BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i104–i114
doi:10.1093/bioinformatics/btm166

Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants

Samuel A. Danziger[1,6], Jue Zeng[2], Ying Wang[3], Rainer K. Brachmann[2,4,6,*] and Richard H. Lathrop[1,5,6,*]

[1]Department of Biomedical Engineering, [2]Department of Medicine, [3]Department of Molecular Biology & Biochemistry, [4]Departments of Biological Chemistry, and Pathology & Laboratory Medicine, [5]Department of Computer Science and [6]Institute for Genomics and Bioinformatics, University of California, Irvine, California, 92697, USA

ABSTRACT

Motivation: Many biomedical projects would benefit from reducing the time and expense of *in vitro* experimentation by using computer models for *in silico* predictions. These models may help determine which expensive biological data are most useful to acquire next. Active Learning techniques for choosing the most informative data enable biologists and computer scientists to optimize experimental data choices for rapid discovery of biological function. To explore design choices that affect this desirable behavior, five novel and five existing Active Learning techniques, together with three control methods, were tested on 57 previously unknown p53 cancer rescue mutants for their ability to build classifiers that predict protein function. The best of these techniques, Maximum Curiosity, improved the baseline accuracy of 56–77%. This article shows that Active Learning is a useful tool for biomedical research, and provides a case study of interest to others facing similar discovery challenges.

1 INTRODUCTION AND BACKGROUND

Ideally, an accurate classifier would be built in the shortest time possible using the least amount of expensive biological data. To achieve this goal strategies are needed that select and assay only the most informative data points first. Active Learning methods iteratively determine the most informative new data points. The p53 cancer rescue mutants present an ideal test case for Active Learning while also engaging in useful cancer research.

Over 6 million people worldwide die of cancer each year (Parkin *et al*., 2002). The central tumor suppressor protein p53 is an important part of cancer prevention mechanisms in healthy human cells. The p53 protein induces cell growth arrest or apoptosis (programmed cell death) in response to cellular stresses (Prives and Hall, 1999; Vogelstein *et al*., 2000). Close to half of all human cancers contain inactivating *p53* mutations. Despite progress, the cure rate of cancers remains around 60% (http://www.cancer.org/). Resistance of human cancers to standard treatments correlates with mutations of *p53* (Seemann *et al*., 2004; Soussi and Beroud, 2001).

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 (Bullock and Fersht, 2001; Hamroun *et al*., 2006; Olivier *et al*., 2002; Sigal and Rotter, 2000).

These full-length p53 cancer mutants provide an exciting opportunity to specifically target cancers. Restoring normal function to mutant p53 would trigger apoptosis in affected cells, thus shrinking or killing the tumor. One strategy is to seek small molecule drugs that stabilize mutant p53 in a native-like conformation (Brachmann, 2004; Bullock and Fersht, 2001; Bykov *et al*., 2003; Wang and Rastinejad, 2003). While this strategy remains to be realized, some p53 cancer mutants are rescued *in vitro* by intragenic second-site cancer suppressor mutations (Baroni *et al*., 2004). In these mutants, a second *p53* mutation restores active wild-type p53 function.

Studies of *p53* second-site suppressor mutations indicate that a large percentage of p53 cancer mutants can be rescued (Baroni *et al*., 2004; Brachmann *et al*., 1998; Danziger *et al*., 2006). In particular, changes of amino acids 235, 239 and 240, alone or combined, result in the rescue of 16 out of 30 of the most common p53 cancer mutants tested (Baroni *et al*., 2004). Thus, intragenic second-site suppressor mutations identify p53 cancer mutants that are likely to be amenable to functional rescue; uncover regions of the p53 core domain that, upon alteration, lead to functional rescue; and, combined with structural and other experimental studies, help to elucidate the basic mechanisms of p53 functional rescue.

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 (Baroni *et al*., 2004; Danziger *et al*., 2006). Which expensive data points should be acquired next in order to rapidly discover biological function?

To this aim, this article explores different Active Learning methods and addresses the following questions:

(1) How well do different Active Learning methods guide the exploration of p53 cancer rescue mutants?

(2) Is one Active Learning method better than others?

*To whom correspondence should be addressed.

1.1 Active Learning

Active Learning refers to iterative machine learning techniques for building a classifier by choosing the most informative examples from a space of unlabeled examples (Cohn *et al*., 2006; Jones *et al*., 2003; Roy and McCallum, 2001; Saar-Tsechansky and Provost, 2001). This strategy constructs a classifier using few examples, and is useful in biological problems where data is expensive. In the case of p53 cancer mutants, Active Learning is intended to select mutants that both quickly improve the classifier and speed the search for previously unknown cancer rescue mutations.

To illustrate the process, we give an example of a Type I Active Learning method. Suppose we have a training set of 204 mutants with known activities. We ask if the laboratory should assay next the activity of mutant R280T + N239Y or of R282W + N239Y. Each mutant, assumed as active and as inactive, is added independently to the training set, resulting in four new sets:

(1) original set plus R280T + N239Y as active;
(2) original set plus R280T + N239Y as inactive;
(3) original set plus R282W + N239Y as active;
(4) original set plus R282W + N239Y as inactive.

A new classifier is built from each new training set and evaluated based on the true positive (*tp*), false positive (*fp*), false negative (*fn*) and true negative (*tn*) receiver operator characteristic (ROC) statistics. For stringency, we use overlap exclusion cross-validation (OECV, Danziger *et al*., 2006), which excludes from the training set all mutants that share more than one mutation with the mutant being tested (i.e. no cancer/rescue mutation pair is shared by tested and training mutants). Finally, the mutant that yielded the best classifier, under either assumed activity, would be chosen.

At each step, typically a small number of mutants are selected, assayed and added to the training set. The new training set begins the next step of *in silico* prediction and *in vitro* experimentation.

The Active Learning methods tested here fall into four categories, called here Type I, Type II, Type III and Type IV. Types I and II are novel methods that extend and adapt the original method of Danziger *et al*. (2006). Types III and IV provide a comparison to existing methods adapted from the machine learning literature.

1.1.1 Type I Type I methods consider each unclassified example, assumed first as active and then as inactive. Type I methods select the example that yielded the most improved cross-validated classifier accuracy under either assumed mutant activity. The Type I methods here are called Maximum Curiosity, Composite Classifier and Improved Composite Classifier (Sections 2.1.1, 2.1.2 and 2.1.3).

A score function, Score(t), evaluates each training set t. The methods differ by choice of the score function. The mutant that gave the highest Score(t) is chosen. In the example, if Score(t_2) were highest, then Type I methods would choose R280T + N239Y.

1.1.2 Type II Type II methods are identical to Type I until the final step. Type II methods choose the mutant with the highest sum of scores across active and inactive. The Type II methods here are Additive Curiosity and Additive Bayesian Surprise (Sections 2.1.4 and 2.1.5). In the example, if $\text{Score}(t_1) + \text{Score}(t_2) < \text{Score}(t_3) + \text{Score}(t_4)$, then Type II methods would choose R282W + N239Y.

1.1.3 Type III Type III methods involve choosing the next unknown example based on how close examples are to the decision boundary. The Type III methods are called Minimum Marginal Hyperplane and Maximum Entropy (Sections 2.1.6 and 2.1.7). These methods are described further in Jing *et al*. (2005), Liu (2004) and Park (2004).

1.1.4 Type IV Type IV methods are similar to Type III methods, but choose examples furthest from the decision boundary rather than closest. Examples furthest from the decision boundary should be those that the classifier is most likely to predict correctly. The Type IV methods here are Maximum Marginal Hyperplane, Minimum Entropy, and Entropic Tradeoff (Sections 2.1.8, 2.1.9 and 2.1.10).

1.2 Motivation for *in silico p53* mutation evaluation

PCR mutagenesis followed by p53 functional assays in yeast (Baroni *et al*., 2004) provides several experimental advantages, such as the ability to combine PCR mutagenesis with repair of a gapped plasmid directly in yeast, and the immediate observation of phenotypes. This experimental strategy rapidly provides an initial training set of positive and negative examples.

For the broader cancer rescue mutant discovery task, however, certain inherent limitations of PCR mutagenesis cannot be overcome. These include the limited number of amino acid changes available per codon, which means that some amino acid changes are essentially inaccessible, and the limited number of coordinated simultaneous mutations, which means that many mutant combinations are unlikely to be seen. Therefore, we began to apply computational strategies to the problem of discovering novel intragenic second-site suppressor mutations, with the long-term goal of a complete functional census of *p53* cancer rescue mutations.

1.3 Relation to previous p53 classifier work

Danziger *et al*. (2006), developed a structure-based p53 classifier that was used to predict a previously unknown set of putative p53 cancer rescue mutants. The principal technical challenge was to extract structure-based features from atomic level molecular models in a way that was useful to feature vector based learning methods. Briefly, 1D, 2D, 3D and 4D features were extracted, filtered and concatenated. 1D features came from the mutation type and location in the p53 core domain. 2D features came from steric and electrostatic properties measured at points on a cartographic projection of the molecular surface. 3D features came from spatial displacements of mutant residues relative to wild-type. 4D features came from the time course of protein unfolding in a simulated heat bath, plus other computational estimates of protein thermostability.

The key finding was that classifiers built from features extracted from atomic level molecular models out-performed classifiers built from features extracted from string-based representations of the same mutant amino acid changes, when compared head-to-head on the same mutant data set. Structure is closer to protein function than is sequence, so that result was satisfying but not unexpected.

Danziger *et al.* (2006) provided a proof in principle that such a structure-based classifier could guide biological discovery, by exhibiting one Active Learning method that out-performed random selection. Several variants and extensions had theoretically desirable properties. Existing machine learning methods were attractive as well. Which should be used, in practice, to guide the upcoming expensive and time-consuming biological experimentation?

2 METHODS

To explain the Active Learning methods used in this article it is helpful to introduce some notation. T is the total set of all p53 mutants under consideration. During each active learning iteration, i, T is broken up into three groups: (1) $T_{K,i}$, p53 mutants with known activities; (2) $T_{U,i}$, p53 mutants with unknown activities and (3) $T_{C,i}$, p53 mutants from $T_{U,i}$ chosen to be assayed.

For the first experiment, the initial training set $T_{K,1}$ contained 204 p53 mutants of known activity, and the initial test set $T_{U,1}$contained 57 putative cancer rescue mutants of unknown activity. Subsequently, after all mutants were predicted and assayed, the total mutant set T was divided into different initial training and test sets to explore how Active Learning methods behaved under different conditions.

2.1 Active Learning implementation

At the beginning of each iteration, three mutants are chosen as $T_{C,i}$ by the Active Learning methods described in this section.

2.1.1 Type I: Maximum Curiosity Maximum Curiosity scores each potential new training set, t, by its cross-validated correlation coefficient (r), calculated as in (1).

$$r_t = \frac{(tp_t \cdot tn_t) - (fp_t \cdot fn_t)}{\sqrt{(tp_t + fp_t)(tp_t + fn_t)(tn_t + fp_t)(tn_t + fn_t)}} \quad (1)$$

The r for each potential new training set is used to determine which unclassified mutants resulted in the largest increase of r.

Maximum Curiosity optimistically assumes that the highest r for each mutant, m, occurs when that mutant is correctly paired with its true activity, as per (2).

$$\text{curiosity}_m = \max\begin{bmatrix} r_{T_{K,i}+m(\text{active})} - r_{T_{K,i}}, \\ r_{T_{K,i}+m(\text{inactive})} - r_{T_{K,i}} \end{bmatrix} \quad (2)$$

2.1.2 Type I: Composite Classifier Composite Classifier scores identically to Maximum Curiosity but the classifier is constructed using the methods discussed in Danziger *et al.* (2006). This was done mainly to provide a direct comparison to our previously published results. Briefly, the Composite Classifier breaks the 1D, 2D, 3D and 4D attributes up into four separate component classifiers. 3000 attributes are selected by Mutual Information (MI, Section 2.3) from the 2D set and the four classifiers are combined into all 15 possible combinations. Each of these combinations is used to build a support vector machine component classifier and given a vote weighted by its cross-validated accuracy in a composite naïve Bayes classifier.

2.1.3 Type I: Improved Composite Classifier The Improved Composite Classifier differs from the original Composite Classifier in how it selects attributes from the 9778 in the 2D component classifier. In this method, parts of the surface not in a promiscuous binding domain (defined in Friedler *et al.*, 2005) are compressed resulting in 4826 attributes. 400 features are then selected from those 4826 attributes using MI.

2.1.4 Type II: Additive Curiosity Additive Curiosity scores much like Maximum Curiosity except in the final step curiosity is calculated by adding the scores for each training set (3). In this way, the mutant chosen may be the most beneficial to the classifier regardless of its revealed activity.

$$\begin{aligned}\text{curiosity}_m = &\left(r_{T_{K,i}+m(\text{active})} - r_{T_{K,i}}\right) \\ &+ \left(r_{T_{K,i}+m(\text{inactive})} - r_{T_{K,i}}\right)\end{aligned} \quad (3)$$

2.1.5 Type II: Bayesian Surprise Bayesian Surprise (Itti and Baldi, 2006) calculates the scores by summing the Kullbach-Leibler (KL) distance between *a priori* probability and *a posteriori* probability (4,5) across assumed active and inactive mutants.

$$KL = \text{posterior} \times \log\left(\frac{\text{posterior}}{\text{prior}}\right) \quad (4)$$

In this implementation, the prior probability is the cross-validated accuracy of the training set ($T_{K,i}$) and the posterior is that of the training set with the unclassified mutant ($T_{K,i} + m(\text{activity})$).

$$\text{surprise}_m = KL_{T_{K,i}+m(\text{active})} + KL_{T_{K,i}+m(\text{inactive})} \quad (5)$$

2.1.6 Type III: Minimum Marginal Hyperplane 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. In its simplest, linear form, a support vector machine creates a hyperplane and stores it as an associated normal vector (w). New mutants, described by an attribute vector (D), are evaluated by (6).

$$\mu = w \cdot D \quad (6)$$

If $\mu > b$, where b is some threshold, then the new mutant is assigned one class, otherwise it is assigned the other. The margin for a new example is the difference between μ and b. It may be helpful to think of the margin as the distance from the new example to the hyperplane. The Minimum Marginal Hyperplane algorithm assumes that the unclassified mutants closest to the dividing hyperplane will be the most informative to the classifier once the true class is known. See Platt (1998) for more details on support vector machines.

2.1.7. Type III: 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 (Witten and Frank, 2006).

Formally, given the attributes for an unclassified mutant (D) and the model (M) constructed from the training set, H is calculated as shown in (7) (Jing *et al.*, 2005).

$$\begin{aligned}H(D) = &-p(\text{active}|D,M) \times \log(p(\text{active}|D,M)) \\ &- p(\text{inactive}|D,M) \times \log(p(\text{inactive}|D,M))\end{aligned} \quad (7)$$

By choosing mutants with the highest H, Maximum Entropy assumes that the most informative mutants are those that the classifier is most uncertain about.

2.1.8 Type IV: Maximum Marginal Hyperplane Maximum Marginal Hyperplane scores as in Section 2.1.6, but chooses mutants that are furthest from the dividing hyperplane.

2.1.9 Type IV: Minimum Entropy Minimum Entropy scores as in Section 2.1.7, but chooses mutants that have the lowest possible H.

2.1.10 Type IV: Entropic Tradeoff Entropic Tradeoff scores using entropy (H) as in Sections 2.1.7 and 2.1.9, giving highest scores to training sets that include mutants with both high and low H.

If a classifier is both to learn quickly and predict accurately as it proceeds, the mutants chosen should be a mix of highly informative and easily predicted. Here, the three mutants chosen are:

(1) The mutant with the highest H.

(2) The mutant with the lowest H that is predicted inactive.

(3) The mutant with the lowest H that is predicted active.

If the classifier runs out of mutants that are predicted active or inactive, the algorithm chooses one mutant with the highest H and two mutants with the lowest H.

2.2 *In Vitro* experimentation

All p53 mutants were evaluated using a well-established functional yeast assay for wild-type p53. The assay findings have proven to correlate well with subsequent confirmatory studies in mammalian experiments (Baroni *et al.*, 2004; Brachmann *et al.*, 1996, 1998; Kobayashi *et al.*, 2003; Qian *et al.*, 2002). The p53 mutants are stably expressed from a *CEN*-based plasmid (one copy per cell) using the constitutive *ADH1* promoter, thus resulting in very similar protein levels of p53 mutants. The p53-dependent transcriptional activity results in expression of the *URA3* reporter gene due to an upstream consensus p53 DNA binding site. Yeast expressing *URA3* will grow on yeast plates lacking uracil.

All *p53* cancer mutations not tested with N235K, N239Y and N235K + N239Y in our previous studies (Baroni *et al.*, 2004; Danziger *et al.*, 2006) were cloned into yeast expression plasmids containing both the suppressor mutation(s) and a *p53* cancer mutation marked by a unique restriction enzyme site. *p53* cancer mutations upstream of N235K, N239Y or N235K + N239Y were cloned using EcoR V and Xba I. *p53* cancer mutations downstream of the suppressor mutations were cloned using Nsi I and Sac I. Correct cloning for all constructs was confirmed by loss of the previously present unique restriction enzyme site of a *p53* cancer mutation (Baroni *et al.*, 2004).

2.3 Attribute selection

Attributes are generated from model p53 mutants constructed *in silico* from chain b of the wild-type p53 crystal structure (Cho *et al.*, 1994) using Amber molecular modeling software (Case *et al.*, 2004). These attributes include a 1D sequence perspective, a 2D steric and electrostatic surface map perspective, a 3D distance map perspective and a 4D stability perspective (Danziger *et al.*, 2006).

Compressing, normalizing and then concatenating attributes from these perspectives yields 5867 attributes per mutant to describe <300 mutants. To improve speed and generalization, attributes were selected using the Conditional Mutual Information Maximization (CMIM) algorithm. Prior to learning, selecting X attributes and then cross-validating using OECV Danziger *et al.*, 2006), we determined that 550 out of 5867 was the optimal number of attributes.

CMIM is a method for determining which attributes are most informative for classifying an example. It is based around the MI algorithm. MI quantifies the change in information, I, by measuring the entropy (H) of the class A before and after attribute B is known. Conceptually, MI measures how much less random the class A is if B is known. MI selects attributes with high I values.

MI is of limited utility in that highly correlated attributes will have very similar scores. For example, given a 2D surface map, two points immediately next to each other would very likely provide almost exactly the same information and therefore have almost exactly the same I value. For an effective classifier, attributes should be as independent as possible (Witten and Frank, 2006).

CMIM solves this problem by answering the question 'How much information does attribute C provide about class A given that attribute B is already known?' Formally, this is shown in Equations (8) and (9).

$$I(A,B|C) = H(A|C) - H(A|B,C) \tag{8}$$

$$H(A|B,C) = H(A,B,C) - H(B,C) \tag{9}$$

Given the probabilities $P(e_1 \ldots,e_n)$ of events e_1 through e_n occurring simultaneously, the I provided by CMIM may be implemented as shown in Equation (10).

$$\begin{aligned} I(A,B|C) = &\sum_{a}^{A}\sum_{b}^{B}\sum_{c}^{C} P(a,b,c)*\log(P(a,b,c)) \\ &+ \sum_{c}^{C} P(c)*\log(P(c)) \\ &- \sum_{a}^{A}\sum_{c}^{C} P(a,c)*\log(P(a,c)) \\ &- \sum_{b}^{B}\sum_{c}^{C} P(b,c)*\log(P(b,c)) \end{aligned} \tag{10}$$

The first attribute chosen is that with the highest I value as calculated using MI. All following attributes are scored by calculating the CMIM I value with respect to all attributes already chosen.

More formally, let X_n be the vector of values for all mutants at attribute n, and let v be the sorted score vector for each of F attributes. Start by initializing v using MI (11) and iteratively update the scores as attributes are chosen (12).

$$v(1) = \arg\max_{n}(I(A,X_n)) \tag{11}$$

$$\forall f, 1 \le f \le F,$$

$$v(f+1) = \arg\max_{n}\left(\min_{l<f}\left(I\left(A,X_n|X_{v(l)}\right)\right)\right) \tag{12}$$

The CMIM algorithm was implemented using optimizations described in Fleuret (2004), which gives fast, efficient execution.

2.4 Control

To test if Active Learning methods are indeed useful, three different control methods were used.

The first control, Non-iterated Prediction, tested if Active Learning is more effective than no Active Learning at all. It predicted all unclassified mutants ($T_{U,1}$) using the initial training set ($T_{K,1}$).

The second control, Predict All Inactive, demonstrated the naturally skewed data set by predicting $T_{U,1}$ as the most common class, inactive. Because mutants that rescue p53 functionality are rare, the classifier could be highly accurate but utterly useless by always predicting mutants to be inactive. In examining the results from this control, it is informative to consider the correlation coefficient as well as the accuracy.

The final control, Random (30 Trials) selects three mutants at random from $T_{U,1}$ during each iteration for 30 trials. This tests if Active Learning methods are truly more effective than selecting the 'most informative' examples at random.

Table 1. p53 cancer rescue mutants predicted using Maximum Curiosity

Cancer mutation	N235K	N239Y	N235K + N239Y
C135Y	**I***	**I***	**I***
C141Y	*(A)*	**I***	**A**
P151S	**I***	**I***	**I***
V157F	*(A)*	*(A)*	**A**
R158L	*(A)*	*(A)*	**A**
V173L	*(I)*	*(A)*	**A**
R175H	**I***	**I***	**I**
C176F	**I***	**I***	**I**
H179R	**I***	**I***	**I**
H179Y	**I***	**I***	**I***
Y205C	*(A)*	*(I)*	**A***
Y220C	*(A)*	*(A)*	**A***
G245C	*(I)*	*(A)*	**A***
G245D	**I**	**I***	**I**
G245S	*(I)*	*(A)*	**A***
R248Q	**I***	**I***	**I***
R248L	**I***	**I***	**I***
R248W	**I***	**I***	**I***
R249M	*(I)*	*(A)*	**A***
R249S	**I**	**I***	**I**
V272M	*(A)*	*(A)*	**A***
R273C	*(I)*	*(I)*	**I***
R273H	*(I)*	*(I)*	**A***
R273L	*(I)*	*(I)*	**I***
P278L	**I**	**I**	**I**
R280T	**I***	**I***	**I***
R282W	**I**	**I***	**I***
E285K	*(I)*	*NA*	*NA*
E286K	*(A)*	*(I)*	**I***

'A' indicates active and 'I' inactive as determined by the yeast assays. *Italicized* yeast phenotypes in parentheses are for p53 mutants that were part of the training set. *NA* means not assayed. Mutants annotated with an * were predicted correctly by the Maximum Curiosity Active Learning method.

3 RESULTS

This section shows how well each of the Active Learning techniques predicted a set of 57 unclassified putative p53 cancer rescue mutants using a training set of 204 mutant. The results for the Active Learning method that performed best, Maximum Curiosity, are presented in Table 1; the *in vitro* assay results are presented in Figure 1; and the summary prediction statistics for all Active Learning methods are presented in Table 2.

The bold and italicized Accuracy and Correlation Coefficient highlight the best and second best scores on $T_{C,i}$, respectively

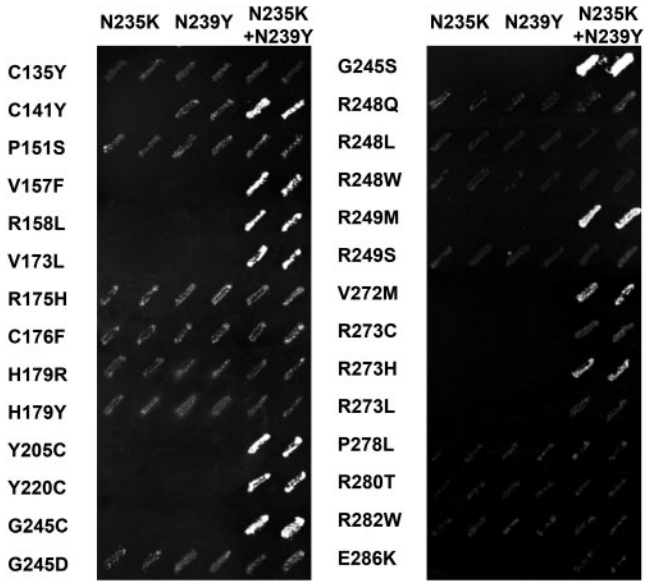

Fig. 1. Yeast growth results for p53 mutants tested in this experiment. p53-dependent transcriptional activity results in *URA3* expression and growth at 37°C on yeast plates lacking uracil.

Table 2. Classifier accuracy predicting three mutants at a time

Type	Method	204 Predicts 57 Accuracy	Correlation coefficient	Student-*T*
I	Maximum Curiosity	77.19% +/− 5.61%	**0.5255**	0.00%
I	Composite Classifier	70.18% +/− 6.11%	0.4447	100.0%
I	Improved Composite Classifier	71.93% +/− 6.00%	0.4637	100.0%
II	Additive Curiosity	73.68% +/− 5.88%	0.3857	99.81%
II	Additive Bayesian Surprise	73.68% +/− 5.88%	0.4342	99.81%
III	Minimum Marginal Hyperplane	64.91% +/− 6.38%	0.2845	100.0%
III	Maximum Entropy	64.91% +/− 6.38%	0.2845	100.0%
IV	Maximum Marginal Hyperplane*	***78.95% +/− 5.45%***	0.3699	90.42%
IV	Minimum Entropy*	77.19% +/− 5.61%	0.3406	0.00%
IV	Entropic Tradeoff*	**80.70% +/− 5.27%**	***0.4860***	99.89%
C	Non-iterated Prediction	56.14% +/− 6.63%	0.2530	100.0%
C	Predict All Inactive	80.70% +/− 5.27%	0.0000	99.89%
C	Random (30 trials)	74.39% +/− 3.87%	0.3550 +/− 0.0992	99.24% +/− 2.89%

The **bold** and *italicized* Accuracy and Correlation Coefficient highlight the best and second best scores on Tc, respectively (excluding Predict All Inactive). All accuracies were calculated by treating each prediction as a separate test for accuracy, and all accuracies show the standard error, except for Random (30 trials) which shows the standard deviation. The Student-T percentage was calculated from the mean and standard deviation to determine statistical difference from Maximum Curiosity. Type C means a control type of active learning. Methods marked * were developed after the double-blind predictions.

(excluding Predict All Inactive). All accuracies were calculated by treating each prediction as a separate test for accuracy, and all accuracies show the standard error, except for Random (30 trials) which shows the standard deviation. The Student-T percentage was calculated from the mean and SD to determine statistical difference from Maximum Curiosity. Type 'C' means a control type of active learning. Methods marked * were developed after the double-blind predictions.

4 ANALYSIS

When exploring a mutation sequence space of medical importance, such as the cancer rescue mutants in this article, there are two things that are desired greatly by the biological researcher: (1) to improve the classifier as quickly as possible (Section 4.1); and (2) to identify as many novel functionally active mutants as possible (Section 4.2). For the most effective discovery, the Active Learning method should perform well in both criteria (Section 4.3).

The Type 1 methods Composite Classifier and Improved Composite Classifier from Danziger *et al*. (2006) were omitted from analysis in this section as they are computationally very expensive and theoretically very similar to Maximum Curiosity.

4.1 The quickest

For most Active Learning, the goal is to train a classifier as quickly as possible by using the smallest number of examples to reach maximum accuracy. Here we define the forward prediction accuracy as the classifier accuracy predicting all remaining unclassified mutants ($T_{U,i}$) during each training iterations. To evaluate the forward prediction accuracy of each Active Learning method while starting with different size training sets, two additional partitions of the 261 mutants in T were constructed. One partition used the 123 mutants known in Danziger *et al*. (2006) as the training set $T_{K,1}$ to predict the other 138 mutants as $T_{U,1}$. The other partition assigned all 25 single amino acid mutants as $T_{K,1}$ to predict the other 236 mutants as $T_{U,1}$. The results of these trials are presented in Table 3.

To better illustrate Active Learning behavior, the forward prediction accuracy for each Active Learning method using 25 mutants as $T_{K,1}$ is plotted in Figure 2. From this figure, Type III methods Minimum Marginal Hyperplane and Maximum Entropy achieve the highest accuracy the most rapidly. This is not surprising as Type III Active Learning methods use the decision boundary to choose the hardest mutants to predict during each iteration. Therefore, as expected, Type IV methods Maximum Marginal Hyperplane and Minimum Entropy failed to achieve high accuracy rapidly. Interestingly, the Type I method, Maximum Curiosity, learns nearly as quickly as the Type II methods.

4.2 The most accurate and the positive predictive value

The long term goals of this p53 cancer rescue mutant study involve iteratively and correctly identifying new active p53

Table 3. Active Learning across varying data sets

Type	Method	25 Predicts 236		123 Predicts 138		204 Predicts 57	
		Accuracy	Area under the curve	Accuracy	Area under the curve	Accuracy	Area under the curve
I	Maximum Curiosity	***0.7274 +/− 0.0046***	0.7700 +/− 0.0094	***0.7410 +/− 0.0077***	0.7891 +/− 0.0140	0.7246 +/− 0.0187	0.7722 +/− 0.0254
II	Additive Curiosity	0.7018 +/− 0.0047	0.7364 +/− 0.0087	0.6835 +/− 0.0082	0.7387 +/− 0.0158	0.7316 +/− 0.0186	0.7900 +/− 0.0291
II	Additive Bayesian Surprise	0.6674 +/− 0.0049	0.7068 +/− 0.0090	0.7095 +/− 0.0080	0.7637 +/− 0.0160	***0.7456 +/− 0.0182***	0.8054 +/− 0.0291
III	Minimum Marginal Hyperplane	0.7250 +/− 0.0046	***0.7780 +/− 0.0124***	**0.7475 +/− 0.0076**	**0.8246 +/− 0.0207**	0.7193 +/− 0.0188	*0.8141 +/− 0.0425*
III	Maximum Entropy	**0.7507 +/− 0.0045**	**0.8118 +/− 0.0122**	0.7348 +/− 0.0078	***0.8089 +/− 0.0199***	**0.7544 +/− 0.0180**	**0.8354 +/− 0.0359**
IV	Maximum Marginal Hyperplane	0.6440 +/− 0.0049	0.6621 +/− 0.0049	0.6432 +/− 0.0084	0.6803 +/− 0.0166	0.6122 +/− 0.0204	0.6192 +/− 0.0178
IV	Minimum Entropy	0.6156 +/− 0.0050	0.5959 +/− 0.0060	0.6392 +/− 0.0084	0.6530 +/− 0.0130	0.6158 +/− 0.0204	0.5902 +/− 0.0222
IV	Entropic Tradeoff	0.6965 +/− 0.0047	0.7139 +/− 0.0058	0.7058 +/− 0.0080	0.7423 +/− 0.0122	***0.7456 +/− 0.0182***	0.8044 +/− 0.0272
C	Random (30 trials)	0.6700 +/− 0.0141	0.6922 +/− 0.0237	0.6931 +/− 0.0231	0.7326 +/− 0.0231	0.6950 +/− 0.0331	0.7392 +/− 0.0372

'25 Predicts 236' uses the 25 mutants with single point mutations as the initial training set and iteratively predicts the remaining 236 mutants in sets of three. Similarly, '123 Predicts 138' uses the 123 mutants known in Danziger *et al*. (2006) to predict the remaining 138 mutants. '204 Predicts 57' is the data set discussed in Section 3. Accuracy is the forward prediction accuracy discussed in Section 4.1 weighted by how many mutants are predicted in each iterations. Area under the curve is the average forward prediction accuracy for all iterations. The bolded and italicized Accuracy and Area under the curve highlight the best and second best scoring classifiers (respectively) for each data set. All errors show the standard error, except for Random (30 trials) which shows the standard deviation. Type 'C' means a control type of Active Learning.

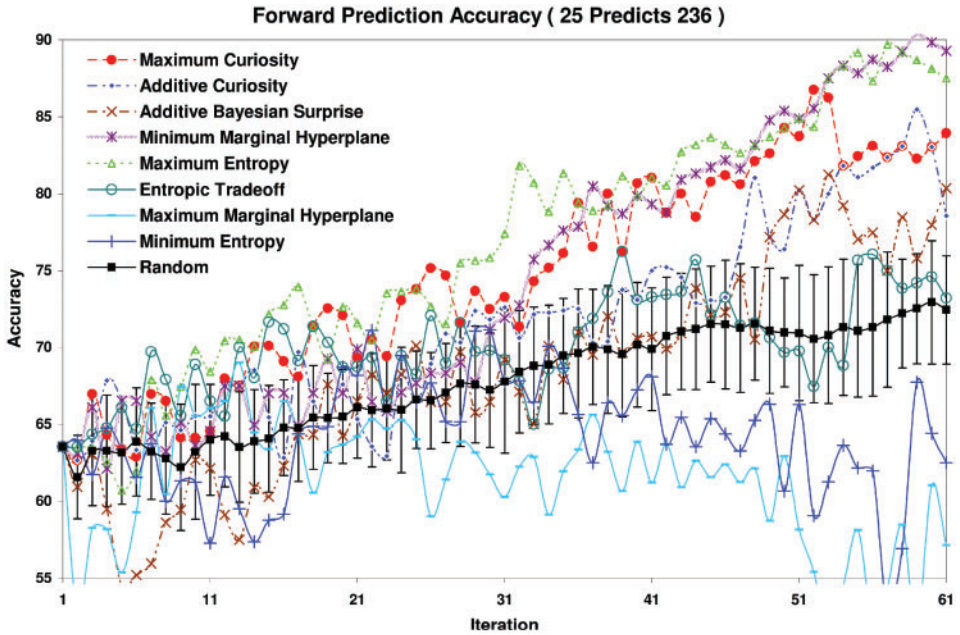

Fig. 2. Accuracy predicting the remaining unclassified mutants at each iteration using an initial set of 25 known mutants. A scree test on the one SD error bars from Random (30 trials) was used to truncate the graph at 61 of 79 iterations.

Table 4. 3-pt. classifier accuracy

Type	Method	25 Predicts 236	123 Predicts 138	204 Predicts 57	Average
I	Maximum Curiosity	0.6525	***0.7391***	0.7719	0.7211
II	Additive Curiosity	0.6313	0.7101	0.7368	0.6927
II	Additive Bayesian Surprise	0.6398	0.7246	0.7368	0.7004
III	Minimum Marginal Hyperplane	0.6483	0.6522	0.6491	0.6498
III	Maximum Entropy	0.6652	0.6667	0.6491	0.6603
IV	Maximum Marginal Hyperplane	0.6992	0.6956	**0.8070**	0.7339
IV	Minimum Entropy	***0.70339***	**0.7681**	***0.7895***	**0.7536**
IV	Entropic Tradeoff	**0.7119**	**0.7681**	0.7719	***0.7506***
C	Random (30 trials)	0.6900	0.7374	0.7439	0.7237

The classifier accuracies predicting the three mutants chosen during each iteration. The **bold** and *italicized* scores highlight the highest scoring and second highest scoring Active Learning methods in each column, respectively.

mutants to be verified using *in vitro* assays. Active Learning, here choosing three unclassified mutants at a time, is essentially a scaled down version of this larger project. Therefore, the 3-pt. Classifier Accuracy, the accuracy predicting the three mutants chosen at the beginning of each Active Learning iteration, $T_{C,i}$ (Table 4), is an indicator of how this study will progress toward accurate classifiers.

The classifier accuracies predicting the three mutants chosen during each iteration. The bold and italicized scores highlight the highest scoring and second highest scoring Active Learning methods in each column, respectively.

However, for a classifier to find new cancer rescue mutants effectively, functionally active mutants must be predicted as accurately as possible. That is to say, true positives (tp) are more important than true negatives (tn) for the classifier to be useful. Therefore, a good way to evaluate a classifier is to use the Positive Predictive Value (PPV), shown in (13), as well as accuracy.

$$PPV = \frac{tp}{(tp + fp)} \tag{13}$$

Table 5 shows the 3-pt. PPV (Section 4.2) from predicting the three mutants selected at the beginning of each iteration.

The PPV calculated on the three mutants chosen during each iteration. The bold and italicized scores highlight the highest scoring and second highest scoring Active Learning methods in each column, respectively.

To a certain extent, the values shown in Table 4 and Table 5 represent an unfair comparison. Since each Active Learning method selects mutants in a different order, all of the mutants

Table 5. 3-pt. PPV

Type	Method	25 Predicts 236	123 Predicts 138	204 Predicts 57	Average
I	Maximum Curiosity	0.4875	***0.4687***	***0.4545***	0.4702
II	Additive Curiosity	0.4533	0.4250	0.4000	0.4261
II	Additive Bayesian Surprise	0.4658	0.4500	0.4090	0.4416
III	Minimum Marginal Hyperplane	0.4789	0.3158	0.3200	0.3716
III	Maximum Entropy	0.5116	0.3333	0.3200	0.3883
IV	Maximum Marginal Hyperplane	0.5672	0.3824	**0.5000**	0.4832
IV	Minimum Entropy	***0.5676***	**0.5333**	0.4167	***0.5059***
IV	Entropic Tradeoff	**0.5857**	**0.5333**	0.4286	**0.5159**
C	Random (30 trials)	0.5377	0.4682	0.3993	0.4684

The PPV calculated on the three mutants chosen during each iteration. The **bold** and *italicized* scores highlight the highest scoring and second highest scoring Active Learning methods in each column, respectively.

Table 6. Unclassified PPV

Type	Method	25 Predicts 236	123 Predicts 138	204 Predicts 57	Average
I	Maximum Curiosity	0.4595	0.3918	**0.3556**	***0.4023***
II	Additive Curiosity	***0.4946***	0.3669	0.3261	0.3959
II	Additive Bayesian surprise	**0.4962**	0.3745	0.3261	0.3989
III	Minimum Marginal Hyperplane	0.4450	***0.3964***	0.3333	0.3916
III	Maximum Entropy	0.4895	**0.4120**	0.3298	**0.4104**
IV	Maximum Marginal Hyperplane	0.4854	0.3148	***0.3529***	0.3844
IV	Minimum Entropy	0.4816	0.3667	0.3253	0.3912
IV	Entropic Tradeoff	0.4258	0.3599	0.3086	0.3648

The PPV was determined for all unclassified mutants at each iteration that did not appear in any training set at that iteration. The bold and italicized scores highlight the highest scoring and second highest scoring Active Learning methods, respectively. Random (30 trials) was omitted from this scoring method because 30 additional classifiers would prematurely eliminate too many mutants from consideration.

except the first three, $T_{C,1}$, are predicted using different training sets. To partially correct for this, Table 6 shows the unclassified PPV, calculated by predicting all unclassified mutants, $T_{U,i}$ that do not appear in the training set for any method at iteration i.

The PPV was determined for all unclassified mutants at each iteration that did not appear in any training set at that iteration. The bold and italicized scores highlight the highest scoring and second highest scoring Active Learning methods, respectively. Random (30 trials) was omitted from this scoring method because 30 additional classifiers would prematurely eliminate too many mutants from consideration.

Across all methods for determining the most useful Active Learning method, the three Type IV Active Learning methods and the Type I, Maximum Curiosity, tend to do best.

4.3 Overall best methods

If some Active Learning methods perform better at learning quickly, but other methods perform well at finding active mutants, which method is best? In this context, there is no clear theoretical framework for quantifying speed versus accuracy. However, there are (at least) three reasonable metrics that combine the accuracy presented in Section 4.1 ($acc1$) with any of the measures presented in Section 4.2 ($acc2$).

(1) **Distance From An Ideal Classifier:** Assume an ideal classifier with the maximum possible accuracies, i.e. $acc1 = acc2 = 1$. Then Euclidean distance from this ideal classifier is described in (14) and shown in Figure 3.

$$\text{Distance} = \sqrt{(1 - acc1)^2 + (1 - acc2)^2} \tag{14}$$

(2) **Maximum Area:** If $acc1$ and $acc2$ are assumed to be orthogonal measures, like the width and height of a rectangle, then a good measure would be the resulting area as described in (15) and shown in Figure 4.

$$\text{Area} = acc1 \times acc2 \tag{15}$$

(3) **Average Accuracy:** If $acc1$ and $acc2$ are assumed to be interchangeable in terms of usefulness, then the best classifier could be found by adding them as shown in (16) and shown in Figure 5.

$$\text{Average} = \frac{(acc1 + acc2)}{2} \tag{16}$$

To merge Figures 3–5, each Active Learning method was ordered and assigned points based on how well it performed for each metric relative to the other methods. For example, if there were eight methods, the highest rated method scored

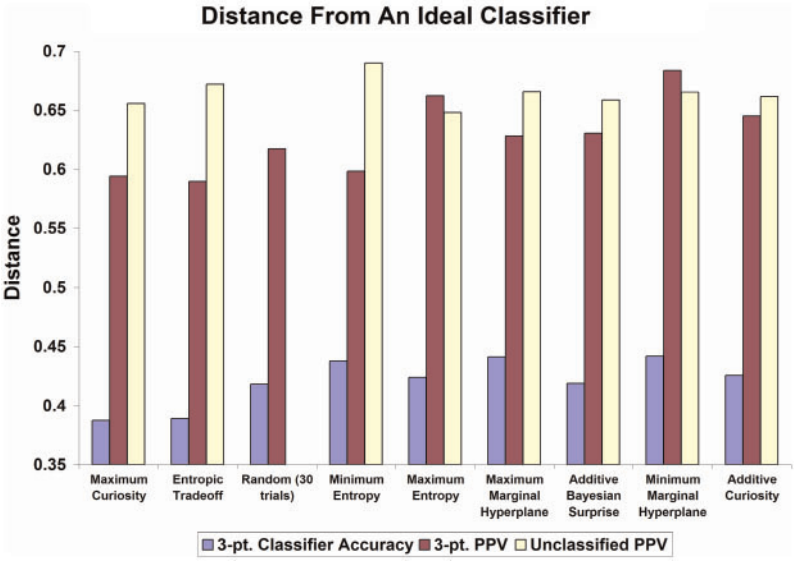

Fig. 3. The Active Learning methods, ordered left to right: Type I: Maximum Curiosity, Type IV: Entropic Tradeoff, Type C: Random (30 trials), Type IV: Minimum Entropy, Type III: Maximum Entropy, Type IV: Maximum Marginal Hyperplane, Type II: Additive Bayesian Surprise, Type III: Minimum Marginal Hyperplane, and Type II: Additive Curiosity, evaluated using Average Accuracy. Shorter distances are better.

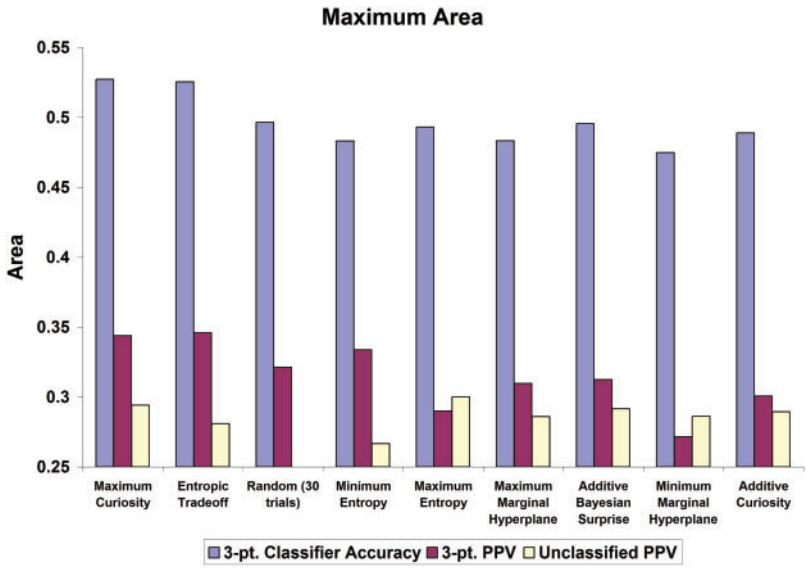

Fig. 4. The Active Learning methods as described in Figure 3 evaluated using the Maximum Area metric. Larger areas are better.

7 points, the second highest 6 and so on, until the lowest rated method received 0 points. Table 7 ranks each Active Learning method based on the average number of points per category, revealing the top 3 Active Learning Methods to be Maximum Curiosity, Entropic Tradeoff, and Random (30 trials).

4.3.1 Maximum Curiosity Maximum Curiosity was the overall best ranked Active Learning method. It performed well, usually at least third best according to any scoring metrics. However it is computationally slow, requiring a separate cross-validation for each unclassified mutant.

The average score for each Active Learning Method, calculated based on the rank of each classifier scored in Figures 3–5.

4.3.2 Entropic Tradeoff As per Figures 4 and 5, Type IV methods, including Entropic Tradeoff, accurately predict the three mutants chosen each iteration while performing no worse than random at learning quickly. Entropic Tradeoff's primary advantages are computational speed and that it can be tuned by adjusting the ratio of high entropy to low entropy mutants chosen.

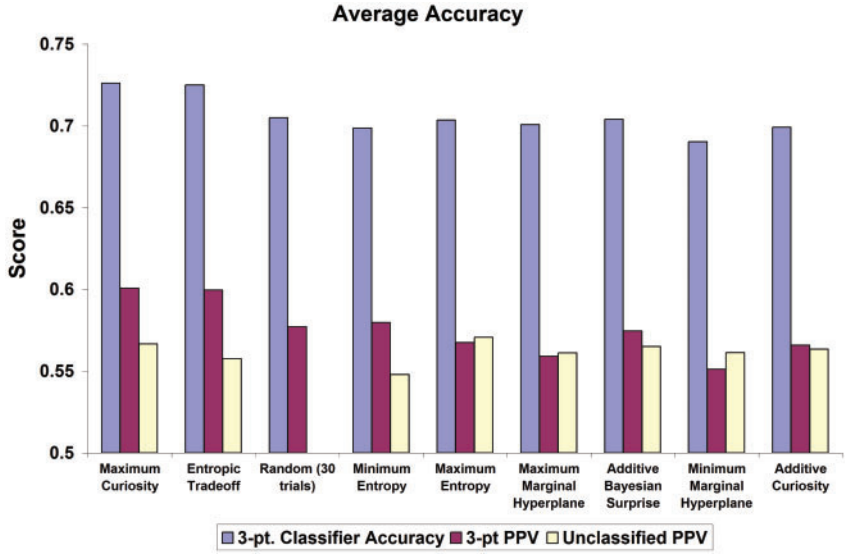

Fig. 5. The Active Learning methods as described in Figure 3 evaluated using the Average Accuracy metric. Larger scores are better.

Table 7. Overall average rank

Rank	Method	Average score
1	Maximum Curiosity	**6.11**
2	Entropic Tradeoff	**5.56**
3	Random (30 trials)	**5.50**
4	Minimum Entropy	**4.44**
5	Maximum Marginal Hyperplane	**3.22**
6	Maximum Entropy	**3.22**
7	Additive Bayesian Surprise	**2.89**
8	Minimum Marginal Hyperplane	**2.33**
9	Additive Curiosity	**1.89**

The average score for each Active Learning Method, calculated based on the rank of each classifier scored in Figs 3, 4 and 5.

4.3.3 Random (30 Trials) Perhaps the most surprising result is that picking mutants at random did so well. We hypothesize that other methods tend to get stuck in local minima, wasting iterations disproving false hypotheses. Previous research has shown that Random Active Learning is only initially more accurate (Tong *et al.*, 2001). Therefore, 261 mutants may be too few data points to reveal this disadvantage.

5 CONCLUSION

This article demonstrated the use of computer models and Active Learning to guide the exploration of p53 cancer rescue mutants. Follow-up studies will proceed by identifying interesting clusters of putative p53 cancer rescue mutants. The p53 classifier will iteratively identify interesting mutants which biologists will synthesize and test them, until cancer rescue mutants have been explored for the top 100 p53 mutants found in human cancers.

It is expected that random Active Learning will become much less useful as the experiments progress into larger mutant spaces. Therefore, further experiments will use Maximum Curiosity and Entropic Tradeoff depending on the computational load to process mutants in the pool currently under consideration.

Ultimately, this research will help others studying mutant protein function using crystal structures. The techniques described here will help reveal mutants with a desired function from a larger pool of candidates. This would be useful for any similar experiment program exploring a large sequence space of expensive biological data.

ACKNOWLEDGEMENTS

Thanks to Pierre Baldi, Richard Chamberlin, Jonathan Chen, Jianlin Cheng, Melanie Cocco, Richard Colman, John Coroneus, Lawrence Dearth, Vinh Hoang, Qiang Lu, Hartmut Luecke, Ray Luo, Gabe Moothart, Hiroto Saigo, Don Senear and Josh Swamidass. Funding provided by the NIH, NSF, UCI Medical Scientist Training Program, UCI Office of Research and Graduate Studies and UCI Institute for Genomics and Bioinformatics.

Conflict of Interest: none declared.

REFERENCES

Baroni,T.E. *et al.* (2004) A global suppressor motif for p53 cancer mutants. *Proc. Natl. Acad. Sci. USA*, **101**, 4930–4935.

Blagosklonny,M.V. (2000) p53 from complexity to simplicity: mutant p53 stabilization, gain-of-function, and dominant-negative effect. *Faseb. J.*, **14**, 1901–1907.

Brachmann,R.K. (2004) p53 mutants: the achilles' heel of human cancers? *Cell Cycle*, **3**, 1030–1034.

Brachmann,R.K. *et al.* (1996) Dominant-negative p53 mutations selected in yeast hit cancer hot spots. *Proc. Natl Acad. Sci. USA*, **93**, 4091–4095.

Brachmann,R.K. *et al.* (1998) Genetic selection of intragenic suppressor mutations that reverse the effect of common p53 cancer mutations. *EMBO J.*, **17**, 1847–1859.

Bullock,A.N. and Fersht,A.R. (2001) Rescuing the function of mutant p53. *Nat. Rev. Cancer*, **1**, 68–76.

Bykov,V.J. *et al*. (2003) Small molecules that reactivate mutant p53. *Eur. J. Cancer*, **39**, 1828–1834.
Case,D.A. *et al*. (2004) ''AMBER 8''. University of California, San Francisco.
Cho,Y. *et al*. (1994) Crystal structure of a p53 tumor suppres-sor-DNA complex: understanding tumorigenic mutations. *Science*, **265**, 346.
Cohn,D.A. *et al*. (1996) Active learning with statistical models. *J. Artif. Intell. Res.*, **4**, 129–14.
Danziger,S.A. *et al*. (2006) Functional census of mutation sequence spaces: the example of p53 cancer rescue mutants. *IEEE T. Comput. Biol. Bioinform*, **3**, 114–125.
Erster,S. and Moll,U.M. (2005) Stress-induced p53 runs a transcription-independent death program. *Biochem. Biophys. Res. Commun.*, **331**, 843–850.
Fleuret,F. (2004) Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.*, **5**, 1521–1555.
Friedler,D.B. *et al*. (2004) Binding of RAD51 and other peptide sequences to a promiscuous, highly electrostatic, binding site in p53. *J. Biol. Chem.*, **280**, 8051–8059.
Greenblatt,M.S. *et al*. (1994) Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res.*, **54**, 4855–4878.
Hamroun,D. *et al*. (2006) The UMD TP53 database and website: update and revisions. *Hum. Mutat.*, **27**, 14–20.
Itti,L. and Baldi,P. (2006) Bayesian surprise attracts human attention. *Adv. Neural Inform Process. Systems*, **18**.
Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
Ho,J. and Benchimol,S. (2003) Transcriptional repression mediated by the p53 tumour suppressor. *Cell Death Differ.*, **10**, 404–408.
Hollstein,M. *et al*. (1991) p53 mutations in human cancers. *Science*, **253**, 49–53.
Jing,F. *et al*. (2005) A unified framework for image retrieval using keyword and visual features. *IEEE T. Image Process.*, **14**, 979–989.
Jones,R. *et al*. (2003) Active learning for information extraction with multiple view feature sets. *ECML-03 Workshop on Adaptive Text Extraction and Mining*.
Ko,L.J. and Prives,C. (1996) p53: puzzle and paradigm. *Genes Dev.*, **10**, 1054–1072.
Kobayashi,T. *et al*. (2003) Genetic strategies in *Saccharomyces cerevisiae* to study human tumor suppressor genes. *Methods Mol. Biol.*, **223**, 73–86.
Liu,Y. (2004) Active learning with support vector machine applied to gene expression data for cancer classification. *J. Chem. Inform. Comput. Sci.*, **44**, 1936–1941.
Muthurajan,U.M. *et al*. (2004) Crystal structures of histone Sin mutant nucleosomes reveal altered protein-DNA interactions. *EMBO J.*, **23**, 260–271.
Olivier,M. *et al*. (2002) The IARC TP53 database: new online mutation analysis and recommendations to users. *Hum. Mutat.*, **19**, 607–614.
Park,J. (2004) Convergence and application of online active sampling using orthogonal pillar vectors. *IEEE T. Pattern Anal. Mach. Learn.*, **28**, 1197–1207.
Parkin,D. *et al*. (2005) Global cancer statistics, 2002. *CA Cancer J Clin.*, **55**, 74–108.
Platt,J.C. (1998) Sequential minimum optimization: a fast algorithm for training support vector machines. *Microsoft Research Technical Report MSR-TR-98-14*.
Prives,C. and Hall,P.A. (1999) The p53 pathway. *J. Pathol.*, **187**, 112–126.
Provost,M.S.-T.a.F.J. (2001) Active learning for class probability estimation and ranking. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, Seattle, Washington, pp. 911–920.
Qian,H. *et al*. (2002) Groups of p53 target genes involved in specific p53 downstream effects cluster into different classes of DNA binding sites. *Oncogene*, **21**, 7901–7911.
Roy,N. and McCallum,A. (2001) Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of 18th International Conference on Machine Learning*, pp. 441–448.
Seemann,S. *et al*. (2004) The tumor suppressor gene TP53: implications for cancer management and therapy. *Crit. Rev. Clin. Lab. Sci.*, **41**, 551–583.
Sigal,A. and Rotter,V. (2000) Oncogenic mutations of the p53 tumor suppressor: the demons of the guardian of the genome. *Cancer Res.*, **60**, 6788–6793.
Soussi,T. and Beroud,C. (2001) Assessing TP53 status in human tumours to evaluate clinical outcome. *Nat. Rev. Cancer*, **1**, 233–240.
Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med.*, **10**, 789–799.
Vogelstein,B. *et al*. (2000) Surfing the p53 network. *Nature*, **408**, 307–310.
Wang,W. *et al*. (2003) Restoring p53-dependent tumor suppression. *Cancer Biol. Ther.*, **2**, S55–S63.
Wei,C.-L. *et al*. (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell*, **124**, 207–219.
Witten, I.and Frank,E. (2005) Data mining: practical machine learning tools and techniques. 2nd edn. Morgan Kaufmann Series in Data Management Systems, San Franciso, CA 94111.