Active Learning Framework of informative p53 cancer rescue mutants

Haohan Wang, Yucong Yang, Varuni Gang

I. INTRODUCTION AND BACKGROUND

1.1 General Overview

For our final project we worked on a protein called P53 which is a tumor suppressor protein. This protein plays an important role in cell cycle control and apoptosis. A defective p53 protein could allow proliferation of abnormal cells thus resulting in cancer. It has been found by previous research that 50% of all human tumors contain p53 mutants[1].

Mostly there is a low presence of p53 protein in normal cells. These come into action when there are stress signals within the body or in case of DNA damage as these are responsible for growth arrest, DNA repair and apoptosis (or as commonly known cell death). The growth arrest means arresting (or stopping) the progression of cell cycle which is responsible for replication of damaged DNA. Thus doing so P53 activates the transcription of proteins involved in DNA repair. If the cell containing damages DNA still persists then p53 helps in apoptosis.

While it is useful to activate p53, it is also advised to regulate concentration of p53 should also be tightly regulated because high level of p53 can accelerate the aging process by excessive apoptosis.

Introduction of p53 in p53 deficient cells has resulted in rapid death of cancer or prevention of further division, in in-vitro. Therefore it has been considered as most important drug target for the past few years. One of the most effective way to destabilize a network is by attacking its most connected node. Following this notion, researchers have found that p53 is one of the most connected node within cell cycle and thus knocking it out would cripple normal function of the cell, which is what's been exploited by by tumor cells.

There are many missense point mutations in p53 genes that could lead to a potential p53 mutant which is not effective to perform its normal function. And one of these functions is

[1] Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. (1991) *Science* 253 , 49-53.

the anti-cancer gene inducing effect. And thus restoring its function would be a major foot forward for scientist in the biological committee and human health as a whole.

1.2 Current progress and problem

Despite progress, the cure rate of cancers remains around 60%. Resistance of human cancers to standard treatments correlates with mutations of p53. Three quarters of p53 mutations result in full-length protein with a single amino acid change. Several hundred clinically important amino acid changes affect p53. These full-length p53 cancer mutants provide an exciting opportunity to specifically target cancers. Restoring normal function to mutant p53 would trigger apoptosis in infected cells, thus shrinking or killing the tumor.

There has been various strategies[2] proposed previously to restore p53 function in cancer cells that include finding molecules that that restore proper tumour suppressor activity of p53 in-vitro. These depend on altering the conformation of mutants of p53 back to its active form. As of now researchers have not been able to find a perfect molecule that can induce such effects but they have identified some lead molecules that are helpful in different situations.

These strategies only seek small molecules / compounds that will stabilize mutant p53 in a native-like conformation. There is another strategy in which the p53 cancer could be rescued by second-site cancer suppressor mutations. It has been found previously that these mutations lead p53 mutants to regain their active wild-type p53 function. Many studies[3] on this topic has already validated this approach. These studies help in uncovering the regions within p53 core domain which when altered would lead to rescue of its functions. Similarly its a known fact that combination of structural expects of p53 helps in understanding of the basic mechanisms of p53 functions.

Unfortunately, in vitro testing of all possible mutation combinations to determine their cancer rescue effects is infeasible due to time and expense. Therefore, it would be very desirable to have a computer model to run in silico experiments on virtual mutants. Such a model could narrow down the list of likely cancer rescue mutants to a number that reasonably could be assayed in the laboratory. To reach the desired predictive accuracy, such a classifier would need a larger training set of known mutants than was provided by the initial experimental screens. Thus its a current research topic to find the expensive

[2] Blagosklonny, M. V. (2002), P53: An ubiquitous target of anticancer drugs. Int. J. Cancer, 98: 161–166.
[3] Danziger SA, Baronio R, Ho L, Hall L, Salmon K, et al. (2009) Predicting Positive p53 Cancer Rescue Regions Using Most Informative Positive (MIP) Active Learning. PLoS Comput Biol 5(9): e1000498.

data points within these that when acquired next would lead to rapid discovery of biological function.

II. CURRENT METHODS

2.1 Strategies

These are some of the in vitro solutions to the current problem:
1) One strategy is to seek small molecule drugs that stabilize mutant p53 in a native-like conformation.

2) Another is to perform intragenic second-site suppressor mutations identify p53 cancer mutants that are likely to be amenable to functional rescue. Thus uncovering these regions of the p53 core domain and altering them would lead to functional rescue. Simultaneously, if this is combined with structural and other experimental studies then it would help to elucidate the basic mechanisms of p53 functional rescue, which can then be treated.

Unfortunately, in vitro testing of all possible mutation combinations to determine their cancer rescue effects is infeasible due to time and expense. Therefore, it would be very desirable to have a computer model to run in silico experiments on virtual mutants. Such a model could narrow down the list of likely cancer rescue mutants to a number that reasonable and could assayed in the laboratory.

2.2 Current Active learning score methods

Previously, there have been effort in using active learning techniques[4] to build classifiers using machine learning techniques. Most of these classifiers are built to choose most informative instances from a space of unlabeled instances. Active learning as an approach is used for p53 to select mutants that both quickly improves the classifier and quickly search for previously unknown cancer rescue mutations.

There are four active learning techniques[5] that have been used before. These are Maximum curiosity, Additive Curiosity, Minimum marginal hyperplane and maximum entropy.

2.2.1 Maximum curiosity

Maximum curiosity scores each new training set by its correlation coefficient (r). The r for each potential new training set is used to determine which unclassified mutants resulted in the largest increase of r. It assumes that the highest r for each mutant, m, occurs when that mutant is correctly paired with its true activity.

[4] Danziger, S.A., Baronio, R., Ho, L., Hall, L., Salmon, K., Hatfield, G.W., Kaiser, P., and Lathrop, R.H. (2009) Predicting Positive p53 Cancer Rescue Regions Using Most Informative Positive (MIP) Active Learning, PLOS Computational Biology, 5(9), e1000498
[5] Danziger, S.A., Zeng, J., Wang, Y., Brachmann, R.K. and Lathrop, R.H. (2007) Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants, Bioinformatics, 23(13), 104-114.

2.2.2 Additive Curiosity

Additive Curiosity is calculated by adding the scores for each training set. In it the Mutant chosen may be the most beneficial to the classifier regardless of its revealed activity.

2.2.3 Additive Minimum Marginal

Minimum Marginal Hyperplane scores training sets based on how far new unclassified mutants are from the boundary (support vector machine hyperplane) separating active from inactive mutants. Here margin could be considered as the distance from the new example to the hyperplane. It assumes that the unclassified mutants closest to the dividing hyperplane will be the most informative to the classifier once the true class is known.

2.2.4 Maximum Entropy

Maximum Entropy scores each training set by using the information theory concept of entropy (H). Entropy is calculated from the probability of class membership for each unclassified mutant, estimated by a support vector machine logistic regression algorithm. It assumes that the most informative mutants are those that the classifier is most uncertain about.

2.3 Current state of instrumentation automation

Computational analyses[6] used molecular model-based representations to create the component classifiers: (1D) genomic sequence, (2D) surface property maps, (3D) protein structure distance maps, and (4D) unfolding trajectories over time.

Information about the location of the mutation and the residue change was used to construct the set of 1D structure features. Secondary structure information of the mutation (alpha helix, beta sandwich, etc.) was recorded with its general location in the p53 core domain. The residue property change was recorded such as polarity, amino acid substitution, size, charge, aromaticity, hydrophobicity, and if in a DNA-binding region. This resulting in 247 features per mutant.

The 2D surface property maps were annotated with surface properties, such as electrostatics or h-bond donor/acceptor status. The molecular surface was mapped to a sphere, steric and depth information was recorded and the sphere was mapped to a plane. The resulting surface map was subtracted from the wild-type map, and a raw set of 4,883 steric surface map features and 4,895 electrostatic surface map features were extracted.

A structural mutation perturbs the molecular structure. The 3D distance map is an NxN matrix giving the Cartesian distance between N residue alpha carbons. It reflects structural shifts induced by the mutation. The p53 core domain has 197 residues resulting in a 197×197 matrix

[6] Danziger, S.A., Swamidass, S.J., Zeng, J., Dearth, L.R., Lu, Q., Chen, J.H., Cheng, J., Hoang, V.P., Saigo, H., Luo, R., Baldi, P., Brachmann, R.K. and Lathrop, R.H. (2006) Functional census of mutation sequence spaces: the example of p53 cancer rescue mutants, IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM, 3, 114-125.

that may be collapsed to a distance vector giving the magnitudes of the distance changes . The resulting 197 length vector map had 3 features for each residue, the directional i, j, and k vectors. This resulted in 591 features per mutant.

The thermodynamic stability of a p53 mutant is an important determinant of cancer. The unfolding of a molecular model in a simulated heat bath is related to thermodynamic stability. The 4D data tracks the 3D structure of the molecule over time.

III. PROPOSED METHOD

3.1 Strategy overview

In this section, we propose a framework for active learning that can be used in any membership model active learning which does not consider predicted class as a criterion.

The goal of our active learning strategy is to solve the problem within the dataset, where the number of positive instances are much greater than the number of negative instances.

During the instances selection process within our algorithm, positive instances are given preference more than negative, however, while the algorithm labels more and more data and the number of positive instances reaches a certain threshold based on the proportion of all the training dataset, the active learning strategy decreases the preference towards positive instances and in the process selects more negative instances.

As a result, we get a relatively balanced training set relative to both positive and negative instances leading to results showing good performance relative to high precision as well as high recall.

To improve performance at each iteration, the active learning strategy dynamically select a subset of features based on the labeled dataset labeled by the active learning algorithm.

3.2 Implementation Strategy

The implementation of the active learning algorithm was performed using Java using machine learning software namely WEKA[7]. WEKA's implementation of Information Gain was used for feature selection while SMO was used as the appropriate classifier.

In our implementation, 50 positive instances and 180 negative instances were selected for initialization. In every round of labeling, 5 instances were selected to label.

[7] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Similar algorithm was also developed using Matlab to improve classifier's accuracy and to validate results.

3.3 Dataset Description:

Our dataset consists of 16772 instances. Each instances are all the mutants for p53. Each instance has a total of 5409 attributes. The attributes 1-4826 represent 2D electrostatic and surface based features, attributes from 4827-5408 represent 3D distance based features and attribute 5409 is the class attribute, which is either active or inactive. The class labels are to be interpreted as follows: 'active' represents transcriptionally competent, active p53 whereas the 'inactive' label represents cancerous, inactive p53. Class labels are determined experimentally.

3.4 Active Learning Strategy

3.4.1 Pseudocode for our strategy

In this section, we will introduce the details of our active learning strategy. Pseudo code for the algorithm is as follows:

Initialization:

Select M features with information gain

randomly select N_p positive instances and N_n negative instances

Main Loop:

While:

1. **Instance Selection**: Select n instances with highest score_k for Instance k,
   score_k = score_classifier + score_balanced
   score_balanced = (#_negative/#_all)^4
2. **Feature selection**: select M features with information gain

3.4.2 Detailed explanation of the algorithm

During the initialization process, a fixed amount of positive and negative instances were selected randomly. These are then used to perform feature selection to select a fixed amount of features to train the classifier. These features are selected based on information gain.

These steps are looped over and over again until and unless a threshold based on the number of positive instances are attained:

Step 1. Instance Selection

1) Calculate Active Curiosity Score set S1 for each instance by labeling it active in training data

and using new train data set to perform 10-fold cross-validation.

Repeat Step 2,3,4,5,6: (till active learner collects 50% of positive instances)

Step 2 :

Label each instance with inactive (i.e. 0) to get Inactive Curiosity scores set S2.

$$r_{score} = \frac{t_p t_n - f_p f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}}$$

Step 3 :

Find the Maximum value for each instance in Set 1 and Set 2 and consider it as its score

$$Score_{ins} = \max \begin{bmatrix} r_{active} \\ r_{inactive} \end{bmatrix}$$

Step 4 :

Use the original train data set to predict each unlabeled instance, if the instance is a positive instance, a weight is added to its score. The value of the weight is equal to the ratio between true positive and true negative in the training set

$$if(predict_{org}(ins) = active)$$

$$score\,s_{ins} = score\,s_{ins} + ratio(\frac{t_p}{N})$$

where N is total number of the instances.

Step 5 :

The instance with the highest score in the previous step is added into the training set.

Step 6 : Features Selection

Again perform feature selection (based on Information Gain/correlation) to select M features from all other features (i.e. 5409).

We select a subset of features with information gain as following:

$$IG(T,a) = H(T) - H(T \mid a)$$
$$= H(T) - \sum_{v \in val(a)} \frac{|\{x \in T \mid x_a = v\}|}{|T|} H(\{x \in T \mid x_a = v\})$$

where:

$$H(T) = P(T) \log_2 \frac{1}{P(T)}$$

IV. RESULTS

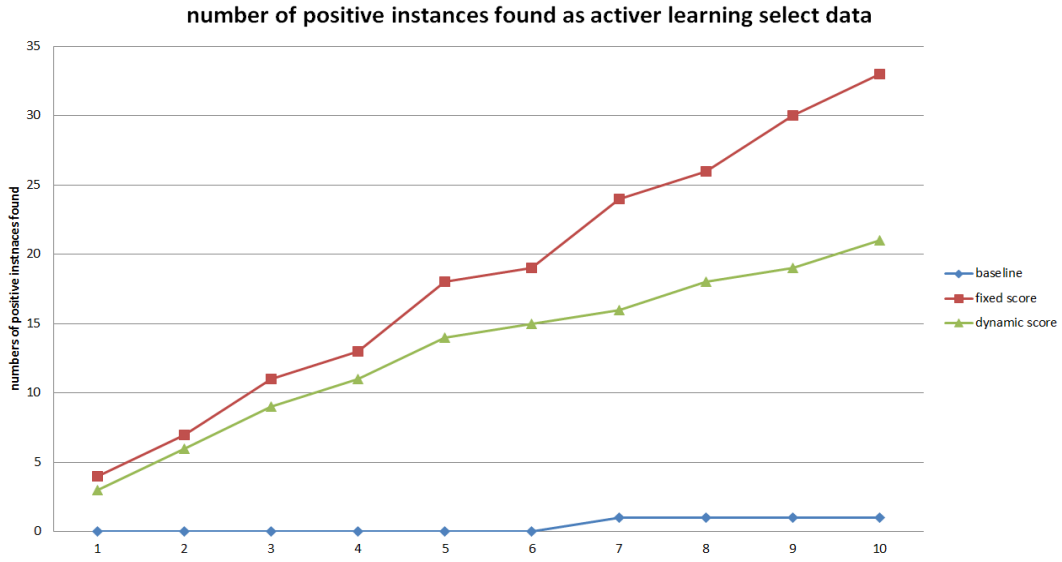

Figure 1 Number of positive instances in each iteration

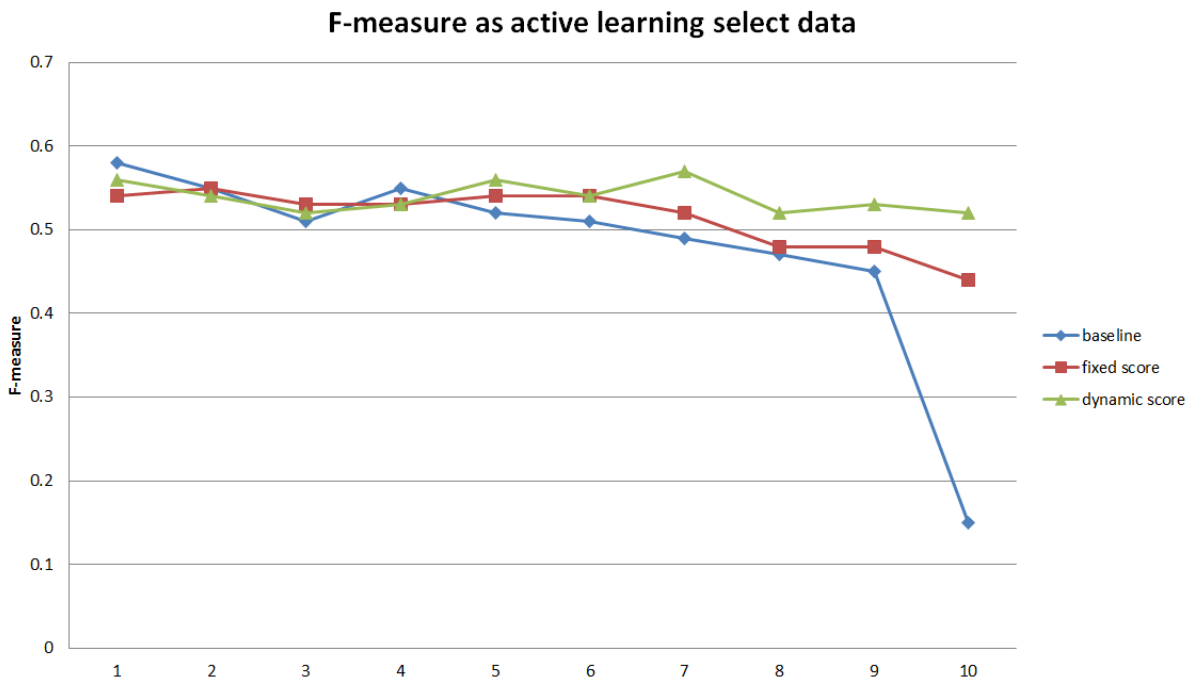

FIgure 2 F-score in each iteration

We compared the framework we used MIP as a baseline applied with score functions such as maximum curiosity and dynamic weight function. To make the score functions comparable dynamic functions were initiated as the fourth power of ratio of negative instances in the training set. These results are visible in Figure 1 and Figure 2 shows.

The figure 1 shows that the current framework works better than other algorithms and searches for positive instances quickly and in much highest rate of accuracy. While it can also be observed that the dynamic weight function extracts the positive instance much slower than the previous algorithms but still much faster than baseline.

In figure 2, we compare the F-measure in each iteration among three method illustrate above. Although MIP(fixed score) can extract positive instance quickly but there is a huge decrease in the overall F-score in each iteration. While the dynamic weight function generates relatively same F score and does not decrease precision as well as recall.

V. CONCLUSION & FUTURE WORK

In this work, we propose an active learning strategy that can dynamically balance the number of positive and negative instances. The active learner actively searches for most informative features to accelerate searching and prediction process. This work is still preliminary phases and requires further improvements to provide significant results.

First and foremost, the algorithm could be improved by providing more robust statistically significant results. On the bright side, the proposed algorithm shows advantage over other methods which is consistent with intuitive thinking of active learning. But it requires further mathematical and experimental proof.

Secondly, there are many other functions that can be implemented other than maximum curiosity, SMO or information gain depending on more granular information on the dataset. Thus other classifiers or feature selection strategy may lead to a better result. These we consider as a priority for the future work.

References

1. Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. (1991) *Science* 253 **,** 49-53.

2. Blagosklonny, M. V. (2002), P53: An ubiquitous target of anticancer drugs. Int. J. Cancer, 98: 161–166.

3. Danziger SA, Baronio R, Ho L, Hall L, Salmon K, et al. (2009) Predicting Positive p53 Cancer Rescue Regions Using Most Informative Positive (MIP) Active Learning. PLoS Comput Biol 5(9): e1000498.

4. Danziger, S.A., Baronio, R., Ho, L., Hall, L., Salmon, K., Hatfield, G.W., Kaiser, P., and Lathrop, R.H. (2009) Predicting Positive p53 Cancer Rescue Regions Using Most Informative Positive (MIP) Active Learning, PLOS Computational Biology, 5(9), e1000498

5. Danziger, S.A., Zeng, J., Wang, Y., Brachmann, R.K. and Lathrop, R.H. (2007) Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants, Bioinformatics, 23(13), 104-114.

6. Danziger, S.A., Swamidass, S.J., Zeng, J., Dearth, L.R., Lu, Q., Chen, J.H., Cheng, J., Hoang, V.P., Saigo, H., Luo, R., Baldi, P., Brachmann, R.K. and Lathrop, R.H. (2006) Functional census of mutation sequence spaces: the example of p53 cancer rescue mutants, IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM, 3, 114-125.

7. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.