Turkish Journal of Computer and Mathematics Education *Vol. 9 No.02 (2018), 488-495*
DOI: https://doi.org/10.17762/turcomat.v9i2.13853
Research Article

An ML-based Cancer Genome Profile Drug Prediction Framework

Rahul Chauhan
Asst. Professor, Department of CSE (Computer Sc), GEHU-Dehradun Campus

Abstract: This article discusses predictive modelling for individualised cancer therapy. We employ machine learning to predict pharmaceutical reactions, drug synergies, drug target-interactions, and cancer classification. This work aims to construct machine learning prediction models for drug sensitivity prediction, medicine combination therapy, drug target interaction prediction, and cancer classification. C-HMOSHSSA, a cancer classification framework using multi-objective meta-heuristic and machine learning, predicts both recognised and new cancer biomarkers. A hybrid feature selection algorithm (HMOSHSSA) for gene selection improves on the multi-objective spotted hyena optimizer (MOSHO) and salp swarm algorithm (SSA). Four classifiers are trained using the HMOSHSSA dataset. The approach uncovers informative gene groupings. KSRMF also predicted missing drug response values. The BE-DTI framework uses dimensionality reduction and active learning to predict drug-target interactions. Active learning helps under-sampling bagging ensembles. High-dimensional data demands unique dimension reduction strategies. Five existing (RF, SVM) feature-based approaches are compared to the proposed framework's performance..

Keywords -Drug Target Interaction Prediction, Ensemble Learning, Dimensionality Reduction, Active Learning

Introduction

Cells are the fundamental structural and functional units of all known living things. They have a unique set of genes that allow them to do a wide variety of sophisticated tasks that set them apart from other people. The genetic material, or gene, is the fundamental structural and functional unit of heredity and the information carrier inside the cell. The variation in genotype and phenotype seen across animals may be traced back to differences in underlying genes. Genes contain the instructions needed to build an organism's phenotype. Since the beginning of genetic research, genetics has grown into its own scientific area. The development of bioinformatics has improved the treatment process for many genetic illnesses and extended the life expectancy of patients. It is now much easier to diagnose serious illnesses like diabetes, cancer, and heart attacks. The healthcare sector, which has already offered lab-on-a-chip equipment, sees chip technology as its future. The genetic profiles of patients may be accurately assessed with the use of these chips. These recent developments in medical technology [3] are facilitating the early identification and prognosis of life-threatening disorders like cancer. Scientists studying genetics have discovered how some traits are passed down from one generation to the next. They are also researching gene expression to identify which stimuli (external or internal) cause certain genes to be up- or down-regulated. We may use statistical and computational methods to conduct a wide range of analyses on this gene expression data. There is a wealth of other omic data (genome, transcriptome, and proteome) that may be used in conjunction with gene expression data, such as copy number variations, gene mutations, etc. Drug pathway analysis, drug target identification, identifying illness biomarkers, and disease categorization all rely heavily on gene expression data. Scientists and researchers are working tirelessly to uncover the underlying mechanisms that may one day aid in the accurate detection and treatment of illnesses like cancer [4, 5, 6, 7]. Such data-driven analysis is getting a helping hand from data mining and machine learning techniques.

Gene Expression Data

Gene expression value refers to the ratio of a gene's expression levels in two distinct environments, as determined by DNA microarray hybridization. Protein synthesis is aided by a process called gene expression, which includes reading instructions from the genome. The gene expression value is the quantity of mRNA (messenger ribonucleic acid) the gene produces at a given period. Values assigned to genes during expression may change in response to both internal and external stimuli, as well as the presence or absence of biological regulators and pathways. The substance known as messenger RNA (mRNA) facilitates the transfer of genetic instructions for protein creation. There are two sub-processes at play here: transcription and translation. The

creation of an RNA transcript is known as transcription. In order to drive the manufacture of the protein it encodes, this copy, known as a messenger RNA (mRNA) molecule, must first leave the cell nucleus and reach the cytoplasm. During protein synthesis, the sequence of amino acids is translated from a messenger RNA (mRNA) molecule. The genetic code provides an explanation for how a gene's sequence of base pairs translates into a specific chain of amino acids. The ribosome, located in the cytoplasm of the cell, reads the mRNA sequence in blocks of three bases at a time in order to construct the protein. The results of many types of biological studies benefit from information about gene expression. It helps distinguish across articulations of a phenotype by providing a map from genotype characteristics to phenotypic traits. It's utilised to identify possible illness biomarkers and categorise disease phenotypes. Genomic analyses, such as m-RNA, DNase-seq, and MNase-seq, provide these information to machine learning models. Exploiting this potential, researchers have generated a surge of new findings on cancer and other chronic illnesses. Cancer is a complex illness with many different forms. Systems or procedures that may aid in early detection and prognosis of cancer type are urgently needed. Several novel strategies for studying and treating cancer have emerged throughout the last decade[13]. Several computational and biological methods are suggested in the literature[14,15] for early cancer detection. Researchers aren't only focusing on finding new biomarkers; they're also trying to figure out how to use computational (in-silico) models and algorithms to forecast how different drugs will react to different diseases and how different targets will react to different drugs.

The accumulation of big cancer data banks has boosted investigation into this field. The likelihood that the tumour is malignant has been predicted using machine learning methods.

Cancer Classification

The interpretation of biological relevance and the association of genes with illnesses using gene expression patterns has been in use for decades. These profiles come from a wide range of patients and are collected in a number of distinct biological settings. By comparing expression patterns in a healthy and cancerous setting, we may learn more about the nature of the illness. Gene expression, or the status (active or inactive) of a gene, is described by the amount of mRNA generated by that gene at a given time. Analysis of microarray data and further study into cancer categorization have been prompted by the development of biological computational tools. Cancer diagnosis and prognosis rely heavily on accurate classification of tumour samples and their subtypes. It aids in the accurate prediction of cancer kinds and the subsequent identification of therapeutic therapies tailored to sub-types of cancer. Classification strategies based on gene expression data have been developed by a number of authors[11]. Cancer classification strategies range from statistical methodologies to machine learning systems. Because of the high dimensionality of gene expression data, classification is a challenging issue, and most classifiers begin with a genes selection phase [12]. It aids in reducing the complexity of the categorization process in terms of both time and accuracy by eliminating superfluous characteristics. Existing "feature selection algorithms" are limited in their capacity to scale and generalise; a classifier constructed using a single feature selection approach on a single dataset may not perform well when applied to other datasets. Automatic feature extraction and the construction of generic, scalable classifiers are two areas where Deep Neural Networks (DNN) [13] may lend a hand. The development of DNA microarray technology has had a profound effect on scientific inquiry in the biological sciences. Researchers may examine the biological activity and importance of thousands of genes simultaneously. In addition, microarray technology allows the parallel screening of genomic profiles, yielding valuable insights into numerous genetic variants and modifications. It aids in the early diagnosis of serious disorders like cancer. Many scientists over the last two decades have helped advance cancer research by contributing standard microarray datasets for a wide range of tumour types. Thousands of genes from various samples are included in the cancer databases. In the several methods suggested for cancer classification based on genetic profiles, these data sets are used as standards. In order to classify cancers, several computer methods have been developed.

Drug Response Prediction

Mutations and variations in tumour cell genes cause cancer, making it a hereditary illness. Mutations in genes have direct effects on cellular functionality because they alter the genes responsible for a variety of cellular processes. Most mutations result from contact with a hostile environment that promotes tumour development. Cancer is a challenging illness to treat because of the complexity of the tumour microenvironment. Patients with the same cancer kind respond differently to the same targeted therapy. Individuals' unique genetic makeups account for these variances in reaction. It is not possible to give the best possible cancer treatment choices based just on the location of the tumour. Taking a patient's unique genetic profile into account, precision medicine strives to provide individualised care that slows or stops the spread of cancer[14]. Although it is difficult, researchers are making strides in identifying the best therapy for different malignancies. Multiple high-throughput drug testing on a massive scale have shown a correlation between patient genes and therapeutic success. Pharmacogenomics databases are generated from these screenings, which include data on a vast number of human cancer cell lines and their responses to various drugs. Cancer Cell Line Encyclopaedia (CCLE) [16] and Genomics of Drug Sensitivity in Cancer (GDSC) [15] are two such comprehensive databases with the same overarching goal of advancing cancer research. These data sets are crucial to current drug development since they are used to forecast drug (responses/combinations/repositioning). Computational approaches need to be developed so that these massive screening datasets can be used to create accurate prediction models. Using the correlation between preexisting malignant genetic profiles and treatment responses, one of the crucial challenges is to forecast which medicines would be effective against a certain cell line.

Literature Review

Chen et al. 2016 have introduced the gene selection strategy for relevant gene clustering based on kernel functions. Weighted learning employs adaptive distance and identifies optimum weights for genes in an iterative process. Two classifiers (SVM, KNN) were used to evaluate the proposed method on eight open-source datasets. The suggested method benefits from not needing to optimise any parameters.

Guoli et al. 2011 partial least squares (PLS) technique has been described as a new gene selection and tumour diagnosis algorithm. Different tumour datasets are used to verify the suggested method, which uses the linear kernel support vector machine.

Martin et al. 2013 use phosphorylation of many proteins and cell receptivity to propose a network model for dedifferentiated liposarcoma (DDLS). Synergy between CDK4 and IGFIR inhibitors was predicted to be mediated through the AKT pathway.

Huang et al. 2009 often referred to as "DrugComboRanker" to predict drug interactions. Topological relatedness between targets, drug-induced changes in gene expression profiles, and gene ontology similarity score are all taken into account to arrive at the synergic score. Researchers have looked at other means of determining synergy scores outside PPI networks. Protein synthesis, for example, may be thought of as the end result of a series of sub-molecular processes or interactions. While PPI methods may aid in the identification of synergistic medication combinations, they cannot provide light on the mechanism by which such effects are generated. Drug synergism is a complex phenomenon, but it may be understood via the routes involved in synergy approaches. Studies of pathways often use a network representation, with nodes standing for components (proteins, genes, metabolites), and edges reflecting interactions between nodes as a function of time derivative.

Polynikis et al. 2009 have summarised many methods for synthesising gene regulatory networks. Several suggested mathematical frameworks use differential equations. Transitions between stable states of different intermolecular interactions may be mapped using differential equations. Additionally, the advantages of different modelling approximations on system dynamics have been examined..

Zhang et al. 2007 designed an innovative approach for selecting features to use in analysing the cancer microarray dataset. The suggested method selects important genes via the use of relevance analysis and the discernibility matrix.

Cancer Classification Using Genomic Profiles

Our goal in this study is to use a bio-inspired meta-heuristic method with machine learning to create a framework for predicting useful and novel cancer biomarkers. The suggested system employs the salp swarm algorithm[16] and the multi-objective spotted hyena optimizer[14]. These algorithms were chosen for their ease of use and quick convergence to the global optimum. The researchers have identified issues with convergence and variety in multi-objective optimisation problems in the actual world. Therefore, it is important to create an algorithm that simultaneously promotes convergence and variety. In order to preserve variety, the salp swarm method is implemented here. However, the cost of keeping records is a burden for SSA. However, MOSHO demands little in the way of computing resources, therefore it is used for data storage. As a result, we present a new hybrid method that draws on elements of both SSA and MOSHO. Here we will introduce some of the fundamental concepts and historical context of multi-objective optimisation methods. In addition, the suggested framework's use of the multi-objective spotted hyena optimizer and SSA is further upon.

Multi-objective Spotted Hyena Optimizer (MOSHO)

The spotted hyena optimisation (SHO) [20] inspiration, the multi-objective spotted hyena optimizer (MOSHO), is explored here. In SHO, the multi-objective optimisation method is proposed with reference to the social behaviour of hyenas. This method modelled an optimisation algorithm after the social and hunting dynamics of spotted hyenas. They also added two additional features to the original SHO: an archive system and a group-based filtering system. The MOSHO algorithm's key benefits over other methods are its strong convergence behaviour and its ability to avoid local optima..

The new components that are included in MOSHO are discussed below:

(a) Archive The collection has the best Pareto optimum solutions. Concave, convex, and unconnected Pareto fronts equally distribute it. Two more parts: Grid, controller.

Archive controller The controller chooses which solutions to archive. Archive update rules: • An empty archive accepts the present solution.

• If a person dominates the archive, the answer is rejected.

• The solution is archived if none of the external population factors dominates it.

• The archive removes solutions dominated by the new element.

(c) Grid An adaptive grid method can locate the Pareto borders. Four subspaces define an objective function. The grid locates people in ungridded areas. Hypercubes form the grid.

(d) Group selection Multi-objective search space's biggest issue is comparing new answers to old ones. Solution: group selection. The group selection method randomly selects the best answer from a pool of options and introduces it to the less densely populated search area.

Experimental Analysis

The purpose of this proposed study is to extrapolate the single-agent dosage response to estimate the effectiveness and synergy of medication combinations. Dose-response curves for a single drug have been shown to be sufficiently predictive of combinational responses [17]. High-throughput drug screening, as described by Held et al. [18], is applied in this study. Forty medications are specifically discussed in the context of BRAF-

melanomas, out of a total of 150 single agent dose-response data. Twenty-seven RAS mutant, RAS wild-type (WT), BRAF mutant, and BRAF wild-type (WT) cell lines are used in the pharmacological screen.

Table 1: Performance comparison of machine learning models in the prediction of synergy labels

Model Name	Accuracy	Specificity	Senstivity
Random Forest	0.8222	0.9091	0.7941
Neural Network	0.7333	0.7105	0.8571
Support Vector	0.7778	0.8889	0.7500
LM	0.7556	0.7353	0.8182
M5P3	0.7111	0.7135	0.6970
Decision Stump	0.6667	0.6782	0.5652
foba	0.7022	0.7150	0.6250
cubis	0.6800	0.6891	0.6250
Decision Tree	0.6978	0.7500	0.6891

By determining the average and standard deviation of each cell line's response to a single dosage of the medication, features may be generated for each drug combination. Most computer models used to foretell medication interactions rely heavily on feature extraction. Prediction accuracy may be increased and the synergy process can be better understood with the use of highly responsive features. In this case, characteristics are estimated in a manner that allows for taking use of the genetic effect of each cell-line on medication combination. This process yielded a total of 54 characteristics, one for each possible drug-and-combination pairing. First, we used a dataset of 750 medication combinations perturbed on RAS and BRAF melanomas to train several machine learning earning models, as shown in Table 2. Accuracy, specificity, and sensitivity are only few of the performance metrics compared across several machine learning models in Table 3. In addition, four distinct partition sets are used during model training to counteract any potential bias introduced by the training-testing split. Among the several machine learning models employed for training, the random forest model performed the best. The lower the mistake or false prediction rate, the more accurate and precise it is. The random forest synergy model has an accuracy of 0.8222 and a specificity of 0.9091.

Table 2: Performance comparison of all 9-models using the different training-testing partition

Models	Training-Testing Partition			
	50-50%	**60-40%**	**70-30%**	**80-20%**
Random Forest	(0.8167,0.7847)	(0.8023,0.7758)	(0.8218,0.7795)	(0.8222,0.7941)
Neural Network	(0.7226,0.8218)	(0.7174,0.8318)	(0.7230,0.8415)	(0.7333,0.8571)
Support Vector	(0.7567,0.6478)	(0.7519,0.7596)	(0.7647,0.7678)	(0.7778,0.7500)
LM	(0.7118,0.7816)	(0.7252,0.7963)	(0.7361,0.7861)	(0.7556,0.8182)
M5P3	(0.6896,0.5787)	(0.6994,0.6486)	(0.7002,0.6689)	(0.7111,0.6970)
Decision Stump	(0.6159,0.4899)	(0.6027,0.4956)	(0.6208,0.4890)	(0.6667,0.5652)
foba	(0.7089,0.6052)	(0.6986,0.5986)	(0.7089,0.6123)	(0.7022,0.6250)
cubis	(0.6548,0.6577)	(0.6689,0.6664)	(0.6672,0.6559)	(0.6800,0.6250)
Decision Tree	(0.7089,0.6473)	(0.6987,0.6323)	(0.7074,0.6537)	(0.7178,0.6891)

DOI: https://doi.org/10.17762/turcomat.v9i2.13853
Research Article

Table 3: Performance of models trained using random forest dataset

Model	Accuracy	Specificity	Sensitivity
BRAF-Synergy(Model-1)	0.8222	0.9091	0.7941
BRAF-Selective-Effective(Model-2)	0.8319	0.8963	0.7102

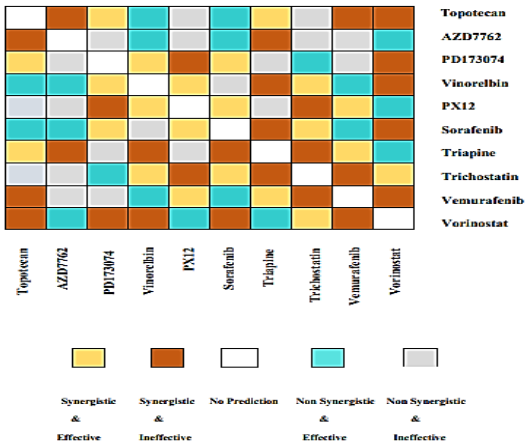

Figure 1: Subset of prediction results using Held et al. dataset

Figure 1 shows We evaluate the proposed method's robustness using a 10-fold cross-validation procedure and find that our method reliably predicts synergy and efficacy (Geno-type-selective).

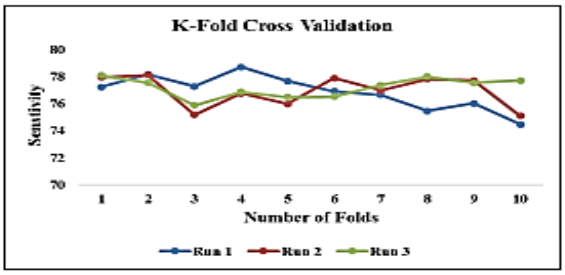

Figure 2: K-fold (K=10) cross validation of Random Forest using Sensitivity

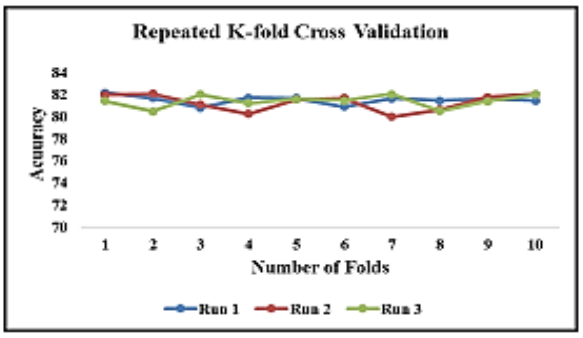

Figure 3: K-fold (K=10) cross validation of Random forest using Accuracy

The results of random forest's 10-fold cross validation with regard to sensitivity and accuracy are shown in Figures 2 and 3, respectively. It is important to stress that our models are developed and tested using BRAF (mutant) cell-lines as their primary data source. However, this does not limit the scope of the suggested approach. Considering the size of the drug screen, the suggested technique may narrow the search space and pinpoint synergistic medication combinations for treating different forms of cancer.

Conclusion

The suggested model accurately anticipates a wide range of medication combinations, as shown in the literature. The suggested strategy has the potential to narrow the search area and locate synergistic medication combinations that are useful in treating different malignancies. This technology has the potential to significantly enhance the prediction of novel medication combinations in light of the massive combinational drug screen. The goal of this effort is to use ML methods to simulate the process through which drugs operate together more effectively. Cancer patients may benefit most from a treatment plan that combines drugs that work well together. Future efforts to improve cancer therapy will need a deeper understanding of medication-disease interaction, which may be gained by extracting characteristics of possible drug combinations. And you can use ensemble machine learning to improve your predictions. Our method has to be tested on a wide range of cancer patients with different genetic profiles.

References

[1] H. Chen, Y. Zhang, and I. Gutman, "A kernel-based clustering method for gene selection with gene expression data," Journal of Biomedical Informatics, vol. 62, pp. 12–20, 2016

[2] L.-J. Zhang, Z.-J. Li, and H.-W. Chen, "An effective gene selection method based on relevance analysis and discernibility matrix," in Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 1088–1095, Springer, 2007

[3] G. Ji, Z. Yang, and W. You, "Pls-based gene selection and identification of tumor-specific genes," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 41, no. 6, pp. 830–841, 2011.

[4] D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander, "Class prediction and discovery using gene expression data," in Proceedings of the fourth annual international conference on Computational molecular biology, pp. 263–272, ACM, 2000.

[5] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, and C. Lau, "Prediction of central nervous system embryonal tumour outcome based on gene expression," Nature, vol. 415, no. 6870, pp. 436–442, 2002.

[6] D. Beer, S. Kardia, C. Huang, A. Gautam, Z. Li, and G. Bepler, "Ten best readings," Group, vol. 343, pp. 1217–1222, 2000. [103] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, and C. Peterson, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," Nature Medicine, vol. 7, no. 6, pp. 673–679, 2001.

[7] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia," Nature Genetics, vol. 30, no. 1, pp. 41–47, 2001.

[8] N. R. Pal, K. Aguan, A. Sharma, and S.-i. Amari, "Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering," BMC Bioinformatics, vol. 8, no. 5, pp. 1–18, 2007.

[9] K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick, "Gene selection: a bayesian variable selection approach," Bioinformatics, vol. 19, no. 1, pp. 90–97, 2003

[10] K. L. Tang, W. j. Yao, T. H. Li, Y. x. Li, and Z. W. Cao, "Cancer classification from the gene expression profiles by discriminant kernel-pls," Journal of Bioinformatics and Computational Biology, vol. 8, pp. 147–160, 2010.

[11] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," arXiv preprint arXiv:1606.05718, pp. 1–6, 2016.
[12] H. Liao, "A deep learning approach to universal skin disease classification," University of Rochester Department of Computer Science, CSC, 2016.
[13] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," in Proceedings of the International Conference on Machine Learning, vol. 28, pp. 1–7, ACM New York, USA, 2013.
[14] A. Chinnaswamy and R. Srinivasan, "Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data," in Innovations in Bio-Inspired Computing and Applications, vol. 424, pp. 229–239, Springer, 2016.
[15] S. S. Shreem, S. Abdullah, and M. Z. A. Nazri, "Hybridising harmony search with a markov blanket for gene selection problems," Information Sciences, vol. 258, pp. 108–121, 2014
[16] L.-Y. Chuang, C.-H. Yang, J.-C. Li, and C.-H. Yang, "A hybrid bpso-cga approach for gene selection and classification of microarray data," Journal of Computational Biology, vol. 19, no. 1, pp. 68–82, 2012.
[17] F. V. Sharbaf, S. Mosafer, and M. H. Moattar, "A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization," Genomics, vol. 107, no. 6, pp. 231–238, 2016.
[18] G. R. Zimmermann, J. Lehar, and C. T. Keith, "Multi-target therapeutics: when the whole is greater than the sum of the parts," Drug Discovery Today, vol. 12, no. 1-2, pp. 34–42, 2007
[19] L. Huang, F. Li, J. Sheng, X. Xia, J. Ma, M. Zhan, and S. T. Wong, "Drugcomboranker: drug combination discovery based on target network analysis," Bioinformatics, vol. 30, no. 12, pp. 228–236, 2014.
[20] A. Polynikis, S. Hogan, and M. di Bernardo, "Comparing different ode modelling approaches for gene regulatory networks," Journal of Theoretical Biology, vol. 261, no. 4, pp. 511–530, 2009.
[21] M. L. Miller, E. J. Molinelli, J. S. Nair, T. Sheikh, R. Samy, X. Jing, Q. He, A. Korkut, A. M. Crago, and S. Singer, "Drug synergy screen and network modeling in dedifferentiated liposarcoma identifies cdk4 and igf1r as synergistic drug targets," Science Signaling, vol. 6, no. 294, pp. 1–14, 2013.