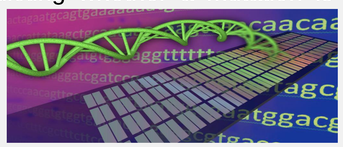
In silico profiling of activating mutations in cancer

Keshav Patil[1], Krishna Suresh[2] and Ravi Radhakrishnan[1,2]

[1] Department of Chemical and Biomolecular Engineering; [2] Department of Bioengineering, University of Pennsylvania, Philadelphia PA

In silico Oncology for Personalized Medicine

- Progress in genomics has raised expectations in personalized cancer research which means it is becoming increasingly feasible to personalize treatment according to the composition of patient's genome.
- The advancement in the sequencing techniques and subsequent large scale sequencing projects have generated copious data on somatic mutations in cancer.
- The majority of these somatic mutations incur little or no functional consequence on tumor progression (neutral passengers), while few (driver mutations) provide a selective advantage to cancer cells thereby making it necessary to distinguish driver mutations from the large number of neutral mutations.
- Profiling the mutations in the tumor cells and targeting the associated signaling pathways guides in rationalized targeted drug design therapy aimed at personalized cancer treatment.
- Kinases are frequently mutated proteins in cancer that account for 2% of all the mutations in the COSMIC database(taking into account clinical data subject to whole genome sequencing only).
- Mutations in the structural domains that activate kinases and upregulate cell proliferation are well represented among known oncogenic driver mutations. Many have been clinically observed and experimentally verified, but determining a priori which mutations in a given kinase are activating (facilitated through computational techniques) is still quite challenging but necessary, as exhaustive experimentation to study the mechanism of each mutation is inefficient and expensive.
- Existing in silico methods mostly use sequence-based methods such as SIFT or additionally information from 3D crystallized structures like PolyPhen-2 for predicting whether a mutation is deleterious to protein function.



Mining Molecular/ Mutation Data from Cancer Atlases

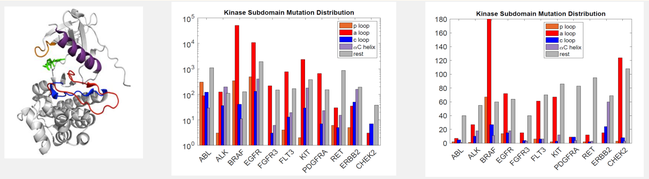
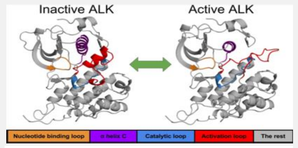
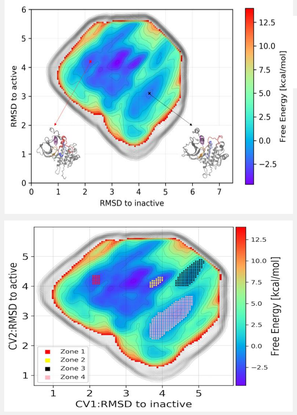


Figure 1. (left) Structure of a tyrosine kinase domain with different subdomains colored; see labels on legends in middle and right panels. The structure derived from the epidermal growth factor receptor tyrosine kinase (PDB ID: 2GS2). The nucleotide binding loop (p-loop), αC helix, activation loop (A loop), and the catalytic loop (C loop) are highlighted. (middle) Histograms of number of clinically observed cancer mutations in kinase domains constructed from COSMIC (version v87, 2018); here each count is an observation in one patient. (right) Histograms of number of amino acids mutated in the kinase domain pooled from clinically observed cancer mutations in kinase domains constructed from COSMIC (year 2018); here each count is a mutation at an amino acid location. We note that the middle panel include patient data from targeted sequencing as well as whole genome sequencing while the right panel includes data from whole genome sequencing only. These figures are provided to motivate the prevalence of mutations in cancer patients only. As a cautionary note, any statistical analysis on such data should consider the bias factors involved in targeted sequencing.

Computational Methods: Enhanced Molecular Dynamics (MD), Machine Learning



Capturing the transition from inactive form of the anaplastic lymphoma kinase (ALK) to its active form (see Fig. on left) involves unattainable simulation times. This is resolved using enhanced MD techniques.



$$V(s,\vec{t}) = \sum_{k\tau<t} W(k\tau)\exp\left(-\sum_{i=1}^{d}\frac{(s_i - s_i(q(k\tau)))^2}{2\sigma_i^2}\right). \qquad V(s,\vec{t}\to\infty) = -F(s)\vec{} + C.$$

In metadynamics, an external history-dependent bias potential is constructed in the space of a few selected degrees of freedom s(q), called collective variables (CVs). This potential is built as a sum of Gaussian kernels deposited along the trajectory in the CVs space. This enables to accelerate the system along CVs in a single simulation. The free energy landscape from metadynamics is depicted in Fig. to the left. Four zones are identified for further analysis.

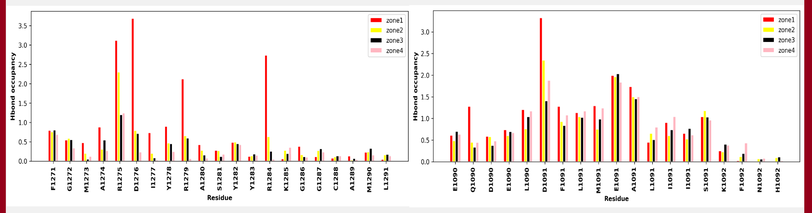
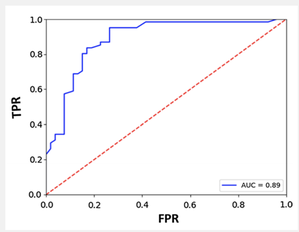
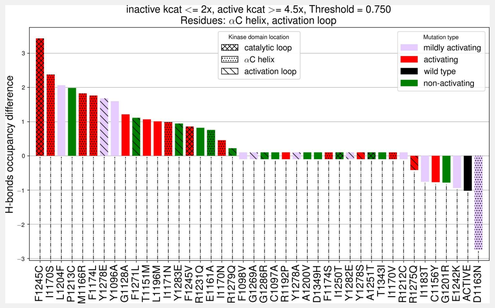


Fig. above – Hydrogen (H) bond occupancy in the activation loop (left) and in the alpha helix (right). Spikes of distinguishing lengths are observed which suggest excellent features for Machine learning algorithm

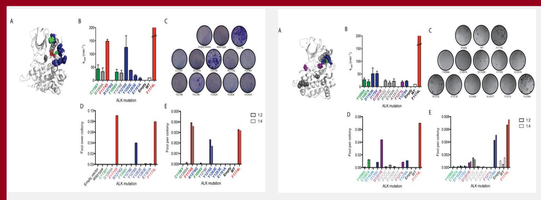


A ROC curve is constructed to analyze the classifying ability of the Support Vector Machine which has only biochemical features like wild type and mutant residue type (aliphatic, acidic, basic, aromatic and polar), sequence conservation (percent of total mutations occurring at residue , amino acid size and charge, free energy of solvation, Kyte-Doolittle hydropathy, SIFT score and secondary structure etc.

Results



Hbond occupancy difference (Fig. left) is computed for 45 ALK mutations that we studied to classify as activating or not. Hbond occupancy as a single feature has a balanced accuracy of 81% as determined from comparison with experiment (Fig. left).



Kinase activity assays and transformation assays were performed on cells transfected with ALK variants for mutations observed in Vietnamese patients (Fig. left).

Synthetically created mutations ((left) and mutations found in Vietnamese patients (right)

Mutation, ErbB2	MD PREDICTION	SVM PREDICTION	SIFT	PolyPhen2	EXPERIMENTAL
L755S	1	1	1	1	1
D769H	1	1	1	1	1
L768S	1	1	1	1	1
V777L	1	1	0	0	1
D769Y	1	0	1	1	1
Y835F	0	0	1	1	0
R896C	0	0	0	0	1
S760A	0	0	0	1	0
I767M	0	0	1	1	0
V773L	0	1	0	1	1
V842I	0	1	1	1	1
TPR	0.625	0.75	0.625	0.75	
FPR	0	0	0.6666	1	
BACC	0.8125	0.875	0.4792	0.375	

Mutation [All]	MD PREDICTION	SVM PREDICTION	SIFT	PolyPhen2
TPR	0.5946	0.8378	0.7568	0.9189
FPR	0.1111	0.2222	0.8333	0.9444
BACC	0.7417	0.8078	0.4617	0.4872

TPR=true positive rate; FPR=false positive rate; BACC= (1-FPR+TPR)/2=Balanced Accuracy

The performance of MD, SVM and existing methods like SIFT, PolyPhen2 are validated against experimentally obtained results and it is found that methods which implement biophysical properties have a significantly better performance and specifically lower false positives which eventually leads to great economical gain; Table, left.

Ongoing Work

To compute the H-bond occupancies and other distinguishing biophysical properties for large datasets of mutated protein structures, it is required that the operation is streamlined with automation; so we have developed an open source computational platform: BioPhysCode (see Fig. left) that essentially facilitates executing and analyzing large numbers of trajectories.

Conclusion and future work

- ❖ Methods that employ biophysical properties as features could boost the prediction performance if used in conjunction with biochemical and sequence based features.
- ❖ The mutations classified as activating need to be investigated in the presence of inhibitor drugs for reduced deleterious activity which thereby would help oncologists in personalizing therapy.

References

1. Jordan EJ, **Patil K,** Suresh K, Park JH, Mosse YP, Lemmon MA, Radhakrishnan R: Computational algorithms for insilico profiling of activating mutations in cancer. *Cell Mol Life Sci. 2019*

2. Computational Studies of Anaplastic Lymphoma Kinase Mutations Reveal Common Mechanisms of Activation Amidst a Varied Mutational Landscape in Neuroblastoma Patients. Jordan EJ*, Park J. H*, **Patil K*** et al. [*in preparation*].

Acknowledgements

We thank Dr. Jin Park, the labs of Mark Lemmon and Yael Mosse for experimental support. The research is supported by NIH, ERC, XSEDE.