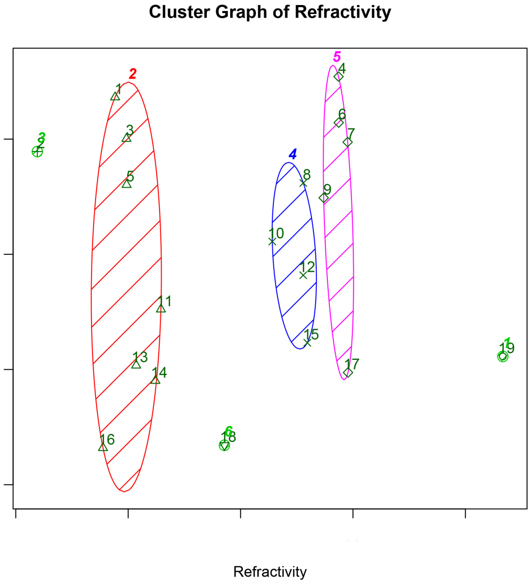


Figure 18: Refractivity Cluster Graph

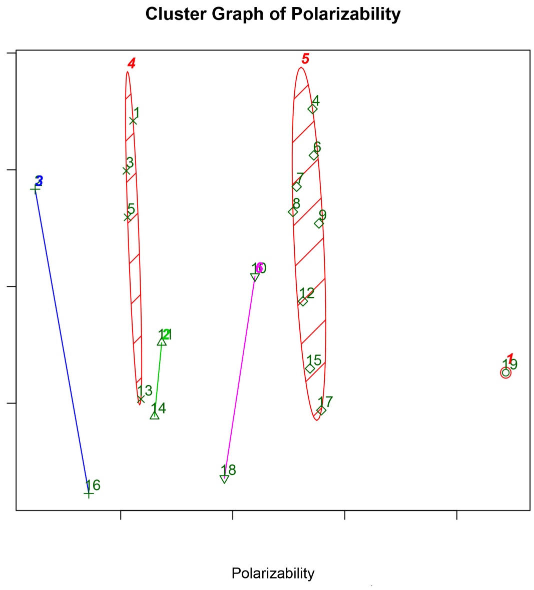


Figure 19: Polarizability Cluster Graph

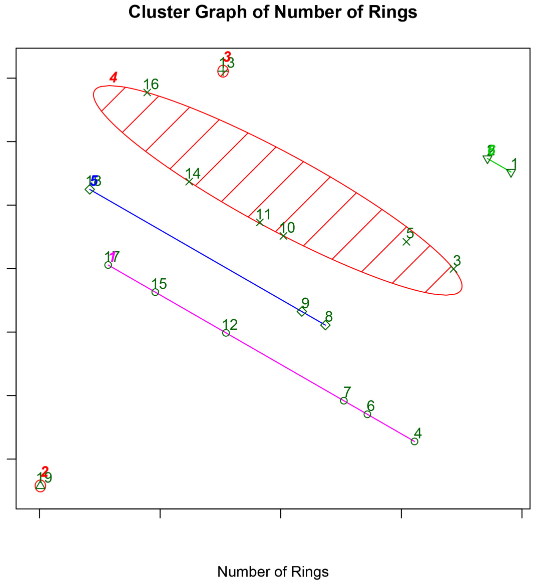


Figure 20: Number of Rings Cluster Graph