HHS Public Access

Author manuscript

Proc 2014 6th Int Adv Res Workshop In Silico Oncol Cancer Investig (2014). Author manuscript; available in PMC 2021 October 06.

Published in final edited form as:

Proc 2014 6th Int Adv Res Workshop In Silico Oncol Cancer Investig (2014). 2014 November ; 2014: .
doi:10.1109/iarwisoci.2014.7034632.

Machine Learning Predictions of Cancer Driver Mutations

E. Joseph Jordan,

The University of Pennsylvania Biochemistry and Molecular Biophysics Graduate Group, PA 19104 USA.

Ravi Radhakrishnan

The University of Pennsylvania Departments of Bioengineering, Chemical and Biomolecular Engineering, Biochemistry and Biophysics, PA 19104 USA.

Abstract

A method to predict the activation status of kinase domain mutations in cancer is presented. This method, which makes use of the machine learning technique support vector machines (SVM), has applications to cancer treatment, as well as numerous other diseases that involve kinase misregulation.

I. INTRODUCTION

Cancer is an evolutionary disease whereby an heterogeneous population of cells acquire a fitness advantage over neighboring cells via such mechanisms as mutations, changes in expression levels, and epigenetic factors, among others. These changes allow cancerous cells to have altered phenotype relative to the parent cells that they derive from, including increased proliferation and invasiveness, as well as less susceptibility to apoptotic signals than non-cancerous cells [1, 2]. The continuing decline in the cost of genome sequencing, as well as the relative ease of interpreting the effects of mutations in many proteins via methods such as activity assays has lead to a sustained drive to understand the effects of cancer derived mutations on cancer progression. The challenge of finding mechanistic links between mutations and cancer progression is made even more imperative by the fact that many cancer drugs target mutations that have specific effects, as well as the observation that many clinical trials fail due to patient cohorts that are not suitable for specific therapies [3]. Sequencing efforts as well as the frequent failure of targeted therapies has lead to an increasingly well-recognized principle that not all mutations confer selective advantage on cancer cells. These mutations are known as passenger mutations while mutations that confer some advantage are commonly referred to as driver mutations, because they can be seen as driving cancer progression [4].

The growing understanding of the importance of mutations on cancer progression is reflected in the rapid increase in available sequencing data via repositories such as The Catalog of Somatic Mutations in Cancer (COSMIC) and The Cancer Genome

Atlas (TCGA). Coterminal with this increase in data has been an increase in efforts to computationally assess the effects of these mutations. These methods generally seek to use machine learning on large datasets of variants that are known to be (or not to be) deleterious. Features that are common among one class of known variants or the other are then used to predict what will be the effect of new variants. One pioneering and still widely cited method to classify mutations called Sorts Intolerant From Tolerant (SIFT) [5]. This tool uses evolutionary sequence conservation to predict whether a single nucleotide polymorphism (SNP), not necessarily cancer related, will be deleterious on protein function. Since the overlap between mutations that have a deleterious effect on protein function and mutations that are cancer drivers may not be complete, this method may not be well suited to predicting the effect of cancer mutations, although many more recent methods have included the results of SIFT as one among many features use to classify mutations as well as used SIFT as a baseline for comparison of classification accuracy and sensitivity.

The most popular technique used to classify mutations that are specific to cancer is a method from machine learning known as support vector machines (SVM) [6]. This essentially geometric method seeks to find combinations of features that are common to mutations of different classes so that mutations of unknown class (i.e. driver or passenger) can be classified. SVM has been used to develop classifiers that predict whether mutations across the whole genome [7] as well as in a specific class of proteins, kinases [8]. These methods report that they make accurate predictions, based on cross-validation, and that they have a high receiver operating characteristic area under the curve (AUC), or the probability of distinguishing between examples of different classes. However, they both suffer from a problem that is seemingly common in the literature whereby they make a priori decisions about which mutations belong in which class, instead of letting the data lead the way. This should make the results of these and any similarly constructed classifiers suspect.

A brief comment on the selection of kinases as a class of protein to receive special attention is warranted. Kinase proteins have an important role in numerous cell signaling processes and have been seen to play an outsized role in human cancers. Indeed, recent work has shown that in COSMIC, kinase domain mutations account for over 20% of non-synonymous coding mutations in cancer, even though they account for less than 3% of protein coding genes in the human genome investigation [9]. Additionally, which kinase domain mutations are driver mutations can be determined by performing kinase activity assays or cell colony transformation assays [10]. These facts, combined with the fact that kinase proteins are the target of numerous drugs both in the clinic as well as in various stages of clinical trial, shows that this is a class of proteins worthy of detailed investigation [9].

II. METHODS

A. Kinase Dataset

The dataset was constructed via mining of uniprot and the literature for mutations in a subset of human kinases consisting of 468 kinase domain containing proteins. Left out were kinases that are uncharacterized in terms of kinase activity, several kinases that have large insertions that cause them to have unique structural features, and kinases that are known

to be constitutively active. The dataset contains 135 mutations that cause kinase domain activation as well as 413 mutations that do not cause activation.

B. Support Vector Machines (SVM) Construction

The kinase dataset was converted into a feature vector using an in house perl script. The elements of the feature vector include: the wild type and mutant residue identity, the wild type and mutant residue chemistry (i.e. aliphatic, acidic, basic, polar, aromatic), change in Kyte-Doolittle hydropathy from wild type to mutant as well as change of mutant from average Kyte-Doolittle hydropathy at that position, similarly for free energy of solvation, change in van der Waals radius, change in charge, and change in polarity. Also included are the SIFT score, whether the mutation occurs in a functionally important region in the kinase domain such as the α C helix, the nucleotide binding loop, the activation loop, or the catalytic loop. Whether the loop is predicted to be a β sheet, α helix, or unstructured loop, and the predicted solvent accessibility as given by Rost and Sander [11]. Finally, the proportion of a specific mutant occurring at a specific location relative to the total number of mutants in the whole kinase domain of the protein that a mutation occurs in was also included as a measure of oncogenicity. These features were encoded in a form readable by the program SVM^{perf}, which was used for all model creation and validation. The tradeoff between training error and margin (the parameter c in SVM^{perf}) was set to 0.25, a 1-slack dual structural learning algorithm was used (parameter w_3), a linear classifier without a bias term was used (parameters t_0 and b_0 respectively), and a loss function which maximized the area under the ROC curve was used (parameter l_0). This set of parameters gave the best outcome on cross-validation.

C. Model Validation

Performance was evaluated via cross-validation. The datasets used for cross-validation were always balanced, as very imbalanced datasets used for training purposes can lead to classifiers which predict the dominant class too often. In order to obtain a balanced dataset from the strongly imbalanced dataset described above, the number of members of minority class, in this case activating mutations, was used to determine the number of members in the majority class for training and testing. This was achieved by taking all members of the minority class and randomly selecting the same number of members from the majority class. Cross-validation was performed by taking 75% of the balanced dataset and using it for model training while using the remaining 25% of the data for testing. This was performed 300 times to account for the fact that data is left out of the balanced training/testing set. For the purposes of evaluating the results, the following abbreviations are used: true positives (TP), true negatives (TN), false positives (FP), false negatives (FN). Here, positive is taken as kinase activating and negative is taken as kinase non-activating. As well, the following definitions are useful.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

In words, accuracy is the percentage of predictions which are correct, sensitivity is the percentage of activating mutations predicted correctly relative to the total number of activating mutations, and precision is the percentage of activating mutations predicted correctly relative to the total number of activating mutations predicted. One more measure is worth mentioning here, the AUC, which as noted in the introduction is the probability that given an activating mutations and a non-activating mutation, the classifier will correctly identify which is which. All cross-validation was done with an in house perl script. As a final test of the method, the results were compared to those of a more mechanistically detailed, and computationally expensive, study of several mutations in the kinase domain of the anaplastic lymphoma kinase (ALK) [10].

D. Ranking Feature Vectors

One potentially important result of this study is which features are most useful for classifier performance. In this study, we use the computationally simple method of comparing differences in feature vector means. The mean and standard deviation of each class is used to determine the distance between distributions based on the formula $|m_1 - m_2| / (s_1 + s_2)$, where m_1 and m_2 are the sample means and s_1 and s_2 are the sample standard deviations for the class, and the elements in the feature vector have been normalized to lie in $[-1, 1]$. Features can then be ranked based on this metric.

A method known as kernel density estimation can be used to determine non-parametric probability distributions for the two classes of each feature. This method is akin to histograms but does not suffer from issues related to selecting bin width or origin. Kernel density estimates were constructed using linear kernels and a width parameter determined by cross-validation via the python package *sklearn*. Visualization was performed with the python package *matplotlib*.

III. RESULTS AND DISCUSSION

A. Classifier Performance

As seen in Table I, the method is fairly accurate and precise, though sensitivity is somewhat lacking. The area under the curve is also quite good. Table I also shows the importance of using a balanced dataset, as there is no metric on which the whole dataset does better than a balanced one. The chief reason that the model (which uses the whole dataset) does poorly on sensitivity and precision is that it too frequently predicts that a mutation is a member of the dominant class. The model was also constructed using an alternative package, a python implementation of *libsvm*, which allow the user to set a parameter to control for imbalanced datasets. Extensive testing of parameters using this alternate setup did not produce results

that were commensurate with those produced by the balanced dataset in SVM^{perf} (results not shown). It remains to be seen if better results could be obtained via the use of a sampling method that synthetically generates examples of the minority class to arrive a balanced dataset such as the synthetic minority oversampling technique (SMOTE) [12]. This remains a future direction.

B. Feature Rankings

A listing of the top six ranked features is given in table II. These six features show the largest differences in sample mean, together accounting for almost half of the cumulative difference of sample mean across all features.

The largest difference in sample means is found to have alanine as the mutant residue. This is depicted in Figure 1A. The likely reason that the ranked list is topped by having the mutant residue be an alanine is that many of the mutations found in uniprot are the product of alanine scanning mutagenesis. This should make us wary of making assumptions on the character of cancer driver mutations without a closer look at the data and how it is generated. Given the cost of making kinase domain mutants and performing activity assays or cell transformation assays, the reliance on datasets with some systematic bias such as uniprot will continue for the foreseeable future. The second, fifth, and sixth ranked features are all related and tell an interesting story. These features are all related in that they involve a charged residue. This is significant as kinase domains in the inactive state tend to be stabilized by salt bridges that are broken in the active state, which in some cases has its own stabilizing salt bridges. Mutations that result in charge loss or reversal could easily disrupt the delicate balance that holds the kinase domain in the inactive conformation, or alternatively could add stabilizing interactions to the active state. Either of these could result in constitutive kinase activation and aberrant signaling. Interestingly, this result is also in line with a recent study of kinase domain mutations in the COSMIC database. This study showed that many of the most frequently observed driver mutations in kinase domains are the result of loss or reversal of charge and concomitant inactive state destabilization [9]. The distribution of mutations by class for which the wild type residue is lysine is shown in figure 1B. The third and fourth ranked features, shown in figure 1C and 1D, are also related and may offer their own unique insight into which kinase domain mutations are likely to be drivers. These features both are related to polarity. Large changes in the polarity of even one residue can have a destabilizing effect on an entire protein by causing reorganization of a hydrophobic core. This reorganization could easily be responsible for a shift in population from inactive to active in a kinase domain. Figure 1C and 1D show that activating mutations tend to increase hydropathy scores and decrease polarity. Both of these can result in decreased potential for hydrogen bonding or salt bridge formation. This is yet further evidence that kinase domain activating mutations tend to destabilize the interactions that maintain the inactive conformation.

C. Comparison to Mechanistic Methods of Mutation Classification

One recent study of kinase domain mutations involved performing molecular dynamics (MD) simulations on a series of 22 clinically observed mutants of ALK. This study used

various criteria to assess whether mutants were destabilizing the inactive state of ALK, such as changes in hydrogen bonding, solvent accessible surface area, and proximity to conserved kinase structural motifs such as activation loop or α C helix [10]. A comparison of the results of the SVM classifier outlined above and the MD classifier of Bressler et al. [10] is given in Table III.

As is made clear by Table III, these methods both do quite well on predicting kinase domain activating mutations, with the MD method doing slightly better. What is striking about this table though is the computational costs associated with the two methods. The MD method takes thousands of hours of processing time on a scientific compute cluster to generate the MD trajectory, and some additional time to run analysis scripts. This must be done for each mutation. In contrast, the SVM method takes a few minutes of compute time on a commodity desktop to generate the feature vector, and then a few seconds to train the model and test the mutations. Although the MD method offers mechanistic insight that cannot be gained from the SVM method, the speedup of several orders of magnitude seems to point to SVM as the more efficient model, even if it is slightly less accurate. Moreover, as outlined below, a combination of the two approaches can yield an optimal (accurate as well as fast) strategy for profiling the effects of mutations.

IV. CONCLUSION

Here we have outlined a method to predict whether clinically observed cancer mutations are driver mutations. The method, based on support vector machines was shown to be fairly reliable at predicting the effect of kinase domain mutations, especially when a balanced dataset is used. It is also faster than recent efforts to use MD to predict the effect of kinase domain mutations, and does not make any *a priori* decisions about which mutations are driver mutations, as many recent machine learning efforts have. Interestingly, we find that ability to affect salt bridge formation can be an important factor in determining whether a given mutation is likely to be a driver. Similarly, mutations that change residue polarity, possibly leading to altered packing, are also important.

While the method outlined here is robust, it could still be improved upon. One class of features that are notably absent from this classifier is ones relating to the three dimensional structure and dynamics of a protein. Inclusion of this sort of information could result in even more reliable classifiers. One particular avenue of future research that may prove useful is the addition of structural information in the form of protein structure graphs. This could also be a way to include insight from MD simulations into the method, as features of protein structure graphs that vary between simulations of activating and non-activating mutations could be easily included in an SVM as long as the appropriate structural information is available (in the form of crystal or computationally generated structures of kinases). This goal is something that we are actively pursuing.

ACKNOWLEDGMENT

We thank G. S. Stamatakis, N. Graf, and members of the Radhakrishnan Laboratory for insightful discussions.

The research leading to these results has received funding from the European Commission grant FP7-ICT-2011-9-600841 and National Institutes of Health Grant U01-EB016027. Computational resources were provided in part by the National Partnership for Advanced Computational Infrastructure under Grant No. MCB060006 from XSEDE.

REFERENCES

- [1]. Tian T, Olson S, Whitacre JM, and Harding A, "The origins of cancer robustness and evolvability," *Integr Biol (Camb)*, vol. 3, pp. 17–30, 1 2011. [PubMed: 20944865]
- [2]. Hanahan D and Weinberg RA, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, pp. 646–74, 3 4 2011. [PubMed: 21376230]
- [3]. Normanno N, Rachiglio AM, Roma C, Fenizia F, Esposito C, Pasquale R, et al. , "Molecular diagnostics and personalized medicine in oncology: challenges and opportunities," *J Cell Biochem*, vol. 114, pp. 514–24, 3 2013. [PubMed: 22991232]
- [4]. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr., and Kinzler KW, "Cancer genome landscapes," *Science*, vol. 339, pp. 1546–58, 3 29 2013. [PubMed: 23539594]
- [5]. Ng PC and Henikoff S, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res*, vol. 31, pp. 3812–4, 7 1 2003. [PubMed: 12824425]
- [6]. Vapnik VN, *Nature of Statistical Learning Theory*: Springer, New York, NY, 1995.
- [7]. Capriotti E and Altman RB, "A new disease-specific machine learning approach for the prediction of cancer-causing missense variants," *Genomics*, vol. 98, pp. 310–7, 10 2011. [PubMed: 21763417]
- [8]. Izarzugaza JM, del Pozo A, Vazquez M, and Valencia A, "Prioritization of pathogenic mutations in the protein kinase superfamily," *BMC Genomics*, vol. 13 Suppl 4, p. S3–2012.
- [9]. Jordan EJ and Radhakrishnan R, "In silico profiling of activating mutations in cancer," at Review, 2014.
- [10]. Bressler S, Weiser D, Huwe PJ, Park JH, Krytska K, Ryles H, et al. , "Integrative functional assessment of ALK mutations for therapeutic stratification in neuroblastoma," *Cancer Cell*, vol. in press., 2014.
- [11]. Rost B and Sander C, "Improved prediction of protein secondary structure by use of sequence profiles and neural networks," *Proc Natl Acad Sci U S A*, vol. 90, pp. 7558–62, 8 15 1993. [PubMed: 8356056]
- [12]. Chawla NV, Bowyer KW, Hall LO, and Kegelmeyer WP, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

Figure 1: Kernel density estimates for the top ranked SVM features as determined by difference in mean feature value for each class.

TABLE I.

SVM Performance Metrics

	DATASET	
Performance Metric	Balanced	Whole
Accuracy	78%	78%
Sensitivity	67%	44%
Precision	81%	56%
AUC	91%	82%

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE II.

Top Six Ranked SVM features

<i>Feature Name</i>	$\frac{ m_1 - m_2 }{s_1 + s_2}$
Mutant A	0.144
Wild Type K	0.128
Kyte-Doolittle Hydropathy	0.089
Polarity Difference	0.084
Mutant Acidic	0.078
Wild Type Basic	0.075

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE III.

Comparison to MD Based Classifier

Contingency	Method	
	SVM	MD
TP	10	11
TN	6	6
FP	3	3
FN	3	2

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript