**Computational Algorithms for In Silico Profiling of Activating Mutations in Cancer**

E. Joseph Jordan1, Keshav Patil2, Krishna Suresh2, Jin H. Park3,4, Yael P. Mosse5,6, Mark A. Lemmon3,4, and Ravi Radhakrishnan1,2,7,\*

1Graduate Group in Biochemistry and Molecular Biophysics, University of Pennsylvania, Philadelphia PA USA

2Department of Chemical and Biomolecular Engineering, University of Pennsylvania, Philadelphia PA USA

3Department of Pharmacology, Yale University, New Haven CT USA

4Cancer Biology Institute, Yale University, West Haven CT USA

5Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA USA

6Childrens Hospital of Philadelphia, Philadelphia PA USA

7Department of Bioengineering, University of Pennsylvania, Philadelphia PA USA

\* Address correspondence to rradhak@seas.upenn.edu

**Abstract**: Methods to catalog and computationally assess the mutational landscape of proteins in human cancers are desirable, and have implications for cohort selection in clinical trials, and for managing drug resistance to targeted therapies. The currently existing approaches to adapt evolutionary or sequence-based methods for predicting whether a single nucleotide polymorphism (SNP) is deleterious to protein structure and function are less than satisfactory as predictive tools. We present two classes of approaches, the first is structure-based and involves multiscale molecular modeling, and the second is chemical-biology-based and involves data science and machine learning. We compare their performances to those of existing methods on clinical datasets of observed mutations in kinase signaling proteins in human cancers. In cases where understanding the mechanism of protein activation and regulation is desired, structure-based computational approaches to predict the effects of point mutations. Through a case study of mutations in kinase domains of three proteins, namely, the anaplastic lymphoma kinase (ALK) in pediatric neuroblastoma patients, serine/threonine-protein kinase B-Raf (BRAF) in melanoma patients, and erythroblastic oncogene B 2 (ErbB2 or HER2) in breast cancer patients, we compare the two approaches above. We find that the structure-based method is most appropriate for developing a binary classification of several different mutations, especially infrequently occurring ones, concerning the activation status of the given target protein; this is especially true if the effects of mutations on the interactions of inhibitors with the target proteins are desired. However, many patients will present with mutations spread across different target proteins, making structure-based models computationally demanding to implement and execute. In this scenario, machine learning methods, including those based on support vector machines, are most appropriate for recognizing and illuminate mutational patterns. We show, however, that in the present status of the field, the two methods have very different accuracies and confidence values and hence the optimal choice of their deployment is context-dependent. We also identify unique opportunities for combining the two approaches in developing better-predictors in the future

**Funding:** FP7-ICT-2011-9-600841 (European Commission), U54 CA193417, U01 CA227550 (NCI)

**Keywords:** machine learning, molecular dynamics, enhanced sampling, driver mutations, passenger mutations, structural bioinformatics

**References**:

Integrative functional assessment of ALK mutations for therapeutic stratification in neuroblastoma, D. Weiser, S. Bressler, P. J. Huwe, R. Radhakrishnan, M. A. Lemmon, Y. Mosse, Cancer Cell, 2014, 26, 682-694. DOI: 10.1016/j.ccell.2014.09.019.

Machine learning predictions of cancer driver mutations, E. Jordan, R. Radhakrishnan, IEEE Proceedings of the 6th International Advanced Research Workshop on In Silico Oncology and Cancer Investigation, 2014, pp1-4. DOI: 10.1109/IARWISOCI.2014.7034632

In silico profiling of activating mutations in cancer, E. Jordan, K. Patil, K. Suresh, J. Park, Y. Mosse, M. A. Lemmon, R. Radhakrishnan, 2019, Cellular and Molecular Life Sciences, 76(14): 2663-2679. DOI: 10.1007/s00018-019-03097-2