
Instructions for Mac Users

MODELS PREDICTING EFFECTS OF MISSENSE MUTATIONS IN ONCOGENESIS

by

KanakaDurga Addepalli

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioinformatics and Computational Biology

Committee:

_____	Dr. Iosif I. Vaisman, Dissertation Director
_____	Dr. James D.Willett, Committee Member
_____	Dr. Ancha Baranova, Committee Member
_____	Dr. James D.Willett, Director, School of Systems Biology
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science
Date: _____	Spring Semester 2014 George Mason University Fairfax, VA

Models Predicting Effects of Missense Mutations in Oncogenesis

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

KanakaDurga Addepalli
Master of Science
University of Mumbai, India 1999

Director: Iosif I. Vaisman, Professor
Department of Bioinformatics and Computational Biology

Spring Semester 2014
George Mason University
Fairfax, VA

Dedication

This is dedicated to my parents for budding a dream in me and constantly encouraging and showing full confidence in me. I dedicate it to my wonderful husband whose love and support has helped me all through, and to my two beautiful children, Bhargav and Pranav who have shown a great level of love and understanding at a tender age.

Acknowledgements

I would like to very sincerely thank my advisor Dr.Iosif Vaisman whose support, guidance and patience has helped me reach my goal. His support has kept my aspiration alive and has instilled confidence in myself, time and again. Dr.Vaisman's passion towards Structural Bioinformatics has kept my interests in the subject also ignited, even after being exposed to other challenging '-omics' environments at work. I thank him for being supportive, patient and helpful even with the breaks I had to take during my long doctoral study.

I would like to thank my committee members, Dr. James Willett, Dr. Ancha Baranova for their valuable guidance and suggestions. I sincerely thank Mrs.Diane St. Germain and Chris Ryan for all their help.

I would like to thank my loving husband, Srinivas, who has been the most supportive and helpful. His support has kept me guilt free and my family afloat even in the most demanding times. It would be less how much ever I praise and thank my kids to have understood my aspiration and feel proud of my achievements. This wouldn't have happened without all the immense support and help from my parents and encouragement from my brothers. These have been the driving force, while my in-laws have shown equally special support and encouragement. Thanks to all of them.

I would like to sincerely thank all my friends. I thank Majjid Masso for generating a wonderful legacy of research work pertaining to the approach used in this research work here. A special thanks goes to my friends, Suneel Marthi and Venu Gangineni for providing technical guidance whenever needed.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

TCGA ..... The Cancer Genome Atlas
COSMIC ..... Catalogue of Somatic Mutations in Cancer
ICGC ..... International Cancer Genome Consortium
BIC ..... Breast Cancer Information Core
MSKCC ..... Memorial Sloan- Kettering Cancer Center
IDHCMM ..... Integrated Database of Human Cancer Missense Mutations
PDB ..... Protein Data Bank
SVM ..... Support Vector Machine
RF ..... RandomForest
SIFT ..... Sorting Intolerant From Tolerant
PolyPhen ..... Polymorphism Phenotyping
MAPP ..... Multivariate Analysis of Protein Polymorphism
PMut ..... Prediction of Pathological Mutations
SNAP ..... Screening for Nonacceptable Polymorphisms
PoPMuSiC ..... Prediction of Protein Mutant Stability Changes
PhD-SNP ..... Predictor of human Deleterious Single Nucleotide Polymorphisms
PANTHER ..... Protein ANalysisTHrough Evolutionary Relationships
topoSNP ..... topographic mapping of Single Nucleotide Polymorphism
LS-SNP ..... Large Scale Human SNP Annotation
Bongo ..... Bonds ON Graph
MuSiC ..... Mutational Significance in Cancer
CanDrA ..... Cancer Driver Annotation
Condel ..... CONsensusDELeteriousness
CAROL ..... Combined Annotation scoRing tool
CHASM ..... Cancer-specific High-throughput Annotation of Somatic Mutations
TransFIC ..... TRANSformed Functional Impact for Cancer
dbNSFP ..... Database of Human Nonsynonymous SNVs and Their Functional Predictions
CUPSAT ..... Cologne University Protein Stability Analysis Tool
SDM ..... Site Directed Mutator
ANN ..... Artificial Neural Network
HMM ..... Hidden Markov Model
PMSAs ..... Protein Multiple Sequence Alignments
PSSM ..... Position Specific Scoring Matrices
PSIC ..... Position-Specific Independent Counts
wt ..... Wild-Type
RI ..... Reliability Index

GEO ..... Gene Expression Omnibus
SAGE ..... Serial analysis of gene expression
RNA-Seq ..... RNA Sequencing
Weka ..... Waikato Environment for Knowledge Analysis
PAPAyA ..... Physician Accessible Preclinical Analytics Application
ITTACA ..... Integrated Tumor Transcriptome Array and Clinical data Analysis
ANOVA ..... Analysis of Variance
ROC ..... Receiver Operator Characteristic
AUC ..... Area Under Curve

Abstract

MODELS PREDICTING EFFECTS OF MISSENSE MUTATIONS IN ONCOGENESIS

KanakaDurga Addepalli, Ph.D.

George Mason University, 2014

Dissertation Director: Dr. Iosif I. Vaisman

The recent avalanche in high-throughput genotyping, next generation sequencing technologies and re-sequencing of cancer genomes has revolutionized the field of cancer genomics. It has generated a humungous amount of mutational data and changed the way the cancer is being studied. Identification and characterization of these mutations and their mutational effect has become one of the major goals of cancer research. We present here a computational geometry approach based on the application of Delaunay tessellation derived four-body statistical potential function where the potentials are directly derived from the high-resolution protein x-ray crystallographic structures utilizing their atomic coordinates. Proteins and their mutants are characterized by potential topological scores and profiles, which measure the relative change in the overall sequence-structure compatibility. Residual scores and profiles are generated which quantify environmental perturbations from wild-type amino acids at every mutational

position. We also present here an integrated database of human cancer missense mutations linked to their 3D structures, which has been created with the whole motivation of building a one stop shop of human missense mutations data sets huge and versatile enough to be used for training and testing of machine learning methodologies. With protein data from this database, we illustrate the use of potential topological cores and residual profiles in the prediction of mutational effects on protein structure and function and generating predictive models using machine-learning algorithms. We successfully apply supervised learning to training sets of protein mutants and generate models, which make statistically meaningful predictions of effects of missense mutations on cancer proteins.

Introduction

Serious efforts are being made in the field of cancer research to develop predictive models, which accurately estimate cancer development and genetic susceptibility of cancer patients. There have been numerous intensive translational research initiatives that make genomic, proteomic, and pre-clinical knowledge available to decision makers in the clinical research and clinical practice arenas. These programs have markedly improved the understanding of the cancer research community of the molecular processes leading to the initiation and progression of cancer and has shifted the kind of approach followed for therapy by targeting the effects of underlying genomic events driving the pathophysiology of cancer rather than previously used crude procedures like radiotherapy and chemotherapy. Deciphering the underlying molecular basis and genetic patterns of cancer will certainly help improve early cancer diagnosis and treatment. Working towards this, the recent progress in next generation sequencing technologies has revolutionized the field of cancer genomics. These advances in high-throughput genotyping and next generation sequencing have generated a humungous amount of human genetic variation data and therefore changed the way the genetic basis for human complex traits, including disease risk, is studied.

Cancer And Mutations In Human Genome

It is not a trivial task to identify genetic variants responsible for a complex multigenic disease such as cancer where the phenotype is defined by a combination of different genes and environmental factors effecting gene expression. The process of transforming a normal cell to a cancerous cell involves a series of complex genetic changes and single nucleotide polymorphisms are the most common types of genetic variations found in human cancers. Identification and characterization of these variants can provide an insight of the process involved and a basis for assessing susceptibility to cancer and an optimal choice of treatment required. This is being accomplished through various cancer genome-sequencing strategies and technologies[1], [5]. Systematic re-sequencing of the cancer genome has revealed genetic changes that may be responsible for lung, breast and colorectal cancers.[6], [5], [7]

A number of cancer somatic genome sequencing projects have been producing a flood of enormous mutational data. Making portions of the data open and accessible to the research community has made it possible for researchers worldwide to start analyzing the data and identifying genetic alterations in human cancer genomes as well as in normal genomes. Characterizing human cancers has become feasible even at the protein level and at a much lower cost in lots of cancer patients[8], [9]. Some of the large-scale efforts focusing on a ‘start to finish’ characterization of cancer at different levels and emphasizing a lot on

sequencing techniques are, TCGA (The Cancer Genome Atlas)[10] funded by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), COSMIC (Catalogue of Somatic Mutations in Cancer)[11] a database of somatic cancer mutations run by the Cancer Genome Project, based at the Wellcome Trust Sanger Institute and ICGC (International Cancer Genome Consortium)[12] an international organization coordinating genomic projects globally and providing collaboration among the world's leading cancer and genomic researchers, with an aim to generate comprehensive catalogues of genomic abnormalities in tumors from different cancer types.

These huge initiatives and many more smaller cancer genome sequencing projects have shed light onto the heterogeneity of different cancers showing that each has a set of mutations, which differ not just between cancer types or between individual to individual but also are intertumoral as well as intratumoral [13]–[15], [7]. This certainly calls for more systemic analysis of cancer genome mutations with large number of mutations identified in each gene. The tumor sequencing techniques will help clinicians and oncologists to demarcate and differentiate between tumor types and subtypes and therefore assist in formulating better diagnostic methods and treatments for cancer patients. Data from these variation studies would be of minimal use in understanding complex diseases unless the genetic variations are identified, characterized and interpreted. The effects of these mutations, especially the disease causing mutations, need to be annotated and evaluated accurately to

understand and relate them to cancer susceptibility and get a clinical relevant interpretation [16], [17].

The trend has shifted to studying the phenotyped clinical subjects with cancer and sequencing their genomes, transcriptional profiles and also few proteomic patterns, to extract the most informative and easily interpretable protein-coding fractions of the genome, and identifying the Single Nucleotide Polymorphisms (SNPs) or Mutations. Mutations within protein coding regions are of particular importance owing to their potential to give rise to amino acid substitutions that affect protein structure and function, which may ultimately lead to a disease state. It is of immense value to distinguish mutations as functionally relevant or irrelevant while identifying the disease mutations. A mutation within coding regions producing an amino acid substitution is called missense mutation. It gives rise to structural variations, which lead to functional disruptions/phenotypic variations in the protein and ultimately instigate a disease. Such mutations altering structure and function provide insight into the specific molecular mechanism responsible for the disease state. Analysis of such mutational profiles also provides insight into understanding the relationship between protein sequence and function. It has been shown that about 95% of mutation in common solid tumors such as colon, breast, brain or pancreas, are single base substitutions, out of which 90.7% are missense mutations[18]. In earlier studies, Sjoblom *et al.*[19] and Wood *et al.*[7] observed that missense mutations accounted for ~80% of the 1963 distinct somatic mutations found

after the removal of germ-line nucleotide polymorphisms. This clearly shows that missense mutations play a major role in oncogenesis. Shi and Moult[20] also observed that missense mutations in known cancer genes have a high impact on *in vivo* protein function. Missense mutations can affect the structure and/or function of the protein by having a dramatic effect on stability, hydrogen bonds, conformations and many other properties of proteins. Missense mutations are thought to be important factors contributing to the genetic functional diversity of encoded proteins [21]–[23] and have been identified recurrently in cancer genomes, hence are the most investigated group of mutations [14], [18], [7], [24].

Effects on Protein Function

The impact of these missense mutations on a protein's function can vary depending on there positions in the protein, their actual function and even on the mutant amino acid [25]. Several researchers have worked on developing computational methods to predict the effects of missense mutations on the function of the protein[26]–[36], [37, p. -], [38].

Predictive methods not only point out the effects of the mutations on the protein functions but also shed light on gene prioritization. They help in sorting the genes, which are more likely or play active role in causing a functional impact on the protein when mutated. Accurate predictions will help in minimizing the set of missense mutations to be characterized experimentally, in the process saving efforts, time and money. In fact, labels such as 'driver' mutations and 'passenger'

mutations have been used in literature to describe the fact that not all mutations found in the cancer genes are involved or play active role in tumorigenesis. The mutations that confer a selective growth advantage to the tumor cell are called 'driver' mutations. It has been estimated that each driver mutation provides only a small growth advantage to the cell, on the order of a 0.4% increase in the difference between cell birth and cell death. However, over many years, this slight increase can result in a large mass containing billions of cells[39]. It is still a significant challenge and is being seriously pursued to identify driver versus passenger mutations[40]. When it is not clear which of the mutations are driver mutations or passenger, a reasonable and an intuitive approach is to have a set of recurrent or overrepresented mutations in genes, i.e. each gene having at least a 100 mutations. This kind of a data set has been shown to support prediction assessment studies performed by Gnad and *et al.* [40]. In the first part of this dissertation we leverage this theory and collect as many mutations as possible against each gene published in different large-scale cancer gene databases globally. The first task towards this effort was to create an integrated database, IDHCMM, Integrated Database of Human Cancer Missense Mutations by integrating Missense mutations from widely used comprehensive cancer projects/databases such as TCGA, ICGC, and COSMIC. The IDHCMM database is described in detail in Chapter 1.

In this dissertation we explore a computational geometry approach based statistical scoring method, which uses a Delaunay tessellation-derived four-body

potential function to predict the functional impact of missense mutations. This potential is derived via an approach that uses the atomic coordinates of non-homologous, high-resolution protein structures. Since changes in protein structure effects protein function, it follows that the relative structural differences between variant proteins (i.e., single point mutants) and their wild type counter parts also correlate with the corresponding relative functional changes. We use a statistical scoring method for quantifying environmental perturbations expected to occur at all positions in a folded protein structure due to a particular amino acid replacement. Variants will be characterized by specifically focusing on perturbations at the mutated residue position and at the six structurally nearest positions.

Predictive methods

The plausible effects of missense mutations could range from affecting the protein stability to perturbing the protein interactions and cellular localization. An increasing number of computational tools are being developed to determine structurally and functionally unfavorable mutations. These predictive methods can be grouped based on the approach taken. Some are observation based, some probabilistic and some based on machine learning methods with wider set of attributes and training sets [21], [41], [34], [42], [36]–[38]. Most of them are based on sequence homology/ evolutionary sequence conservation methods [31], [43] while some include few structural attributes and some are based on physiochemical attributes. There are some prediction methods, which utilize a

combination of methods and their feature attributes. Supervised machine learning algorithms, such as support vector machines (SVMs) and Random Forest (RF) train models that perform a binary classification of single amino acid mutations in proteins as either neutral or deleterious to function. However, it is frequently the case that the functional effect of a polymorphism on a protein resides between these two extremes. Shi and Moult [20] established that destabilization of three- dimensional structure is the major molecular mechanism underlying driver missense mutations, therefore destabilizing mutations should preferably be determined by a structure-based approach.

The choice of features in any predictive method is of utmost importance as it decides the usefulness as well as the limitations of the method. No single method can consider all possible structural and functional features of a protein. This would not be feasible. For example, SIFT uses sequence homology to classify amino acid substitutions as tolerated or deleterious and prediction is based on conservation built purely on orthologous protein alignments. It does not distinguish intrinsically disordered regions and it has been recently observed that SIFT predictions have more false negatives on annotated disease mutations in disordered, solvent accessible and non-conserved regions [44].

So not all methods can assume a 'complete picture' in order to predict the effects of missense mutations. Traditionally there have been two classes of predictive methods, sequence conservation based and Structure and /or sequences based methods.

Sequence based methods

Sequence based methods exploit the evolutionary conservation of bases assuming that mutations in conserved positions in a multiple sequence alignment, across homologs tend to affect the protein structure and function drastically. It is also assumed that a disease-causing missense mutation in the current population is also disease-causing in homologous genes in other living or extinct species, or in other words the fitness landscape is constant [45]. An early observation that disease-associated missense mutations were overabundant at the evolutionarily conserved positions led to the use of multiple sequence alignment to help analyze missense mutations [46][47]. Protein multiple sequence alignments performed in different sequence based methods are relatively informative and provide reasonable sensitivity and specificity to missense prediction analysis when they have a significant alignment depth, i.e. have sampled enough sequences across evolutionary spread. These methods are then carefully constructed and curated to distinguish between positions that are functionally constrained and distinguish between different effects of different mutations. Therefore the variables that could effect the prediction of these algorithms include the genes involved, the number of sequences involved in the alignment, the evolutionary distances among species, the algorithm used and the importance of absolute amino acid conservation versus relatively conserved mutant amino acid [48]. Different approaches are designed to achieve this goal

ranging from simply listing the different amino acid residues present at that position, estimating the likelihood of the position being functionally constrained through phylogenetic tree based methods [49], calculating average BLOSUM62 scores for all amino acid pairs present in a mutation position [46] and measure the physicochemical variation that has been evolutionarily tolerated at the mutation position [50]. There are typically two steps in performing a conservation-based prediction, first choosing appropriate homologous sequences to build the multiple sequence alignment as the selection of sequences plays a major role in the accuracy of the prediction. A very shallow alignment depth is not informative where as high alignment depth may include very distant homologs and may deviate the prediction results. The second step is to evaluate the alignment. Different approaches have been used such as positional conservation measures, scoring functions, conservation of physicochemical properties (Align-GVGD [50], MAPP [51]).

Some of the earliest predictive methodologies are sequence based such as SIFT [43], MAPP [51] and PMut [52], which calculate sequence weights based on phylogenetic relationships between sequences, Align-GVGD [50] uses an approach based on conservation of amino-acid physicochemical properties, and other tools such as PolyPhen [53] and SNAP [54] use heuristic algorithms. Mutation Assessor [31] has a more elaborate conservation-based approach and has been seen to be yielding constantly high performance and prediction specificity. It distinguishes between conservation patterns within aligned

families (conservation score) and sub-families (specificity score) of homologs and so attempts to account for functional shifts between subfamilies of proteins. These are described in detail in the later sections of this dissertation.

Disadvantage of Sequence based methods are that they need diverse set of sequences and provide no insight into the nature of the underlying functional effect. Predictions using sequence conservation based methods are some times erroneous when some benign mutations are counted as deleterious and vice versa. This happens due to the presence of what are called Compensatory Pathogenic Deviations (CPDs). Compensatory changes in other sites of the same protein or its interaction partner may make a damaging mutation benign in other species. If this CPD is present in the sequences included in the multiple sequence alignment, the mutations might be wrongly predicted to be benign. A high prevalence of CPDs has been observed by a number of studies [55], [56]. Another disadvantage of these methods is the inability to rate the effect of the missense mutations. A very less damaging mutation could segregate within the populations at high frequencies, in such a case the corresponding amino-acid position will be conserved in the phylogeny and the mutations in that position may be predicted as highly damaging by conservation based methods.

Sequence and/or Structure based methods

Structure based methods examine the three-dimensional structural consequences of missense mutations and rely on the assumption that the function of a protein depends on the fundamental physiochemical properties

that can be derived only from the protein structure. Studying the relation of missense mutation to protein structure is a good approach for learning about protein structure and function and has been shown useful for structure based drug design [57]. Use of structural features provides direct insights into the role of mutations in molecular functions of the protein. Protein structures provide a detailed atomic level information and a mechanistic insight into why an amino acid change results in a change in protein properties or why a mutant has a damaging effect on protein function [58].

Although many combinations have been used in different structure based predictive methods, there seem to be three main strategies, namely, decision tree based classifiers, data vectors analyzed by machine learning algorithms to generate classifiers and molecular dynamics simulations. Decision tree based classifiers work with a set of features best extracted from the crystal structure of the protein such as, binding site, solvent accessibility, enzymatic site etc. The mutation is predicted to affect the protein function if the mutation violates an empirically determined condition. PolyPhen [53] is an example of this approach. A disadvantage of decision tree based methods is that they are not good at combining marginal results from two or more inputs [44]. This hurdle is crossed by machine learning algorithms such as Support Vector Machine (SVM), which can analyze multiple data types and consider joint effects of multiple inputs. Most of the methods are using multiple sources of both structural and

phylogenetic information, in a single classifier, to improve upon the prediction accuracy or performance.

PolyPhen-2 [32], PMut [52], MUPro [59], I-Mutant 2.0 [60], SNPs3D [61], LS-SNP [62], PhD-SNP [63], SNAP [64], MutPred [65] and nsSNPAnalyzer [66] are some examples of predictive tools that integrate both sequence and structural features and use machine learning algorithms, for predictions. PMut and SNAP use neural networks, SNPs3D, LS-SNP and PhD-SNP and I-Mutant 2.0 use SVMs, PolyPhen-2 uses Naïve Bayes, MutPred and nsSNPAnalyzer use Random Forest etc. There are other methods, which use other machine learning algorithms such as HMMs (Panther) or specifically designed custom algorithms. These combined classifiers though do not have high predictive accuracy have high success rate at predicting damaging mutations.

There has been a lot of emphasis on predictive tools, which predict the change in stability caused by a missense mutation. Shi and Moult [20] established that destabilization of three-dimensional structure is the major molecular mechanism underlying driver missense mutations, therefore destabilizing mutations should preferably be determined by a structure-based approach. Approximately 70% of monogenic disease mutations and 60% of very damaging germ-line missense mutations act through destabilization of protein three-dimensional structure, rather than via direct effects on molecular function [20] [67]. There are a lot of structure-based tools developed, which measure the change in folding free energy. Traditionally, Molecular Dynamics has been the

most straightforward way to estimate the folding free energy. But since Molecular Dynamics is computationally very extensive, lot of other functions that are not computationally very intensive, have been developed. These include purely statistical, empirical or knowledge-based energy functions. Examples of such tools are AUTO-MUTE [68], PoPMuSiC-2.0 [69], FoldX [70], CUPSAT [71], MultiMutate [72], Dmutant [73] etc. An overview of these tools is found in Table 1. There are methods, which do not look into energy functions but infer the proteins structural properties from its sequence such as MUPro [59], I-Mutant 2.0 [60].

However, some studies [74] showed that combining sequence and structure information can increase prediction accuracy to a certain degree. This gave rise to 'integrated analysis' where methods and approaches were developed where both sequence and structure inputs were either together to predict the mutational effects. The advantage of integrated analysis approach is that it can handle some uncertainty within each input parameter and does not require each method to output a perfect binary classification.

Machine learning algorithms

Machine learning algorithms are widely used for classification purposes in complex bioinformatics methods and approaches [75], [76]. The aim of using machine learning algorithms is to train a computer system to distinguish i.e. classify a set of test cases based on known examples, which is the training set. Typically for the machine learning methods, features related to missense

mutations are extracted, a classifier is trained using label-clear mutations and then classifications for the unknowns on the trained classifiers are performed. Automated learning from training data set is a reasonable alternative to tuning of empirical rules manually. Automated methods can explore more efficiently as to how the attributes of each mutation can be utilized to produce an optimal prediction. Machine learning methods can easily be cross-validated too. Machine learning approaches learn more complex nonlinear functions of input mutation, protein sequence, and structure information, than fitting a linear combination of energy terms. They are more robust in handling of outliers than linear methods, thus, explicit outlier detection used by empirical energy function approaches is not needed. Another advantage of machine learning algorithms is that they are not limited to using energy terms; they can easily leverage any relevant information. A good machine learning algorithm should have a good quality training set as the performance of the classifier depends a lot on the training from the training set. The training data set should represent the space of possible cases. This space of possible cases is too huge in case of missense mutations, therefore making it hard to formulate a benchmark variation dataset. Machine learning methods include several widely different approaches such as support vector machines, neural networks, Bayesian classifiers, random forests and decision trees. The quality of results of each predictor depends upon how the training has been done, what features are used to describe the phenomenon and optimization of the method.

It has been seen in case of functional predictors of missense mutations, that a combinations of methods and approaches yields better predictions. This calls for an increased number of features included for prediction. Caution has to be taken here and only features that best capture the effects of missense mutations should be incorporated to avoid the problem of dimensionality, which means that much more data is needed when the number of features increases.

The volume of the feature space grows exponentially with the dimensionality such that the data become sparse and insufficient to adequately describe the pattern in the feature space. Another problem which could arise is the 'over fitting' which means that the learner, due to sparse data, complex model or excessive learning procedure, describes noise or random features in the training dataset, instead of the real phenomenon. It is crucial to avoid 'overfitting' as it leads to decreased performance on real cases [77]. There are different statistical techniques to evaluate the machine learning algorithms, cross-validation being the most popular of these. Some of the other are random sampling and leave one out validation. Random sampling has a problem that the same cases may appear more than once in the test set and others not at all. Leave out one validation is computationally very intensive. As the name implies, one case at time is left for validation while the remaining cases are used for training. The computational requirements may be prohibitive with large datasets. A lot of the machines learning predictors are binary classifiers, but it is possible to have more than 2 classes of outputs.

Decision tree is a classifier for generating a pruned or unpruned decision tree and it is a mapping of observations to classification. The decision tree represents the classifier as a tree structure in which each node represents a decision based on an attribute value, and it leads to a set of predictive rules that can be interpreted easily [37].Decision trees are built with an inner node representing the variable, an arc to the child, representing a possible value of the variable and a leaf for the predicted value of target variable using the values of the variables represented by the path from the root. Dobson et al. [78] and Krishnan and Westhead [37] used decision trees in there predictive studies.

The Bayesian network uses various search algorithms and quality measures to find a minimum set of direct dependencies that together explain the observed correlations in the data. The best Bayesian network is the one that models the observed data using a measure of scoring metric, a trade-off between complexity and accuracy. Linear regression analysis relates the output with the linear combination of single/multiple input features [79]. Naive Bayes is the simplest Bayesian classifier. It is built upon the assumption of conditional independence of the predictive variables given the class [75]. PolyPhen-2 uses a Bayesian approach and is based on two Bayesian probabilistic models.

Support vector machine (SVM) is a learning algorithm, which from a set of positively and negatively labeled training vectors learns a classifier that can be

used to classify new unlabeled test samples. SVM learns the classifier by mapping the input training samples into a possibly high dimensional feature space, and seeking a hyperplane in this space which separates the two types of examples with the largest possible margin, i.e., distance to the nearest points. If the training set is not linearly separable, SVM finds a hyperplane, which optimizes a trade-off between good classification and large margin i.e. larger the margin, the better the generalization of the classifier. One important feature of SVMs is that computational complexity is reduced because data points do not have to be explicitly mapped into the feature space. Instead SVMs use a kernel function, to calculate the dot product of data vectors in feature space, obtained from a map from input space to feature space. The linear classification or regression function is computed from the Gram matrix of kernel values between all training points. Only data points with positive weight in the training dataset affect the final solution—these are called the support vectors. PHD-SNP, MuStab, MUpro, LS-SNP, SNPs3D, SAPRED, I-Mutant 2.0 and Scpred are some of the SVM based tools to predict functional impact of missense mutations.

Random Forest is a classifier consisting of an ensemble of unpruned classification or regression trees created by using bootstrap samples of the training data and random feature selection in tree induction. It is trained to optimally combine the heterogeneous sources of predictors using a curated training dataset prepared from the SwissProt database. In the training stage, an

RF builds a committee of decision trees and in the test stage it averages the results from all trees as the final output. In the tree-growing procedure, a random subset of attributes is selected at each node and the best one is used for splitting [65].

An Artificial Neural Network (ANN) is an information processing model that is able to capture and represent complex input-output relation- ships. It is a network of non-linear processing units that have adjustable connection strengths, hidden layers and the discrimination is mainly based on feed-forward networks using the back propagation-learning rule [76]. It learns and classifies a problem through repeated adjustments of the connecting weights between the elements. The goal of the method is to find a good input-output mapping, which can then be used to predict the test set [79]. SNAP, SNPdbe, PoPMuSiC-2.0, PMUT, MUpro, PoPMuSiC-2.0 and I-Mutant are some of the neural networks based tools to predict functional impact of missense mutations.

Mutational Data Sets

The performances of most of the predictive methods employing machine learning algorithms have a strong dependency on training data sets, i.e. the selected set of neutral and disease-causing missense mutations. It is not a straightforward task to generate an optimal set of either deleterious or neutral mutations for any predictive analysis, as there cannot be a uniform definition of

functionality across all the proteins. Protein function is context dependent, and so is the effect of a missense mutation on the protein function. Missense mutations may directly affect the normal function of proteins by altering binding sites in proteins such as protein, nucleic acid, ligand or ion binding sites. Protein function may also be affected by missense mutations that alter protein stability, protein aggregation or posttranslational modifications. In either case, where protein malfunction occurs, disease may result. Cases with experimentally validated known functional effects, which represent the real world i.e. having a distribution of the missense mutations closely resembling the distribution in the real world data, would form the ideal benchmark dataset. Datasets used for training or testing the predictive methods should be large enough to cover mutations related to all the features included (sequence based or structure based) and to have predictive statistical power. The datasets need to be non-redundant and need to contain both disease-causing and neutral mutations.

A dataset has been released recently, VariBench [80], which is a benchmark database suite comprising of variation datasets for testing and training methods for variation effect prediction. VariBench can be used for developing, optimizing, com- paring and evaluating the performance of computational tools that predict the effects of variations.

Gnad *et al* [40] observed that when CHASM was run using a test data set, which did not have mutations matching COSMIC mutations, the accuracy of CHASM dropped from 89% to 50%. CHASM was explicitly trained on COSMIC

mutations. They also observed that on a data set, which included only oncogenic driver mutations, CHASM showed a poor performance. This was thought to be because of, CHASM being trained to predict both tumor suppressor mutations as well as oncogene mutations. This clearly shows the importance of training set used to train various different predictors. Thusberg *et al.* [81] observed that poor performance by the predictive methods they tested in their study was not because of the differences in the size of the data sets but was because of other factors such as differences in the type of data.

Overview of Predictive tools

An overview of different predictive methods is presented below.

1. **SIFT**□

One of the earliest tools developed in this area, SIFT, (Sorting Intolerant From Tolerant) [43], uses sequence homology to classify amino acid substitutions as tolerated or deleterious and prediction is based on conservation built purely on orthologous protein alignments. Owing to its impressive predictive power and simplicity, SIFT continues to be used as a benchmark for other methods and approaches. SIFT considers the position at which the change/mutation occurred and the type of amino acid change. Given a protein sequence, SIFT chooses related proteins and obtains an alignment of these proteins with the query. SIFT uses Dirichlet mixtures extracted from these protein multiple sequence alignments (PMSAs) to create position specific scoring matrices (PSSM) and

score missense substitutions. Based on the amino acids appearing at each position in the alignment, SIFT calculates the probability for each of the 19 amino acid changes to be tolerated relative to the most frequent amino acid being tolerated. If this normalized value is less than a cutoff, the substitution is predicted to be deleterious. However, such a prediction could be unreliable if there are few homologs available. Better predictions are obtained if the users can provide their own curated alignments. SIFT scores of ≤0.05 are usually taken as indicative of deleterious substitutions. However, the authors point out that in some situations higher or lower cutoffs might give a more accurate result for binary deleterious/neutral classifications [43]. The method is easy to install and use. An overview of SIFT workflow is shown in Figure1 below, taken from SIFT publication by Kumar *et al.* [33]:

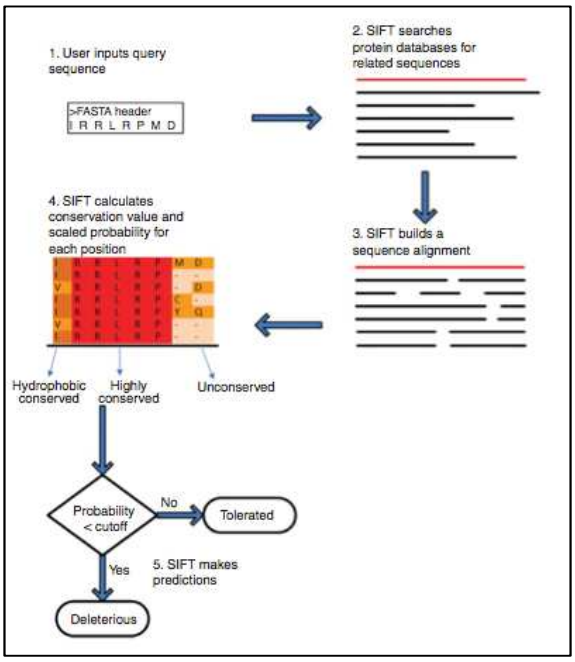


Figure 1: SIFT Workflow

1.1. B-SIFT

B-SIFT, Bi-directional SIFT [82] is a modified version of SIFT algorithm that utilizes protein sequence alignments with homologous sequences to identify functional mutations based on evolutionary fitness. B-SIFT attempts to classify both gain- and loss-of-function mutations. By calculating SIFT scores for both the mutant and wild-type alleles, it identifies potential gain-of-function mutations where the mutant residue is more similar to those found in homologous proteins.

2. PolyPhen□

PolyPhen (Polymorphism Phenotyping) [32] demonstrated that the combination of structural and evolutionary attributes, improve prediction. It predicts possible impact of an amino acid substitution on the structure and function of a human protein by using straightforward physical and comparative considerations. It uses a rule-based cutoff system to classify variants. It initially characterizes the input missense mutations by various, sequence, structure, and phylogeny based descriptors. The sequence-based characterization includes SWALL database [83] annotations for sequence features, a transmembrane predictor TMHMM [84] and PHAT [85] transmembrane-specific matrix score for substitutions at predicted transmembrane regions, the Coils2 program [Lupas et al., 1991] for prediction of coiled coil regions, and the SignalP [86] program to predict signal peptide regions. Phylogenetic information is derived by constructing a profile matrix from aligned sequences by the PSIC (Position-Specific Independent Counts) software [87]. The structural descriptors are obtained by mapping the missense variant onto the corresponding or similar protein and then using the DSSP program [88] for secondary structure information, solvent-accessible surface, and j–c dihedral angles. In addition, PolyPhen calculates the normalized accessible surface area and changes in accessible surface propensity resulting from the amino acid substitution, change in residue side chain volume, region of the Ramachandran map, normalized B factor, and loss of a hydrogen bond according to the Hbplus program [McDonald and Thornton, 1994]. The SWALL database annotations are utilized in the structure analysis such that the program

checks whether the substitution site is in spatial contact with critical residues annotated to be involved in forming binding sites or active sites. Additionally, the contacts of the substituted residue with ligands or subunits of the protein molecule are checked. After characterizing the variant, PolyPhen applies empirically derived rules based on the characteristics to predict whether a missense variant is damaging or benign. Figure2 below shows PolyPhen's query processing flowchart, taken from its publication by Ramensky *et al.* [53]:

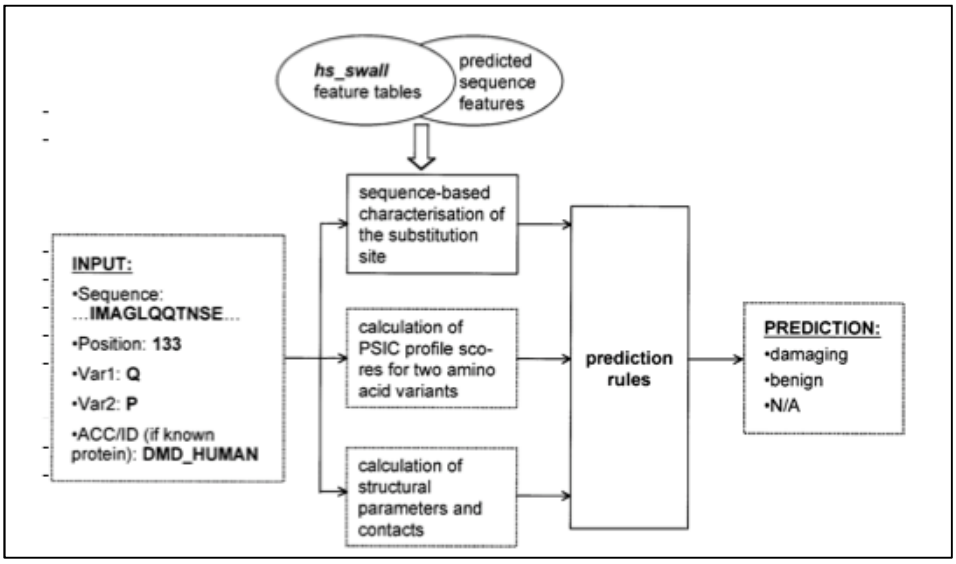


Figure 2: PolyPhen query process workflow

PolyPhen-2 differs from the early tool PolyPhen-1 in the set of predictive features, alignment pipeline, and the method of classification [53]. PolyPhen-2 uses eight sequence-based and three structure-based predictive features, which were selected automatically by an iterative greedy algorithm. The sequence-

based features include PSIC scores and MSA properties, and position of mutation in relation to domain boundaries as defined by Pfam [89]. The structure-derived features are solvent accessibility, changes in solvent accessibility for buried residues, and crystallographic B-factor. Majority of these features involve comparison of a property of the wild-type (ancestral, normal) allele and the corresponding property of the mutant (derived, disease-causing) allele, which together define an amino acid replacement. However, when there are not enough structural parameters, its classification is based predominantly on comparative analysis. Thus, structural attributes are complementary to evolutionary ones, rather than overlapping. PolyPhen2 predicts the effect of mutation using a naive Bayesian classifier.

3. MAPP□

MAPP, (Multivariate Analysis of Protein Polymorphism) [51] considers the physicochemical variation present in a column of a protein sequence alignment and, on the basis of this variation, predicts the impact of all possible amino acid substitutions on the function of the protein. MAPP quantifies constraint in terms of biochemical properties rather than substitutions. Analysis rests on two complementary ideas: 1. That, differences in standard physicochemical properties between the “wild-type” amino acid and the missense mutation are the root cause of functional impairment; and 2. that evolutionary variation among orthologs in the affected position is a sample of the physicochemical

properties that are tolerated at that position. MAPP was designed using these two ideas as a premise, and which quantifies the physicochemical variation in each column of a multiple sequence alignment and calculates the deviation of candidate amino acid replacements from this variation. The greater the deviation, the higher is the probability that a replacement impairs the function of the protein, and the greater is its predicted effect on the function of the protein. MAPP uses quantitative scales measuring six physicochemical properties to evaluate missense variants: (1) hydropathy [90] (2) polarity [91] (3) charge [91] (4) side-chain volume (Zamyatin 1972); (5) free energy in alpha-helical conformation [92] and (6) free energy in beta-sheet conformation [92].

MAPP consists of seven steps, shown in the MAPP analysis workflow in Figure3 taken from the MAPP publication by Stone *et al.* [51]. First it builds a multiple alignment of orthologs or closely related paralogs; distant paralogs are excluded to avoid including evolutionary variation that specifies functional differences. The sequences' evolutionary relationships are inferred by standard likelihood analysis, which also yields the branch lengths in substitutions per site, for the tree. Based on the topology and branch lengths of the tree, weights are calculated for each sequence that control for phylogenetic correlation among the sequences. Multiplication of the weights with the fraction of sequences carrying a particular amino acid yields the alignment summary. This is interpreted by using a matrix of physicochemical property scales. The result is an estimate of the physicochemical constraints on each position in terms of the mean and

variance of the property distributions observed in its alignment column. The statistics are stated to be biologically meaningful; the mean hydropathy value at a position estimates its hydrophobic character, while the variance measures the strength of that constraint. Deviations from the alignment column are obtained for each possible variant by calculating its property difference from the mean and dividing by the square root of the variance. This statistic is interpreted as a signed measure of constraint violation. To compute a single score measuring the violation of constraint across all properties, it first decorrelates the properties themselves by using a principal component transformation. The decorrelation gives rise to a new coordinate system in which each axis is a principal component; the distance from the origin to any variant is the variant's decorrelated impact score. An impact score is thus assigned to every possible variant in the protein. A high impact score identifies a potentially deleterious variant by virtue of its physicochemical dissimilarity to the observed evolutionary variation, whereas low-scoring variants are less likely to compromise protein structure or function.

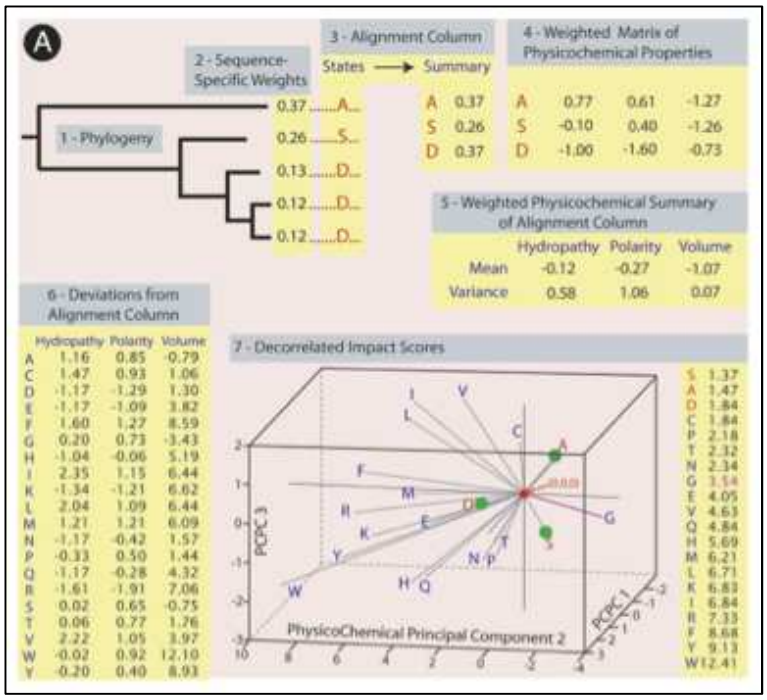


Figure 3: MAPP Analysis Steps

MAPP can distinguish intermediate from negative variants and allows a continuous classification, because its impact scores are widely spread across the sub-functional spectrum. MAPP outperforms SIFT in distinguishing positive from deleterious variants, even for the data set upon which SIFT was trained (LacI). MAPP's predictive accuracy is complemented by the interpretability of its impact scores, which provide a transparent rationalization of predictions in terms of physicochemical properties. Each variant's impact score can be dissected into individual components that measure property-specific constraint violations effectively assigning a rationale to every prediction.

4. AlignGVGD□

Align-GVGD [50] is a freely available, web-based program that combines the biophysical characteristics of amino acids and protein multiple sequence alignments to predict where missense substitutions in genes of interest fall in a spectrum from enriched deleterious to enriched neutral. Align-GVGD is an extension of the original Grantham differences to multiple sequence alignments and true simultaneous multiple comparisons. The biochemical variation at each alignment position is converted to a Grantham Variation score (GV). The difference between these properties and those of the variant amino acid being assessed are calculated and a Grantham Difference score generated (GD). These values are used, as a measure of how likely the substitution is to be deleterious or neutral on a classification spectrum. Align-GVGD requires an alignment as input. Predictions are found to be highly varied depending on the alignment used. Using highly divergent sequences in an alignment can introduce gaps and will result in all amino acids being classed as neutral at that position. In an attempt to overcome the problem of all amino acids being classed as neutral at a position because of the use of highly divergent sequences in an alignment that introduce gaps, manually curated alignments are available to use which contain 8-14 orthologous sequences from a range of species. These alignments cover ATM, BRCA1, BRCA2, CHEK2, TP53, MLH1, MSH1, MSH6, PMS2, RAD51 and XRCC2. These alignments have been carefully constructed to provide the correct amount of sequence divergence whist also using computational algorithms to

improve sequence coverage and reduce the number of gaps. When using Align-GVGD on the genes listed the provided alignments should always be used (http://agvgd.iarc.fr/alignments.php). For each of these gene alignments, the user must select the species depth of the alignment. For example for BRCA1, the alignment can span species from human to frog, human to puffer fish or human to sea urchin. The depth of the alignment will influence sequence diversity over the sites and thus effect the prediction. Mutation interpretation software, Alamut (version 2.1 - Interactive Biosoftware, Rouen, France), supplies alignments to the Align-GVGD. A-GVGD, can be used to identify sets of missense mutations that are either enriched for deleterious mutations or enriched for neutral mutations. However, A-GVGD does not account for the possibility that sequence variation that has been permissible during the evolution of BRCA1 in one group of non-human vertebrates is not permissible in human BRCA1. It also does not take into account that the nucleotide substitution underlying a missense variant may interfere with mRNA splicing or have some other deleterious effect at the level of DNA or RNA [50].

5. nsSNPAnalyzer

nsSNPAnalyzer, [66] is a web-server implementing a machine-learning method that combines the biophysical characteristics of amino acids and protein multiple sequence alignments to predict where missense substitutions in genes of interest fall in a spectrum from enriched deleterious to enriched neutral. Align-GVGD is, an extension of the original Grantham difference to multiple

sequence alignments and true simultaneous multiple comparisons. nsSNPAnalyzer takes a protein sequence and the accompanying nsSNP as inputs. The input protein sequence is searched against the ASTRAL database [93] for homologous protein structures, and calculates three types of information from user's input: (i) the structural environment of the SNP, including the solvent accessibility, environmental polarity and secondary structure (ii) the normalized probability of the substitution in the multiple sequence alignment using the SIFT method and (iii) the similarity and dissimilarity between the original amino acid and mutated amino acid. The program then uses a Random Forest classifier trained by a dataset prepared from the SwissProt database to classify the variant to be disease-associated or functionally neutral. Figure4, taken from nsSNPAnalyzer publication [66] shows the workflow.

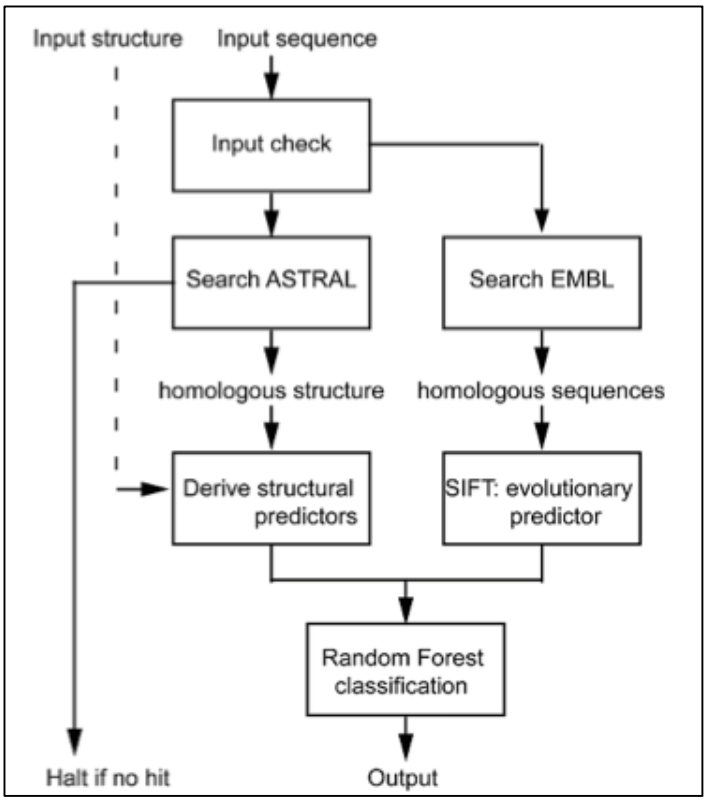


Figure 4: nsSNPAnalyzer Workflow

Two inputs are mandatory: protein sequence in FASTA format and the nsSNP identities to be analyzed. An nsSNP is denoted as X#Y, where X is the original amino acid in one letter, # is the position of the substitution (starting from 1), and Y is the mutated amino acid in one letter. Multiple nsSNPs in a protein need to be separated by new-line characters in the input. In addition to the two mandatory inputs, an accompanying protein structure file in PDB format can be uploaded, if the users want their own structure to be used. The output includes several calculated features of the nsSNP: (i) predicted phenotypic class (disease-associated versus neutral); (ii) a hyperlink to the homologous structure with a SCOP identifier; (iii) the normalized probability of the substitution calculated by

the SIFT program; (iv) area buried score, a measure of the solvent accessibility; (v) fraction polar score, a measure of environmental polarity related to hydrogen bond formation; (vi) secondary structure (helix, sheet and coil); and (vii) the structural environment class, a discrete environment class definition by combining features (iv)–(vi). The area buried score and fraction polar score are calculated by the ENVIRONMENT program [94], and the secondary structure is calculated by the STRIDE program [95].

6. SNPs&GO

SNPs&GO [96] is a web server for the prediction of human disease-related single point protein mutations, based on support vector machines. SNPs&GO is an SVM classifier based on the local sequence environment of the mutation, features derived from sequence alignment, prediction data provided by the PANTHER classification system [97] and a functional-based log-odds score calculated considering the GO classification. The main novelty of SNPs&GO is the use of functional GO terms. The final input vector consists of 52 values, 40 components encode for the mutations and sequence local information (Seq); four inputs concern features derived from sequence profile plus an extra one (a bit) codifying the presence/absence of the features themselves (Prof); four values represent selected parameters of PANTHER (prediction output plus an extra node encoding the presence/ absence of PANTHER output) (PANTHER); two components encode for the GO log-odd score (LGO) and for its

presence/absence (LGO). Results that are obtained adopting a cross validation procedure under different implementation conditions on the same training/testing set. Performance is measured by computing different scoring indexes: Q2, the overall accuracy; P(D), the rate of correct predictions for the disease- related mutations (D); Q(D), the coverage (number of correctly predicted mutations) for the disease-related mutations; P(N), the rate of correct predictions for the neutral mutations (N); Q(N), the coverage for the neutral mutations; AUC, an estimate of how the predictor is different from a random predictor characterized by AUC =0.5. The study states that with increasing complexity of information, the performance is enhanced, suggesting that in addition to the sequence profile, the LGO data derived from GO annotation improves the ability to discriminate neutral and disease-related SNPs.

SNPs&GO^{3d} is an extension of SNPs&GO including information extracted from protein 3D structure. It is stated that although SNPs&GO^{3d} has been tested on a smaller set of mutations, it results in a better accuracy with respect to the sequence based method. It predict deleterious single point mutations considering in a unique framework protein structure information, used for the prediction of stability changes in I-Mutant [60] [98], and protein sequence, evolutionary and functional information, used in the recently developed SNPs&GO algorithm [96]. The final input vector consisted of 52 elements, 20 components encoding for the mutations (Mut); 21 features representing local protein structure (Structure Environment); 5 features derived from sequence

profile (Prof); 4 features from the output of PANTHER method (PANTHER); 2 elements encoding the number of GO terms associated to the protein and the GO log-odd score (LGO). It differs from SNPs&GO only in the 21 elements encoding for the local protein structure environment (Structure Environment). These replace the 20 elements encoding for the sequence environment used by the sequence-based SVM predictor.

7. SNAP

SNAP (Screening for Nonacceptable Polymorphisms), [54] [64] is a neural network-based method for the prediction of the functional effects of non-synonymous SNPs. It perhaps spans the most comprehensive feature space [40]. SNAP needs only sequence information as input, but benefits from functional and structural annotations, if available. The method utilizes evolutionary information from PSI-BLAST [99] frequency profiles and PSIC [87], transition frequencies for mutations, biophysical characteristics of the substitution, secondary structural information, and relative solvent accessibility values predicted by PROFsec/ PROFacc [100] [101], chain flexibility predicted by PROFbval [102], protein family evolutionary information, and information about domain boundaries from Pfam [89], and SwissProt annotations to classify a missense variant. The training sets for the NN were constructed from Protein Mutant Database (PMD) data complemented by a set of neutral pseudomutations generated by the authors of the method as described in

Bromberg and Rost. A number of networks were trained before the optimal architecture and feature space were obtained for each data set. The only feature that was not altered in the network selection process was the presence of two output nodes, each ranging from 0 to 100. The difference between two outputs, sampled at a particular cutoff, determined the classification of the mutant. When additional features no longer improved performance window length, hidden node number, learning rate and momentum were varied. Further runs were only attempted if any of the changes stimulated an increase in overall accuracy. The results of these runs determined the architecture and input vectors for the final networks. As an input SNAP takes the wild-type sequence along with their mutants. A comma-separated list gives mutants as: XiY, where X is the wild-type amino acid, Y is the mutant and *i* is the number of the residue (i = 1 for N-terminus). X is not required and a star (*) can replace either *i* or Y. Any combination of characters following these rules is acceptable; e.g. X** = replace all residues X in all positions by all other amino acids, *Y = replace all residues in all positions by Y. Users may provide a threshold for the minimal reliability index (RI) and/or for the expected accuracy of predictions that will be reported back. These two values correlate so when both are provided, the server chooses the one yielding better predictions. For each instance SNAP provides a reliability index (RI), i.e. a well-calibrated measure reflecting the level of confidence of a particular prediction. As an output, for each mutant, SNAP returns three values: the binary prediction (neutral/non-neutral), the RI (range 0–9) and the

expected accuracy that estimates accuracy on a large dataset at the given RI (i.e. accuracy of test set predictions calculated for each neutral and non-neutral RI.

In a cross-validation test on over 80 000 mutants, SNAP identified 80% of the non-neutral substitutions at 77% accuracy and 76% of the neutral substitutions at 80% accuracy. This constituted an important improvement over other methods. The improvement rose to over ten percentage points for mutants for which existing methods disagreed. Possibly even more importantly SNAP introduced a well-calibrated measure for the reliability of each prediction. This measure will allow users to focus on the most accurate predictions and/or the most severe effects. The most important single feature for SNAP prediction is conservation in a family of related proteins as reflected by PSIC scores [87]. SNAP depends on many tools owing to its extensive feature space and is therefore not easy to install compared to other tools. For limited set of mutations it is preferable to use its website { https://www.rostlab.org/services/snap/}

8. PMUT

PMUT [52] is a software aimed at the annotation and prediction of pathological mutations and is based on the use of neural networks (NNs) trained with a large database of neutral mutations and pathological mutations. PMUT uses different kinds of sequence information to label mutations, and the neural networks to process this information. PMUT server works at two different levels 1. It retrieves information from a local database of mutational hotspots and 2. It

analyzes a given SNP in a specific protein. The first input to PMUT is either the sequence of the protein or its SwissProt/trEMBL code. The user has to select the mutation site and whether to analyze a single mutation, which is the default or to perform a complete mutation scan at this position. PMUT can simulate massive single-point mutation along the whole sequence (Mutation Hot-Spot analysis), helping to detect regions where mutations are expected to have a large pathological impact. Irrespective of the selection, the program retrieves a series of parameters describing the mutation [103][104] from (1) its internal databases, (2) PHD output [101] and (3) multiple alignments. The latter are either introduced by the user [e.g. from the PFAM database] or automatically generated by the program from a two-iterations PSI-Blast run on a non-redundant SwissProt/trEMBL database. Two neural networks are implemented as predictor engines, a large one with 1 hidden layer, 20 nodes and 15 descriptors and a small one with 20 nodes, no hidden layer and with 3 parameters. Results are displayed in the form of various text files and, when the structure is experimentally known, 2-D and 3-D plots are also available. It provides a very simple output: a yes/no answer and a reliability index (0-9). Additionally, the program allows users to retrieve all the intermediate information (alignments, Blast and PHD outputs, etc.) used in PMUT predictions. The cross-validated performance of the method is 84 % overall success rate, and 67 % improvement over random. PMUT also has a database, PMUT Database, which comprises of pre-computed mutation profiles of all the

proteins in the 90% identity cluster of the PDB database. All the residues of each protein were mutated to all 19 possible alternative amino acids. The mutation matrix is manipulated to define mutation hot spots in different ways: 1. Maximum, mean and minimum pathogenicity indexes in each mutation site, 2. the pathogenicity index associated with the mutation to Ala (alanine-scanning) of all the residues and 3. the maximum, mean and minimum pathogenicity indexes associated with the genetically accessible mutations (i.e. those implying only one nucleotide change) in each position of the protein. PMUT is freely accessible through a web interface at the Molecular Modeling and Bioinformatics website (http://mmb2.pcb.ub.es:8080/ PMut/). A limited version of PMUT Predictor providing a hot spot analysis is also avail- able as a web service running according to the BioMoby standard (http://www.biomoby.org; http://www.inab.org).

9. PhD-SNP

PhD-SNP (*Predictor of human Deleterious Single Nucleotide Polymorphisms*) [63] is a prediction method based on single-sequence and sequence profile based support vector machines trained on SwissProt variants. The single-sequence SVM (SVM-Sequence) classifies the missense variant to be pathogenic or neutral based on the nature of the substitution and properties of the neighboring sequence environment. The profile-based SVM (SVM-Profile) utilizes sequence profile information taken from MSAs, and classifies the variant according to the

ratio between the frequencies of the wild-type and substituted residue. A decision tree algorithm chooses which one of the two SVMs described above is to be used at each case based on the occurrence of wild- type and mutant amino acids at the given position. The PhD-SNP SVM input is build in three steps: for a given mutation the substitution form the wild-type residue to the mutant is encoded in a 20 elements vector that have -1 in the position relative to the wild-type residue, 1 in the position relative to the mutant residues and 0 in the remaining 18 positions; a second 20 elements vector encoding for the sequence environment is build reporting the occurrence of the residues in a windows of 19 residue around the mutated residue; Both the frequency of the wild type (Fi(WT)) and mutated (Fi(MUT)) residues at position I are evaluated from the sequence profile. The latest version of PhD-SNP uses the same input described for the SVM-Sequence method and 4 more profile based features. The sequence profile is calculated according to the procedure used for the SVM-Profile method but in this case the input vector is composed by the frequencies of wild-type and mutant residues, the number of aligned sequences and the conservation index in the mutated position. The output consists of a table listing the number of the mutated position in the protein sequence, the wild-type residue, the new residue and if the related mutation is predicted as disease-related (Disease) or as neutral polymorphism (Neutral). The RI value (Reliability Index) is evaluated from the output of the support vector machine.

10. SNPs3D

SNPs3D [61], [105] is a web resource and is organized into three modules, each one accessible via a separate simple search window on the user interface. One module generates lists of candidate genes for any specified disease, based on an analysis of the relationship between the disease and genes, as reflected in the literature making use of simple text mining techniques. Concept profiles are constructed for each disease and for each gene. Each concept, a disease or a gene is represented by an ordered list of words and terms most closely associated with the concept. The set of words and terms is complied from the contents of the approximately 80,000 PubMed abstracts that have been manually associated with one or more human genes in the NCBI Entrez Gene database, using natural language processing. The second module provides an interactive graphical gene-gene network, built from literature associations, known protein-protein interactions from BIND (Biomolecular Interaction Network Database), and existing pathways (KEGG). The third module provides information on the relationship between non-synonymous SNPs and protein function. SNP/protein function relationships are derived by two methods ([58], [67], [105]), one using principles of protein structure and stability, the other based on sequence conservation. Access to details of both analyzes is provided through the web interface. Both methods mentioned above make use of a machine-learning algorithm, the support vector machine (SVM), to assign each SNP as deleterious or non-deleterious to protein function. The SVMs are trained on monogenic disease data. Five parameters: probability of accepting that amino acid

substitution, entropy, mean entropy, standard deviation of the entropy and the entropy Z score, were used as features to train a SVM. Bootstrapping, with 30 SVMs for each method, was used to obtain the accuracies and confidence limits. That is, each SVM was trained on data points drawn randomly from the disease and control sets, with the total number of points equal to the size of each set. The training and testing procedure was repeated 30 times. Details of the analysis of each SNP are provided via the user interface. For the profile model, a user can inspect the protein MSA from which the result is derived. For the structure/stability model, feature values (for example, surface accessibility, electrostatic interactions and hydrophobicity) are provided, as well as an interactive molecular graphics interface powered by Jmol, displaying the affected residue in its three dimensional structural context is provided. SNPs3D had a pre-complied candidate genes lists for a set 76 diseases, taken from the NCBI online book, 'Genes and Disease' at the time of its publication.

11. stSNP

Structure SNP [106] is a webserver, which provides the ability to analyze and compare human nsSNP(s) in protein structures, protein complexes and protein–protein interfaces, where nsSNP and structure data on protein complexes are available in PDB, along with the analysis of the metabolic data within a given pathway. StSNP allows users to analyze data using different inputs, by utilizing different search capabilities, by keyword, NCBI protein accession numbers, PDB

IDs and NCBI nsSNP ids quickly retrieve targeted information. StSNP utilizes three major data sources: (1) Protein sequences from NCBI, (2) the reference and nsSNPs locations from NCBI's dbSNP and (3) structures and sequences from the PDB. A pre-calculated list of structural modeling templates found by BLAST has been generated for every protein sequence, and stored in a database for quick retrieval. stSNP enables researchers to map nsSNPs onto protein structures by comparative modeling of structure with nsSNPs using MODELLER and visualize their structural locations by using the multiple structure-sequence viewer. Pathway information is provided from KEGG database. The modeling part of StSNP is interactive and allows the user to choose a template from the list, select particular mutations to be modeled, calculate the model and subsequently visualize the superimposition of the models and template in the Friend software application applet.

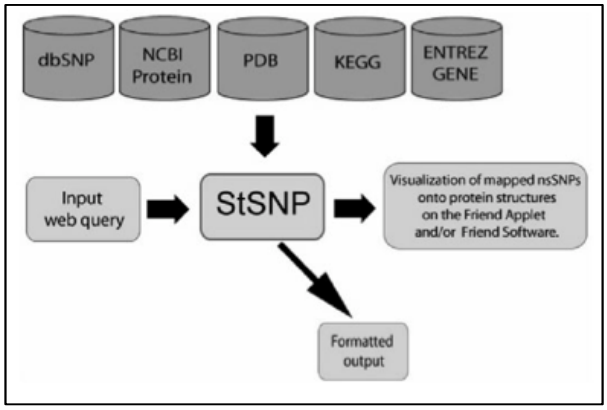


Figure 5: stSNP Workflow

12. PoPMuSiC

PoPMuSiC [107] the first version is an efficient tool for rational computer-aided design of single-site mutations in proteins and peptides. It evaluates the changes in stability of a given protein or peptide under all possible single site mutations, either in the whole sequence or user specified region and returns a list of most stabilizing or destabilizing mutations or of the mutations that do not effect stability. Two types of queries can be submitted. The first option allows to estimate the changes in folding free energy for specific point mutations given by the user. In the second option, all possible point mutations in a given protein or protein region are performed and the most stabilizing or destabilizing mutations, or the neutral mutations with respect to thermodynamic stability, are selected. For each sequence position or secondary structure the deviation from the most stable sequence is evaluated, which helps to identify the most suitable sites for the introduction of mutations. It uses different combinations of database-derived potentials according to the solvent accessibility of the mutated residues. The input for PoPMuSiC is the wild type protein or peptide structure in PDB format. First it computes the effective potentials from a set of known protein structures. It then reads the coordinates of the protein to be mutated from the PDB file, positions the average side chain centroids and computes the backbone torsion angle domains. Then it mutates that position with the 19 other amino acids and evaluates the changes in folding free energy caused by these mutations. These mutations are then classified as a function of smallest folding

free energy changes in absolute value. The output contains the number of mutations that are most destabilizing, most stabilizing or neutral. By default mutations are performed on the whole sequence but the user can limit the specified regions. Figure6, taken from the PoPMuSiC publication [107] shows a schematic description of PoPMuSiC workflow. PoPMuSiC-2.0 [69] is a neural network based tool which has the same basic idea as that of its first version stated above, but has a newly designed energy function. It has a whole new set of 24 statistical potentials, as well as terms modeling the volume changes upon mutation, and express the folding free energy change as a single linear combination of these terms, with weighting coefficients that depend on the solvent accessibility.

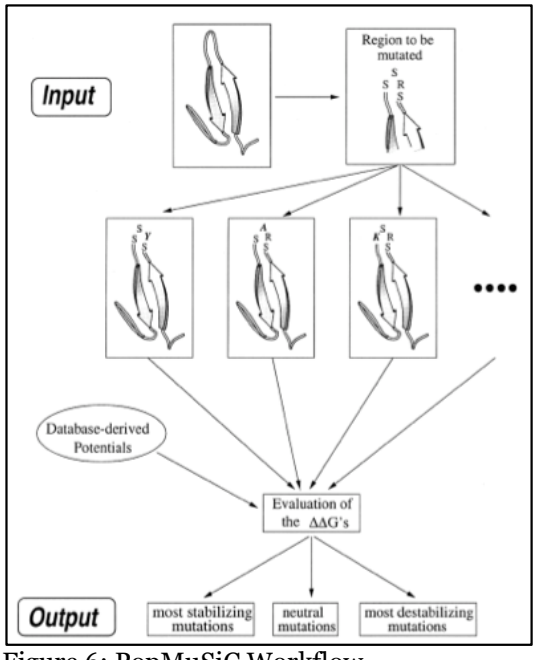


Figure 6: PopMuSiC Workflow

They assume that the weighting coefficients have a sigmoid shape, identified using neural network, to reproduce the smooth transition between the core and the surface of the proteins and to generalize the step functions used in the original version. Mutations whose impact on stability of the protein structure has been measured experimentally were taken from ProTherm database [108]. Different interactions contributing to protein stability are described by a set of statistical potentials, extracted from a database of known protein structures. It uses 24 different potentials, with n ranging from 2 to 7 and grouped in several subsets according to their complexity. They can be divided into two major classes: local and non-local potentials, which describe the correlations between descriptors attached to residues close to each other along the sequence, or close to each other in space, respectively. Another parameter used to predict the mutant stability is the volume difference between the mutant and wild-type amino acids. The estimated stability change upon mutation is expressed as a linear combination of the 26 energy functions and proportionality coefficients. The neural network is trained on a 5-fold validation. The performances are assessed using the root mean square error and the Pearson correlation coefficient between the measured and predicted values of the folding free energy changes.

13. MutPred

MutPred is a Random Forest-based classification method that utilizes several attributes related to protein structure, function, and evolution. It predicts molecular cause of disease/deleterious mutations. MutPred builds upon SIFT and a gain/loss of 14 different structural and functional properties along with PSI-BLAST, transition frequencies and Pfam profiles [89]. For instance, gain of helical propensity or loss of a phosphorylation site. It was trained using the deleterious mutations from the Human Gene Mutation Database [2] and neutral polymorphisms from SwissProt [3]. In MutPred, structural descriptors include prediction of secondary structure and solvent accessibility by the method PHD [100], transmembrane helix prediction by TMHMM [Krogh et al., 2001], coiled-coil structure prediction by MARCOIL [Delorenzi and Speed, 2002], stability prediction by I-Mutant 2.0 [Capriotti et al., 2005], B-factor prediction [Radivojac et al., 2004], and disorder prediction by DisProt [Peng et al., 2006]. Function-related attributes include predictions of DNA-binding residues [Ahmad et al., 2004], catalytic residues, calmodulin-binding targets [Radivojac et al., 2006], and posttranslational modification sites [Daily et al., 2005; Iakoucheva et al., 2004; Radivojac et al., 2010]. A collection of five data sets of human amino acid substitutions were constructed from online databases and the literature. Four of these data sets composed of disease-associated mutations (Cancer, Kinase, HGMD and SPd), whereas the remaining data set contains inherited, putatively neutral polymorphisms. To discriminate between disease-associated mutations and neutral polymorphisms, MutPred applied and

compared support vector machine (SVM) and random forest (RF) classifiers, which were evaluated using per-protein 10-fold cross-validation. Since RFs performed better than the SVMs, further analyses and the predictive model, MutPred, were based on these classifiers. MutPred modeled the loss and gain of each structural and functional property directly via posterior probabilities, thereby directly enabling estimation of the contribution of a gain/loss of a given property in order to deduce the underlying mechanism of disease. The output of MutPred contains a general score (g), i.e., the probability that the amino acid substitution is deleterious/disease-associated, and top 5 property scores (p), where p is the P-value that certain structural and functional properties are impacted. Scores with $g > 0.5$ and $p < 0.05$ are referred to as actionable hypotheses; Scores with $g > 0.75$ and $p < 0.05$ are referred to as confident hypotheses; Scores with $g > 0.75$ and $p < 0.01$ are referred to as very confident hypotheses.

Current version of MutPred, at the time of writing this dissertation, 1.2, has some updates which include replacing SIFT score by a more stable version of code that calculates evolutionary conservation, the I-mutant software replaced by a more stable MUpro [59], by the Baldi group and the training data set updated to contain 39,218 disease-associated mutations from HGMD and 26,439 putatively neutral substitutions from Swiss-Prot.

14. FastSNP

FastSNP [109] is a web-based application, which prioritizes SNPs according to 12 phenotypic risks and putative functional effects, such as changes to the transcriptional level, pre-mRNA splicing, protein structure, etc. It extends, with recent findings and a decision tree, the strategy of Tabor *et al.* who studied the functional effects of polymorphisms and presented a prioritization strategy that associates the relative risk of a SNP with its location and the type of sequence variants. FastSNP uses a decision tree, stated to be complete in the sense that it considers all known functional roles of a SNP in a gene, to assess the risk of a SNP into 1 of 13 types of the functional effects, each of which is assigned a risk ranking number between 0 and 5. A high risk rank implies a high-risk level. It uses eight services that provide databases and analytical tools to predict functional effects for SNP prioritization. dbSNP [110] provides the location of a SNP in a gene and its alleles, allele frequency and context sequence, Ensembl provides a cross-reference/alternative data source for dbSNP, TFSearch [111] predicts if a non-coding SNP alters the transcription factor-binding site of a gene, PolyPhen [53] predicts if a non-synonymous SNP alters an amino acid in a protein resulting in structural changes (damaged or benign) in a protein, ESEfinder [112] predicts if a synonymous SNP is located in a exonic splicing enhancer motif, which would diminish the motif with a different allele, Rescue ESE [113] provides a cross-reference/alternative data source for ESEfinder, FAS-ESS [114] predicts exonic splicing silencer for each SNP allele and SwissProt provides the information about protein domains to

determine if a SNP causes an alternative splicing that leads to a protein domain being abolished. UCSC Golden Path [115] and NCBI Blast [116] are two services for quality control of candidate SNPs and haplotype database from HapMap [117] is used for further reducing the number of candidate genes for genotyping. A unique feature of FASTSNP is that the prediction of functional effects is always based on the most up-to-date information, which FASTSNP extracts from the above mentioned 11 external web servers at query time using a team of re-configurable web wrapper agents. These extendable web wrapper agents automate web browsing and data extraction and can be easily configured maintained and extended with a tool that uses a machine-learning algorithm. The input format has three different methods, 'Query by Candidate Gene.' where user can choose to specify a gene symbol, SNP reference cluster ID (rsid), or a chromosome position as the query. User can select the transcripts of the queried gene, if SNPs are coding or non-coding and the allele frequency. Once a final set of candidate SNPs is selected, FASTSNP performs the SNP prioritization and return the prioritization results in a risk ranking report, and provide a function report for each candidate SNP. The second method is 'Query by SNP' which allows the user to specify a single SNP rsID or upload an excel file containing their entire candidate SNPs for prioritization. In the third method it accepts novel SNP sequences along with the position and the substitution, as input. The function analysis module consists of three agent pipe-lines corresponding to decision paths in the decision tree. The first pipeline is for non-coding SNPs. The

input sequence pair will be sent to TFSearch to obtain the predicted transcription factor-binding sites. The second pipeline handles non-synonymous SNPs. In this pipeline, the agent queries PolyPhen to obtain its prediction on whether the SNP will alter an amino acid in a protein and result in structural changes (damaged or benign) in the protein. The third agent pipeline obtains information to predict if the alternative splicing caused by a synonymous SNP may abolish a protein domain. FASTSNP performs the necessary post-processing for the data returned from the agent pipelines and submits the results to the prioritization module, which then classifies the SNP, assigns it a risk ranking according to the decision tree, and compiles the results into a function report. The function report on a SNP contains seven sections on the SNP's functional effects, namely (i) genomic information, presents the nearby sequence, the alleles and the allele frequency among different ethnic groups; (ii) functional effects summary, presents the risk assessment; (iii) transcription regulatory, shows the predicted transcription factor binding sites generated or disrupted by the different SNP alleles; (iv) alter- native splicing regulatory, reports exonic splicing enhancer/ silencer motifs changed by the SNP alleles leading to exon skipping or inclusion; (v) mRNA/protein domain effects, presents all spliced forms of mRNAs and protein variants extracted from GenBank. The protein domains that the SNP locates in are highlighted; (vi) protein structure effects, reports whether the SNP may cause a significant structural change in a protein; and (vii) SwissProt feature table, provides

information regarding other known mutations or variations of the translated protein of mRNAs related to the SNP. Some of these sections are specific to coding or non-coding SNPs and they will appear or not appear in the function report accordingly. Figure7 taken from the FastSNP publication [109] shows data flow of FastSNP.

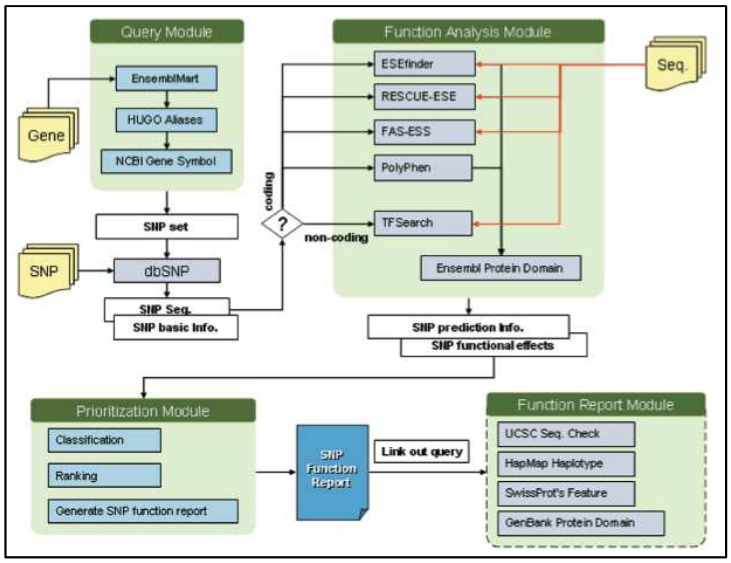


Figure 7: FastSNP Data Flow

15. Bongo

Bongo, Bonds ON Graph [118] is based on graph theoretic measures that analyze the likelihood of a missense mutation to cause diseases by affecting its corresponding protein structures. It considers a target protein as a residue–residue interaction graph in which vertices represent residues and edges represent interactions between residues, and applies graph theoretic measures

to estimate the topological change due to single point mutations. The novelty lies in the application of a graph theory concept, vertex cover, by which key residues are identified for analyzing structural effects of single point mutations. For a target mutation, Bongo identifies two sets of key residues from the residue interaction network of its corresponding wild-type and mutant protein structure. Then, Bongo quantifies the structural effect of the mutation via comparing the difference of the two key residue sets. Bongo derives the interaction graph of a protein by considering each residue as a vertex and each residue-residue interaction, including hydrogen bonds, π–π, π–cation, and hydrophobic interactions, as an edge. The weight on each edge differs according to the total number of cross-secondary structure interactions as well as number of interactions with individual residues. The interactions are then normalized between the two secondary structures by dividing the weight with the total number of cross-secondary structure interactions. Based on the weighting scheme, Bongo defines the key residues as the minimum weighted vertex cover, which represents the minimum necessary residues to establish the interaction network. Then it uses a selection scheme, which adopts an approximation algorithm based on the greedy principle to identify the key residues. Bongo first uses Andante to model the mutant-type protein structure by rearranging the side chain around the mutation site. The structural effects of a mutation are then analyzed by comparing the wild-type and mutant-type key residues, denoted as Kwt and Kmt, respectively. If a key residue in Kwt is not found in Kmt, then it is

considered to be affected by the mutation. Consequently the overall impact of a mutation is calculated according to the key residues affected by the mutation. On deriving the impact value, Bongo considers mutations with I>1 to cause structural effects, which is the criterion calibrated over mutations in the p53 core domain. Bongo has been calibrated using experimental data on the tumor suppressor p53 core domain. Figure8, taken from its publication [118] shows the Bongo work flow.

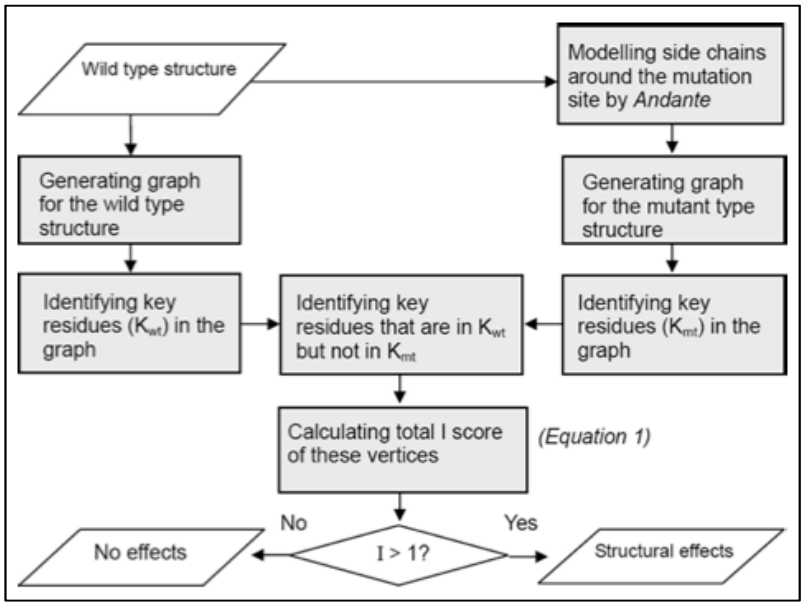


Figure 8: Bongo Work Flow

While identifying disease-associated nsSNPs, it is stated that Bongo yields similar accuracy as PolyPhen with a PPV (positive predictive value) of 78.5% to 77.2% and a NPV (negative predictive value) of 34.5% and 37.6% respectively.

16.Panther□

Panther, Protein ANalysis THrough Evolutionary Relationships [119][120] is a comprehensive software system for inferring the functions of genes based on their evolutionary relationships. It is a database of phylogenetic trees of protein-coding gene families from all kingdoms of life. Phylogenetic trees of gene families form the basis for PANTHER and these trees are annotated with ontology terms describing the evolution of gene function from ancestral to modern day genes. PANTHER is composed of two main components: the PANTHER library (PANTHER/LIB) and the PANTHER index (PANTHER/X). PANTHER/LIB is a collection of "books," each representing a protein family as a multiple sequence alignment, a Hidden Markov Model (HMM), and a family tree. Hidden Markov models (HMMs) are constructed for all families and subfamilies, which can be used for genome annotation projects. The PANTHER/LIB HMMs are used as a statistical method for scoring the "functional likelihood" of different amino acid substitutions on a wide variety of proteins. Panther predictions have a score based on an alignment of evolutionarily related proteins. Estimates are incorporated with the development of the Substitution Position-Specific Evolutionary Conservation (subPSEC) scores utilizing more sophisticated alignments based on hidden Markov models from protein families, in the PANTHER database. The primary mission of the PANTHER database is to organize genes into families and subfamilies and to classify them according to inferred function. Much of the

organization achieved by this database relies on making PMSAs across a large number of gene subfamilies and families. The alignments are obtained from the PANTHER library of protein families based on Hidden Markov Models (HMMs). The subPSEC score describes the amino acid probabilities, in particular, positions among evolutionarily related sequences, and the values range from 0 (neutral) to about -10 (most likely to be deleterious). The cutoff for classifying a missense variant to be pathogenic can be defined by the user, but it is advised to use a cutoff of -3 for classification. One important limitation, however, is that PANTHER's PMSAs generally cover only the most conserved portions of genes, limiting the fraction of missense substitutions to which it can be applied.

17. LS-SNP

LS-SNP [62] is a genomic-scale, computational pipeline that comprehensively maps human nsSNPs in NCBI's dbSNP database onto protein sequences in the SwissProt/TrEMBL databases, functional pathways and comparative protein structure models, and predicts positions where nsSNPs destabilize proteins, interfere with the formation of domain-domain interfaces, have an effect on protein-ligand binding or severely impact human health. The automated computational pipeline consists of three modules: In the first module, it extract the genomic locations of human SNPs from dbSNP and maps these SNPs onto human protein sequences in SwissProt/TrEMBL to identify the SNPs that result in an amino acid residue substitution. The primary output of

the SNP-to-protein mapping module is a list of protein sequences from SwissProt/TrEMBL and the positions of all amino acid residue substitutions produced by the SNPs found in dbSNP. In the second module, each of the SwissProt/TrEMBL protein sequences are input into MODPIPE, an automated system for comparative protein structure modeling, Sequence–structure matches are identified by aligning the PSI-BLAST profile of each sequence (built with 10 iterations and E-value cutoff 0.0001) against a library of candidate template sequences extracted from PDB and by scanning the sequence against a database of template profiles with IMPALA (Schaffer *et al.,* 1999). Each significant alignment (E-value cutoff 0.0001) that covers distinct regions of the target sequence is chosen for modeling. Models are calculated for each of the sequence–structure matches using the default 'model' routine of MODELLER (Sali and Blundell, 1993). A statistical scoring function is used to assess each model (Melo *et al.,* 2002). The output of the sequence-to-structure module is a collection of fold assignments, alignments of target sequences and template structures, comparative structure models for SwissProt/TrEMBL sequences and mutated sequences, and model assessments. In the third module, the output of the first two modules is used to help compute a variety of annotations for human nsSNPs. The nsSNPs are annotated with respect to genomic sequence, protein sequence, protein structure and function to identify nsSNPs that generally impact human health and specifically nsSNPs that interfere with the formation of domain–domain interfaces or have an effect on protein–ligand binding.

Finally, it combines a rule-based approach to identify putatively destabilizing nsSNPs and a supervised machine learning approach to identify nsSNPs likely to have an impact on human health. To identify destabilizing effects on protein structure, four structural rules are applied that are based on the preferences of each amino acid residue type to be in any of the secondary structure and solvent accessibility states. DSSP program [88] is used to compute secondary structure state and solvent accessible surface area at each position. Destabilization is predicted when Relative solvent accessibility, RSA is <25% and difference in accessible surface propensities is >0.75; (2) RSA is >50% and difference in accessible surface propensities is >2; (3) RSA is <25% and formal charge change (histidine is assigned a +1 charge); (4) the variant involves a proline in a helix. Interference with domain–domain interface formation or protein–ligand binding is predicted when any of the four conditions listed above occur at a putative domain–domain interface or ligand binding site. To find such nsSNPs, template residues at domain–domain interfaces and in proximity to small molecule ligands are identified using PIBASE and the LIGBASE table (Stuart *et al.,* 2002) of MODBASE, respectively. A template residue is considered to be at an interface if it is within 6 Å of an atom in an adjacent domain. It is considered to be ligand binding if it is within 5 Å of a HETATM (i.e. an atom not covalently bonded to the protein, not in one of the standard 20 residue types, nor in a water molecule) in the PDB structure. Figure9 taken from LS-SNP publication [62], shows the LS-SNP computational pipeline.

A supervised machine learning approach is applied next, that combines information from multiple sources: amino acid residue side chain properties, comparative structure models of the SwissProt/TrEMBL sequences and mutated sequences and evolutionary properties extracted from MSAs.

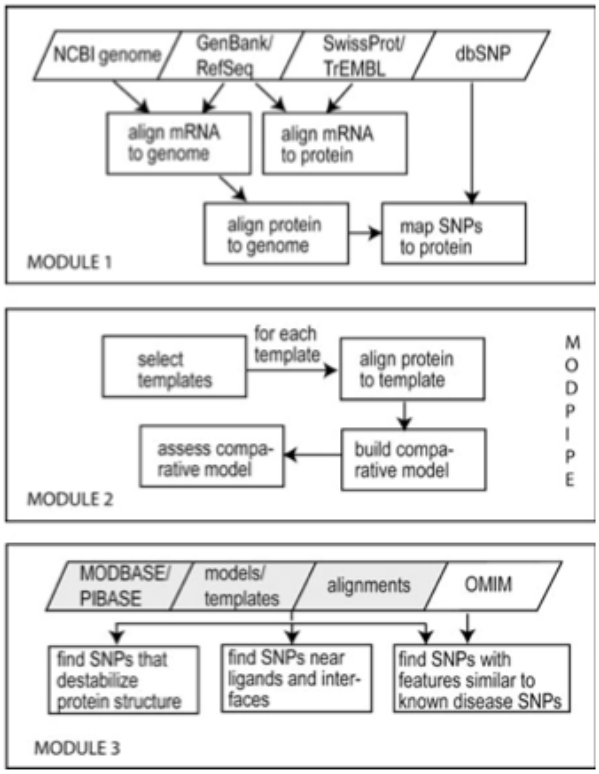


Figure 9: LS-SNP Work Flow

To compute evolutionary properties based on amino acid residue conservation and substitution likelihoods, an MSA for each SwissProt/TrEMBL protein sequence is constructed via iterative search of NCBI's nr database using the SAM-T2K algorithm. For each nsSNP, 13 features are computed using a variety

of programs, including MODELLER, MODPIPE (Sanchez and Sali, 1998), DSSP (Kabsch and Sander, 1983), and SAM (Krogh *et al.,* 1994). Two new features that measure strain when a mutant side chain is introduced into the native sequence are added; the strain is quantified by the number of violated spatial restraints used in the construction of the mutant model. A support vector machine (SVM) was trained and evaluated using a 3-fold cross-validation protocol. The SVM classifies each example with a discriminant score where negative scores predict disease association while positive scores predict a neutral or positive nsSNP. The absolute value of the score provides a confidence measure for the prediction.

18.topoSNP

topoSNP [121][122], topographic mapping of Single Nucleotide Polymorphism is a database that provides an online resource for analyzing non-synonymous SNPs derived from the Online Mendelian Inheritance in Man (OMIM) database and other nsSNPs derived from dbSNP, and be mapped onto known 3D structures of proteins. The web interface produces an interactive visualization of disease and non-disease associated non-synonymous single nucleotide polymorphisms (nsSNPs) and displays geometric and relative entropy of SNPs calculated from multiple sequence alignment as obtained from the Pfam database as well as manually adjusted multiple alignments obtained from ClustalW. TopoSNP classifies each nsSNP site into three categories based on their geometric location: those located in a surface pocket or an interior void of the protein, those on a convex region or a shallow depressed region, and those

that are completely buried in the interior of the protein structure. It attempts to gain an insight by correlating geometric locations of disease SNPs and the degree of their conservation in the protein family. The geometric sites are determined using the alpha shape theory, which is based on a weighted Delaunay tessellation scheme from which topological and metric properties of the molecular shapes are extracted. Once classified, SNPs are mapped to known SNP sequences using a hidden Markov Model to determine whether or not the mutation occurs on a conserved or more variable residue. It has been found that disease-associated nsSNPs found in the interior of proteins are more likely to be conserved and that nsSNPs not in the interior have no strong tendency to occur at a conserved or non-conserved residue. It was also found that compared to control nsSNPs, 88% of disease-associated nsSNPs (derived from the online mendelian inheritance in man (OMIM) database) are more likely to be located in well-formed surface pocket or void locations. In an attempt to overcome the fact that relatively few alleles can be mapped to 3-D protein structures, a bootstrap method was used to calculate 95% confidence intervals.

Integrated Predictive Methods

1. Condel

Condel, CONsensus DELeteriousness score [123] of missense mutations, is a weighted average of the normalized scores from multiple methods. The idea behind it is to integrate the output of computational tools aimed at assessing the impact of non synonymous SNVs on protein function. To do this, it computes a weighted average of the scores (WAS) of these tools. The scores of different methods are weighted using the complementary cumulative distributions produced by the five methods on a dataset of approximately 20000 missense SNPs, both deleterious and neutral. The probability that a predicted deleterious mutation is not a false positive of the method and the probability that a predicted neutral mutation is not a false negative are employed as weights.

The original idea for developing Condel was to integrate into a unified classification, the outputs of five tools: SIFT, Polyphen2, MAPP, LogR Pfam E-value and MutationAssessor. First, the five methods are used to score and classify HumVar, a comprehensive dataset of deleterious and neutral mutations. Then, the outputs of the five methods are combined in four different ways. It was found that a weighted average of the normalized scores (WAS) of the five methods outperforms each individual tool—and the other three combining operations assayed—in the task of classifying SNVs as deleterious or neutral. The process of integrating the scores of individual methods in the WAS uses the probabilities of the complementary cumulative distribution of scores produced

by each method to compute their weights. The score of each method is thus penalized in an inverse manner with respect to its confidence. Subsequently, it infers both a consensus prediction and a score. In order to operate with comparable scores, the internal scores of MAPP, Logre, and MutationAssessor were normalized to values between 0 and 1 and the complement of the SIFT probability was taken as the normalized score of this tool.

Four datasets were obtained from different sources: HumVar, HumDiv and two datasets containing only deleterious mutations. First, two datasets, HumVar and HumDiv, obtained from the website of PPH2, composed of positive and negative examples, were used to run five programs aimed at separating deleterious from neutral missense SNVs and assessing the performance of different ways to integrate their outputs. The other two datasets containing only deleterious mutations, were used to assess whether the WAS correlated with the recurrence of cancer mutations and with the degree of impairment of the biological activity caused by the mutations. All the five tools were run on the four datasets and their outputs were integrated. After running the five tools on HumVar and HumDiv, the complementary cumulative distributions of the scores of deleterious and neutral mutations produced by each tool were constructed. The corresponding receiver operator characteristic (ROC) curves were also built. The dependence of their accuracy with respect to their sensitivity was assayed to determine the optimal cutoff of each tool. A cutoff the score that produced the sensitivity yielding the maximum accuracy at classifying each dataset was

selected. Four different ways to integrate the outputs of the tools in these two data- sets were assayed: a simple vote score (SVS), a simple average score (SAS), a weighted vote score (WVS), and a weighted average score (WAS). Integration was pursued at two levels: classification and score. At the first level, the classifications of different methods were integrated by using both an SVS and a WVS. At the second, the internal scores calculated by each method to achieve a classification were combined through an SAS and a WAS. For a predicted deleterious mutation, the weight—the probability that it is not a false positive— increases with the score, thus inflicting a higher penalty on scores that are closer to the cutoff and lower costs to scores closer to the tail of the complementary cumulative distribution of true neutral mutations. For a mutation predicted to be neutral, the lower the score, the smaller the weight—the probability that it is not a false negative— and thus, the higher the penalization. Condel scripts can be downloaded and run locally. Condel scores can be derived for a limited set of specified mutations via the corresponding web application. The Ensembl database provides position-specific Condel predictions that combine SIFT and Polyphen-2 for every possible amino acid substitution in all human proteins.

2. Carol

CAROL, Combined Annotation scoRing tool, [124] is a combined functional annotation score of non-synonymous coding variants which combines information from 2 predictive tools: PolyPhen-2 and SIFT, in order to improve

the prediction of the effect of non-synonymous coding variants. In this scoring method, a weighted Z method that combines the probabilistic scores of PolyPhen-2 and SIFT were used. Two dataset pairs, positive (known disease-causing) and negative (postulated non-disease-causing) were used to train and test CAROL using information from the dbSNP: 'HGMD-PUBLIC' and 1000 Genomes Project databases. To compare PolyPhen2.0 and SIFT scores, the probability of the complement of the SIFT scores was calculated. The scaled scores range between 0 and 1, in which scores closer to 1 indicate that the amino acid substitution is deleterious, and scores closer to 0 that it is neutral. The CAROL algorithm is based on a weighted Z method, which combines the probabilistic score for each annotation tool. For investigative purposes PANTHER and Genomic Evolutionary Rate Profiling (GERP) were incorporated into the functional annotation tool, but it is stated that PolyPhen-2 and SIFT produced the most robust combination. The authors state that CAROL has higher predictive power and accuracy for the effect of non-synonymous variants than each individual annotation tool (PolyPhen-2 and SIFT) and benefits from higher coverage.

3. dbNSFP

dbNSFP [125] [126] is a database developed for functional prediction and annotation of all potential non-synonymous single-nucleotide variants (nsSNVs) in the human genome. Its current version, ver 2.0 is based on the Gencode

release 9 / Ensembl version 64 and includes a total of 87,347,043 nsSNVs and 2,270,742 essential splice site SNVs. In its first version, the genes and their corresponding codons were determined based on CCDS version 20090327, latest version based on the human reference sequence build hg18. It compiles prediction scores from six prediction algorithms, SIFT, Polyphen2, LRT, MutationTaster, MutationAssessor and FATHMM, three conservation scores, PhyloP, GERP++ and SiPhy and other related information including allele frequencies observed in the 1000 Genomes Project phase 1 data and the NHLBI Exome Sequencing Project, various gene IDs from different databases, functional descriptions of genes, gene expression and gene interaction information, etc. The database is separated into two parts, dbNSFP_variant and dbNSFP_gene. The former focuses on variant annotations including prediction scores and conservation scores, and the latter focuses on gene annotations. As to variant annotation, the database has expanded its SNV collections not only based on a more up-to-date GENCODE 9 annotation but also included all potential essential splice site SNVs (ssSNVs), which are another type of candidate variants in exome sequencing studies. To facilitate filtering common SNVs observed in human populations, allele frequencies from the 1000 Genomes Project phase 1 data (Abecasis et al. 2012) and the NHLBI Exome Sequencing Project data (Fu et al. 2013) were also added. Figure10 taken from dbNSFP publication shows summary of functional prediction scores and conservation scores.

Score	Training data	Information used	Prediction model
PolyPhen-2	UniProtKB/UniRef100; PDB/DSSP; UCSC alignments of 45 vertebrate genomes	eight sequence-based and three structure-based predictive features	naive Bayes classifier
SIFT	SWISS-PROT/TrEMBL	sequence homology based on PSI-BLAST	position specific scoring matrix
Mutation Taster	UniProt; homologous genes in humans and 10 other species; dbSNP; HapMap	conservation, splice site, mRNA features, protein features;	naive Bayes classifier
LRT	coding sequences of 32 vertebrate species	sequence homology	likelihood ratio test of codon neutrality
Mutation Assessor	homologous sequences from Uniprot identified by BLAST	sequence homology of protein families and sub-families within and between species	combinatorial entropy formalism
FATHMM	homologous sequences from UniRef90, SUPERFAMILY and Pfam	sequence homology	hidden Markov models
SiPhy	genomes of 29 mammals	multiple alignments	inferring nucleotide substitution pattern per site
GERP++	genomes of 34 mammals	multiple alignments and phylogenetic tree	maximum likelihood evolutionary rate estimation
PhyloP	genomes of 33 placental mammals	multiple alignments and phylogenetic tree	distributions of the number of substitutions based on phylogenetic hidden Markov model

Figure 10: dbNSFP summary of functional prediction scores and conservation scores

In the original version, the PhyloP [127] scores were extracted from the placental subset of the precomputed phyloP44way scores [Pollard et al., 2010] provided by the UCSC Genome Browser. Original SIFT scores were got from ANNOVAR [128], which were originally from a local database format of SIFT 4.0.3. Original LRT scores (LRTori) were downloaded from the LRT Webserver. Polyphen2 scores were manually queried and downloaded as ~500 batches from its batch query server with default query settings. MutationTaster scores were queried from its Webserver (http:// www.mutationtaster.org/) using its batch query Perl scripts. Each nsSNP had links to, chromosome number, physical position on the chromosome as to hg18 (1-based coordinate), reference nucleotide allele (as on

the 1 strand), alternative nucleotide allele (as on the 1 strand), reference AA, alternative AA, physical position on the chromo- some as to hg19 (1-based coordinate), gene name, gene Entrez ID, CCDS ID, reference codon, position on the codon (1, 2, or 3), degenerate type (0, 2, or 3), AA position as to the protein, coding sequence (CDS) strand (+ or -), estimated nonsynonymous-to-synonymous-rate ratio (o, reported by LRT), PhyloP score, PhyloP prediction, SIFT score, SIFT prediction, Polyphen2 score, Polyphen2 prediction, LRT score, LRT prediction, MutationTaster score, MutationTaster prediction. The Spearman's rank correlation coefficients (RCCs) and the Pearson's correlation coefficients were calculated for each pair of the methods. The program BPCAfill was used to impute the missing scores in dbNSFP. A companion search program is provided to search for a nsSNP a chromosome position or a gene. By default, the program searches all chromosomes with the positions according to the human genome reference sequence hg18. The search program now supports vcf format for the input file. Users can specify both the chromosomes to search for and the reference sequence version. Some dbNSFP contents can also be accessed through variant tools, ANNOVAR, KGGSeq, UCSC Genome Browser's Variant Annotation Integrator, Ensembl Variant Effect Predictor and HGMD.

4. F-SNP

F-SNP [129], [130] is a database which provides integrated information about the functional effects of SNPs obtained from 16 bioinformatics tools and

databases. The functional effects are predicted and indicated at the splicing, transcriptional, translational, and post-translational level. As such, the F-SNP database helps identify and focus on SNPs with potential pathological effect to human health. Each SNP is examined for deleterious effects with respect to each functional category (i.e., protein coding, splicing regulation, transcriptional regulation, and post-translation – as shown in the top part of the figure). Another distinguishing feature of the F-SNP database is its integration of human-disease databases to facilitate identification of potential disease-causing SNPs as genetic markers in association studies. The F-SNP database provides a web interface that takes as input either a disease, a gene, a genomic region or a SNP identifier.

The sources of the dataset of human SNPs and their annotations are the dbSNP (build 126) ,NCBI Entrez Gene and Ensembl (release 42) databases. To link SNPs with specific genes, for each gene, SNPs located along the gene region (including 5 kb upstream and 5 kb downstream) were identified. To link candidate genes with the 85 diseases the dataset of a gene-disease map from NCBI's OMIM database was downloaded.

For each category a series of tests is executed to determine whether the SNP has a functional impact. First the type (coding, intronic etc.) of the genomic region is identified, using data from dbSNP and Ensembl. Once this is determined, other tests are performed, TFSearch (ver. 1.3) and Consite are used to identify transcriptional regulatory SNPs in promoter regions; The Ensembl

(release 42) and GoldenPath databases are used to identify SNPs in other transcriptional regulatory regions (e.g. microRNA, cpgIslands); KinasePhos, OGPET (ver. 1.0) and Sulfinator are used to examine post-translation modification sites. In addition, genomic regions that are conserved across multiple species are identified using GoldenPath. To assess if a SNP has a deleterious effect on protein coding, it first must be located on a coding region. Ensembl is used to examine if this is a Nonsense mutation, in which case the SNP is considered to be deleterious. Otherwise – if the SNP is a Missense mutation, it is further tested by five different tools (PolyPhen, SIFT, SNPeffect, SNPs3D, LS-SNP to check if the non-synonymous substitution is deleterious. Figure11 taken from F-SNP publication shows a list of all the bioinformatics tools and databases integrated in it.

Functional category	Tool	URL
Protein coding	PolyPhen (6)	http://genetics.bwh.harvard.edu/pph/data/index.html
	SIFT (7)	http://blocks.fhcrc.org/sift/SIFT.html
	SNPeffect (8)	http://snpeffect.vib.be/index.php
	SNPs3D (9)	http://www.snps3d.org/modules.php?name=SNPtargets
	LS-SNP (10)	http://alto.compbio.ucsf.edu/LS-SNP/Queries.html
	Ensembl (4)	http://www.ensembl.org/index.html
Splicing regulation	ESEfinder (11)	http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi
	RescueESE (12)	http://genes.mit.edu/burgelab/rescue-ese/
	ESRSearch (13)	http://ast.bioinfo.tau.ac.il/
	PESX (14)	http://cubweb.biology.columbia.edu/pesx/
	Ensembl (4)	http://www.ensembl.org/index.html
Transcriptional regulation	TFSearch (15)	http://www.cbrc.jp/research/db/TFSEARCH.html
	Consite (16)	http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite/
	GoldenPath (17)	http://genome.ucsc.edu/
	Ensembl (4)	http://www.ensembl.org/index.html
Post-translation	KinasePhos (18)	http://kinasephos.mbc.nctu.edu.tw/
	OGPET (19)	http://ogpet.utep.edu/
	Sulfinator (20)	http://www.expasy.ch/tools/sulfinator/
Conserved region	GoldenPath (17)	http://genome.ucsc.edu/

Figure 11: Bioinformatics tools and databases integrated into F-SNP

A majority vote between these tools concludes the process, and identifies the SNP as either having a potentially deleterious functional impact (denoted functional in the figure) or not. Figure12 shows the Decision procedure for functional SNP assessment in F_SNP.

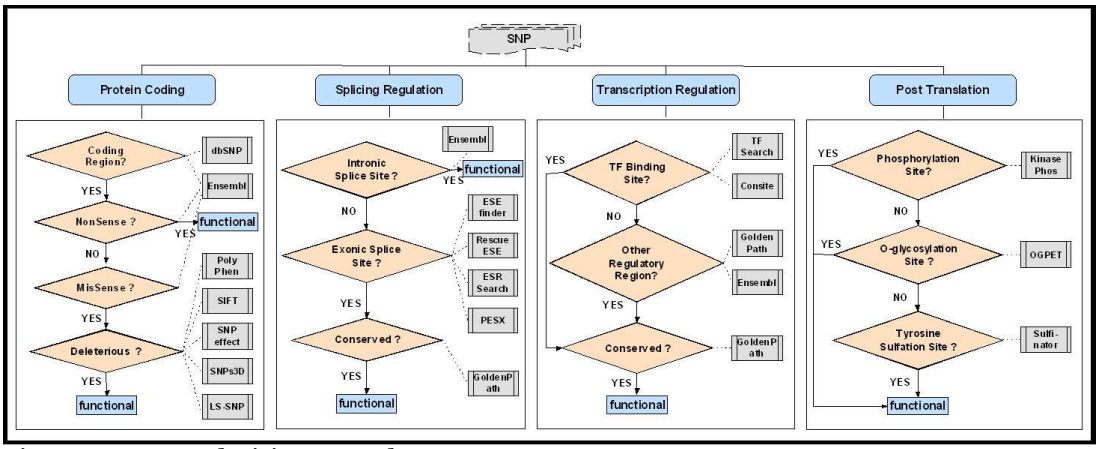


Figure 12: F-SNP decision procedure

Cancer Specific Prediction tools

1. Mutation Assessor

Mutation Assessor [31] is a computational protocol based on an elaborate conservation-based approach. The use of evolutionary information in Mutation Assessor differs from other sequence-based predictors. It distinguishes between conservation patterns within aligned families (conservation score) and sub-families (specificity score) of homologs and so attempts to account for functional shifts between subfamilies of proteins. The novelty of the approach, as stated by

the authors, is in exploiting the evolutionary conservation in protein subfamilies, which are determined by clustering multiple sequence alignments of homologous sequences on the background of conservation of overall function. Given a mutated protein name and a mutated residue position, the computational protocol searches for sequence homologs, builds a multiple sequence alignment, clusters sequences into subfamilies and scores a mutation by global and sub-family specific conservation patterns. Mutations affecting either type of conserved residue are likely to be functional. Based on the assumptions that evolutionarily unfavorable residues are not observed or observed less frequently than neutral or critically important residues, while critically important residues are conserved in diverse evolutionary settings and that the distribution of residues in any (aligned) sequence position of a protein family can be treated independently of other positions, the protocol uses the entropy of the residue distribution in an alignment column as a measure of residue conservation and estimates the mutation impact, named the conservation score, using the difference of the entropy caused by the mutation. To refine the assessment of conservation patterns, patterns of a subtler type are considered, in which the evolutionary constraint on a residue type in a particular position is not constant in the entire family, but only appears to operate in a protein subfamily. A combinatorial entropy approach id used to quantify subfamily conservation patterns which simultaneously determines protein subfamilies, by clustering, and residues, called specificity residues, which

characteristically differ between these subfamilies. Specificity residues are conserved within a subfamily but differ between subfamilies presumably encoding functional diversity. Interestingly, specificity residues were found to be predominantly located in binding interfaces on the protein surface implicating them in protein interaction [131].For example, a D125N mutation in CDKN2A (cyclin-dependent kinase inhibitor 2A) from liver cancer is scored as deleterious by Mutation Assessor, because this residue is absolutely conserved as D in mammalian homologs, but is scored as neutral by other methods that include more distant homologs, such as those of fishes, where the wild-type residue is N. The co-crystal of CDKN2A with cyclin-dependent kinase-6 (CDK6) shows that D125 is at the binding interface of the two proteins, close to Serine 155 (4.9A) of CDK6. Loss of this negative charge in the D125N mutant may substantially alter the binding affinity and so promote tumorigenesis [40]. Figure12 taken from Mutation Assessor publication [31] shows the work flow.

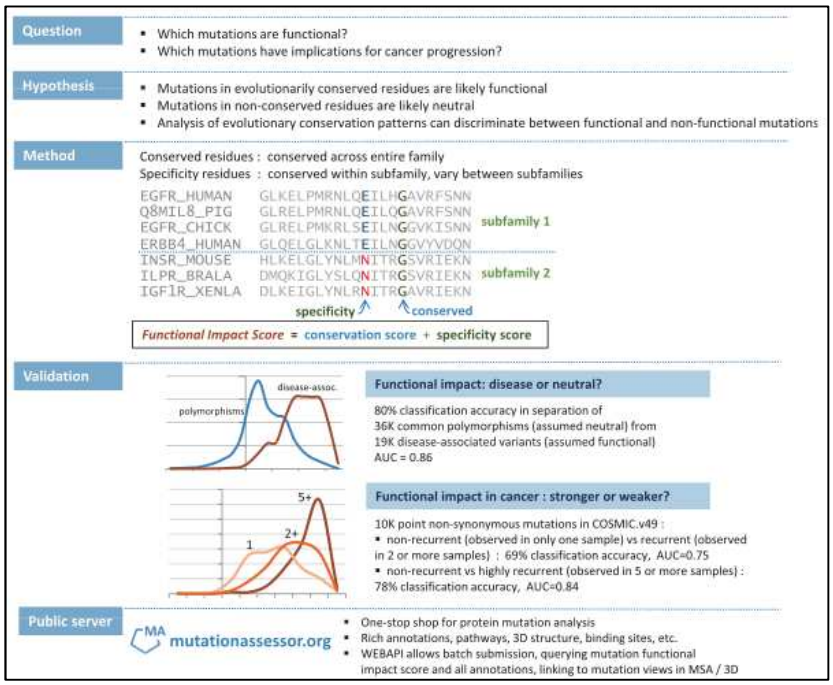


Figure 13: MutationAssessor Work Flow

The clustering algorithm groups the sequences of a protein family alignment into distinct subfamilies, so as to minimize the sequence diversity within subfamilies and to maximize the overall difference between subfamilies at a select number of 'specificity' positions. To quantify the entropy difference resulting from a mutation that affects conserved residue patterns in protein subfamilies, it (i) determines distinct sequence subfamilies from protein family alignments and (ii) computes a specificity conservation score in analogy to the family conservation score mentioned above. The conservation and the specificity scores are then averaged to get the combined score of the functional impact, functional impact score (FIS) that gave a higher prediction accuracy in the validation tests, as assessed in the context of evolutionary patterns in a multiple

sequence family alignment. The validation of FIS was done on experimentally tested TP53 mutations and found that the functional impact score is correlated with experimentally measured functional impact of mutations and the score is higher for mutations that result in 'loss of function' and in 'gain of function' of TP53. The scoring function was validated by separation of a large set of disease-associated variants from common (benign) polymorphisms with the accuracy of 79% and the area under the curve (AUC) 0.86 in the receiver-operation-characteristic (ROC) analysis for a two-class distinction. The predictive power of the score for cancer mutations from COSMIC database was assessed by separating assumed to be driver mutations (1800 recurrent mutations - observed in two or more samples and 700 highly recurrent mutations - observed in 5 or more samples) from assumed to be passenger mutations (8200 single mutations - observed only in one sample). The maximal separation accuracies are 69% (AUC 0.75) for recurrent vs. single and 78% (AUC 0.84) for highly recurrent vs. single. Mutations in multiply mutated genes and mutations in known cancer genes tend to have significantly higher functional impact scores than control sets.

2. CanPredict

CanPredict [132] [42] is a web application built using the Gene Ontology and data from the SIFT [43] and LogR.E-value metric [133], that predicts cancer-associated missense mutations with a very low false-positive rate. It allows users

to determine if particular mutations are likely to be cancer-associated. As an input a single full-length RefSeq protein sequence or accession and multiple associated mutations can be submitted. The impact of each mutation is measured using two known methods: Sorting Intolerant From Tolerant (SIFT) and the Pfam-based LogR.E-value metric. The SIFT algorithm uses similarity between closely related proteins to identify potentially deleterious changes. SIFT scores <0.05 are predicted to be deleterious and only SIFT scores with a median information content score <3.25 are included for predictions since higher values likely indicate unreliable SIFT scores. The Pfam-based logR.E-value score, derived from values provided by the HMMER 2.3.2 software, predicts whether a mutation will alter protein function by determining the difference in fit of a wild-type version of the protein to a particular Pfam model. A third method where the log-odds scores, the Gene Ontology Similarity Score (GOSS) [42], are calculated to represent the relative frequency with which a Gene Ontology (GO) term was used to annotate cancer or non-cancer gene sets. Scores from these three algorithms are analyzed by a random forest classifier, which then predicts whether a change is likely to be cancer-associated. The training data set used to construct the classifier, downloaded from COSMIC database, composed of 200 randomly selected known somatic cancer mutations and 800 non-cancer, non-synonymous variants. The non-cancer variants were selected randomly from SNPs stored in dbSNP with a minor allele frequency >20%. RF classifiers divide

a large pool of data into smaller subsets based on characteristics of each datum. Figure13 taken from CanPredict website shows the work flow.

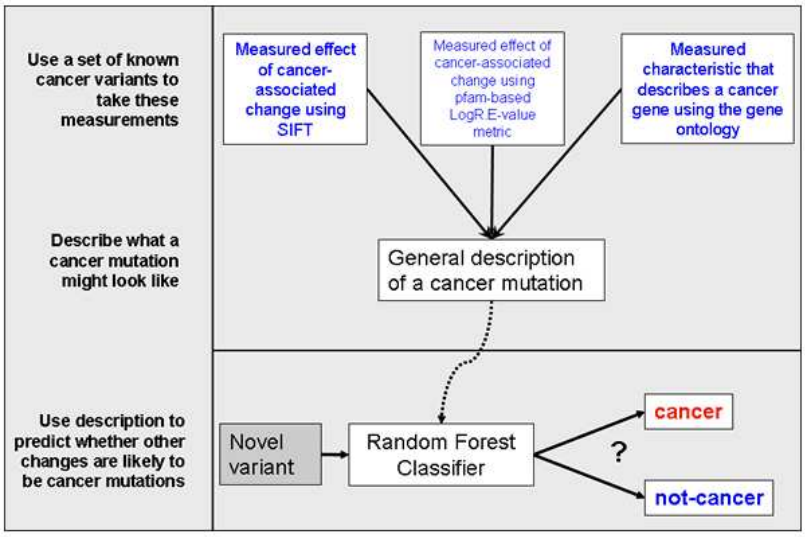


Figure 14: CanPredict Work Flow

Further validation of the classifier was achieved by performing a cross-validation experiment in which a group of known variants was entirely excluded from the training process during the construction of the classifier. The out-of-bag error, an internal measure of the rate of misclassification of the classifier, was determined to be 3.19% suggesting that the classifier is very effective. Results of the analysis are returned to the user in a summary page. There is also a link directing users to a detailed description of the scores produced from each metric. Within the submission summary is a prediction from the classifier indicating likely cancer, likely non-cancer or not determined.

3. CHASM

CHASM, Cancer-specific High-throughput Annotation of Somatic Mutations [23] is a computational method that identifies and prioritizes the missense mutations most likely to generate functional changes that enhance tumor cell proliferation. It is an open-source software, a collection of Python and C++ programs that takes a list of somatic missense mutations as input and ranks them according to their likely tumorigenic impact. The reasoning for creating CHASM was to train a classifier with improved specificity by representing passenger missense mutations not by high MAF nsSNPs, as done previously, but rather by *in silico* simulations using mutation profiles that reflected tumor type as well as mutation context. The classifier was trained on 49 predictive features. Feature selection was done with a protocol based on mutual information, which is a generalized version of correlation that does not make assumptions about linear relationships between two variables of interest. The authors claim that some of the features have not been used previously for missense mutant function prediction. These features include the average nucleotide-level conservation of the exon in which a mutation occurs in 17-way vertebrate Multiz alignments, estimated by PhastCons; SNP density (the number of SNPs in the exon where the mutation occurs, normalized by exon length); and frequency of missense change type in the COSMIC database of somatic variation in cancer. The driver mutation data set comprised of 2,488 missense mutations previously identified as playing a functional role in oncogenic transformation from breast, colorectal,

and pancreatic tumor re-sequencing studies [7], [19], [134] and the COSMIC database. The synthetic passenger mutations were generated in these genes *in silico*, using an algorithm that recapitulated the type of base substitutions found in brain tumors. Genes that were mutated as the substrate for the *in silico* generation of synthetic mutations were purposefully chosen. This increased the likelihood that the new classifier would detect mutations that were extraordinary rather than detect genes that were extraordinary. Before training, all features were standardized with the Z score method. To avoid overfitting, known driver mutations and synthetic passenger mutations were divided into two partitions, one for feature selection and one for classifier training. The mentioned features and data sets were used to design a new classifier using two state- of-the-art machine learning methods, SVMs, and Random Forests. Although both methods were able to define good classifiers, the Random Forest proved superior and was used for the analyses. The final score yielded for each mutation is the fraction of an ensemble of "decision trees," specifically classification and regression trees, each of which uses a hierarchical set of rules to decide whether a mutation is a driver or a passenger that voted for the passenger class. A forest with 500 trees, and default parameters were used to run the classifier. Random Forest classifier performance was assessed by two threshold-independent measures, receiver operating characteristic (ROC) and Precision-recall (PR) curves. CHASM yielded AUCs (area under the curve) of 0.91 and 0.79 for ROC and PR, respectively. CHASM was compared and proved to be superior to other methods, including

PolyPhen's PSIC score, SIFT, CanPredict, KinaseSVM in the fraction of mutations that could be evaluated, specificity, sensitivity, and precision.

4. CanDrA

CanDrA, Cancer Driver Annotation, is a tool that predicts missense driver mutations based on a set of 95 structural and evolutionary features computed by over 10 functional prediction algorithms such as CHASM, SIFT, and MutationAssessor. Two missense mutation datasets, GBM and OVC, were curated from those reported in COSMIC (V58), TCGA, and the CCLE project. Passenger mutations were selected from hyper-mutated samples, which have deficiency in DNA damage repairing and have much higher fractions of passenger mutations than non-hyper-mutated samples. In summary, four stringent sets were formed: GBM.S1, GBM.S2, OVC.S1 and OVC.S2. These sets were used as independent test sets to measure CanDrA's performance against those of other tools. To represent commonality across cancer types, a cancer-type-specific set was constructed with expanded set of drivers and passengers using the empirical rules. For a given cancer type, missense mutation is called a driver mutation if it occurs in a gene mutated in this cancer type and 1) it is observed in at least 3 primary tumor samples (regardless of cancer type), or 2) its site intersects at least 4 mutations (including indels, dinucleotide or trinucleotide mutations), or 3) it is centered in a 25 bp region that intersects at least 5 mutations in the COSMIC database. Passenger mutations of a cancer type

were chosen as those that occur only once in primary tumor samples of this cancer type, not in any COSMIC cancer census gene, and do not coincide with any other mutations within a 31-bp window in the entire COSMIC database. By combining the above putative drivers and passengers for each cancer type, two expanded datasets were formed: GBM.Ex and OVC.Ex. They were used as training sets for feature selection and supervised training. For each missense mutation, 95 features were acquired from four data portals: CHASM's SNVBOX [135] , ENSEMBL Variant Effect Predictor [136], Mutation Assessor [31] and ANNOVAR [128]. Among them are UniProtKB annotations, evolutionary conservation scores, protein physicochemical properties, sequence context indices, and functional impact scores computed by algorithms such as SIFT [43], PolyPhen-2 [53], CONDEL [123], Mutation Assessor [31] , PhyloP [127] , GERP++ [137] and LRT [138]. The predictive performance of each feature was evaluated based on the Mann–Whitney U test and the area under the curve (AUC) of the receiver operating characteristic curve. The feature combinations were assessed using a hybrid feature selection algorithm. The feature set that achieved the maximum AUC in cross-validation was selected as the optimal set.

CanDrA classifies a mutation into 3 categories: driver, no-call, and passenger, based on scores computed by the SVM. The SVM method used by CanDrA is more robust against the COD than other classifiers, including the random forest algorithm used by CHASM. According to the score distributions, a mutation is classified as a driver if its score is greater than the 90th percentile of

those of the passenger mutations in the training set, as a passenger if its score is less than the 10th percentile of those of the driver mutations, or as a no-call otherwise. In addition, CanDrA computes a confidence score for each prediction, defined as the fraction of mutations that have more extreme scores in the same class in the training data. These confidence scores are thus significant P values estimated from the empirical class-wise score distribution in the training dataset. CanDrA was trained using the optimal set of 21 features, and evaluated the performance on the two independent validation datasets (GBM.S1 and GBM.S2). CanDrA achieved AUCs of 0.911 and 0.941, respectively, which compared favorably with those obtained from either CHASM (0.890 and 0.923, respectively) or MutationTastor (0.892 and 0.909) respectively. For evaluating its performance on the two independent validation datasets (OVC.S1 and OVC.S2) CanDrA was trained using 22 features. On both sets, CanDrA achieved AUCs of 0.953, which again compared favorably to those of either CHASM (0.936 and 0.940) or MutationTastor (0.910 on both test sets).

5. mCluster

mCluste [139] is a framework for identifying and elucidating functionally important protein-altering mutations. The mCluster analysis leverages somatic mutation data from large-scale discovery projects and combines it with curated data from both cancer and germline mutation databases. The mCluster approach is based on the hypothesis that functionally significant mutations will fall into

clusters more frequently than functionally insignificant ones. Disease mutations are by definition functionally significant, and cancer mutations should also be enriched for functional mutations because driver mutations are under positive selection. If the hypothesis is correct, then real cancer and disease mutations should fall in clusters more frequently than nsSNPs or simulated random cancer mutations. mCluster projects all of these mutations onto a comprehensive set of conserved protein domains and identifies candidate domain hotspots (clusters): conserved locations that are enriched for mutations across multiple proteins. This can identify important protein regions, suggest functionally significant mutations, and provide insight into how mutations may exert their effects. Figure14 taken from the mCluster publication [139] shows its workflow.

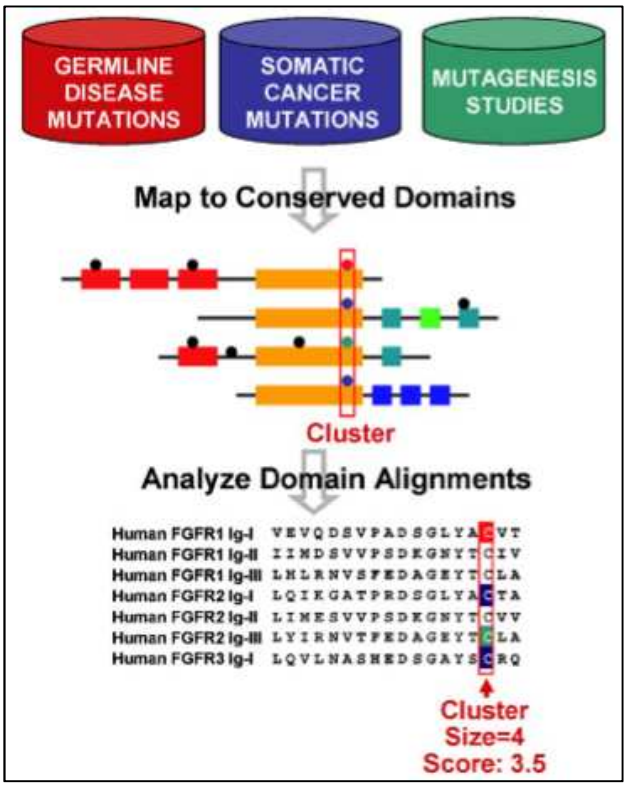


Figure 15: mCluster Work Flow

The datasets used contained missense somatic mutations from the COSMIC database, from cancer genome sequencing projects, the SwissProt database using "Mutagen" as the keyword and from dbSNP database with corresponding reference protein sequences retrieved from the RefSeq database. Germline disease mutations were extracted from the SwissProt database using "Variant" as the keyword. Passenger mutations were simulated by randomly introducing missense mutations into the 1,898 coding sequences in which mutations were identified by Wood et al. [7]. Mutations were generated by applying the nucleotide context-specific mutation rates observed in colorectal cancer samples [19], and only missense changes were kept. The Pfam domain boundaries of the

coding sequences were retrieved from Unison and domains predicted using Hmmsearch [140]. Boundary coordinates were used to extract domain-specific subsequences, which were then aligned to the Pfam HMM profiles using Hmmalign [140] to generate multiple sequence alignments. Mutation clusters were identified as sets of mutations located in the same column in a Pfam domain alignment. The size of a cluster is the total number of mutant samples for cancer mutations and number of unique mutations in case of germline mutations. The mCluster score reflects the probability of observing a cluster of a given size given the number of available positions in a domain and the total number of mutations observed. The mCluster score is an important aid to the interpretation of mutation clusters. By taking into account the total number of mutations observed in a domain it can indicate the significance of each cluster regardless of variation in sequencing coverage and mutation rates.

The scoring scheme considers all positions equally likely to be mutated. Within a given domain, clusters were ranked by their sizes in a decreasing order, and then a step-down approach, which improves the ability of the score to distinguish real cancer and germline disease mutations from simulated random mutations. Pathways enriched for genes with mutations in high-scoring clusters relative to genes with any mutations in conserved domains are identified by Fisher's Exact Test. The results showed that the class of variants that fell most frequently within conserved protein domains was germline disease mutations, which are all expected to have significant effects on protein function.

Furthermore, nsSNPs and random cancer mutations, which are not enriched for functional variants, fell within conserved domains least frequently. Cancer mutations from COSMIC and resequencing studies fell within domains with intermediate frequencies, consistent with them consisting of a mixture of functional and nonfunctional variants.

6. transFIC

TransFIC, TRANSformed Functional Impact for Cancer [141], is a method to transform Functional Impact scores taking into account the differences in basal tolerance to germline SNVs of genes that belong to different functional classes. This transformation allows to use the scores provided by well-known tools (e.g. SIFT, Polyphen2, MutationAssessor) to rank the functional impact of cancer somatic mutations. Mutations with greater transFIC are more likely to be cancer drivers. TransFIC takes as input the Functional Impact Score of a somatic mutation observed in cancer provided by one of the aforementioned tools. It then compares that score to the distribution of scores of germline SNVs observed in genes with similar functional annotations (for instance genes with the same molecular function as provided by the Gene Ontologies). The score is thus transformed using the Zscore formula. The result is that mutations in genes that are less tolerant to germline SNVs are amplified, while the scores of mutations on relatively tolerant genes are decreased.

The approach attempts to rank cancer somatic mutations and is based on the observation that genes with dissimilar functions show different tolerance to germline variants, measured as the distribution of functional impact scores of variants accepted during human evolution. This was called the *baseline tolerance* of genes. Then the functional impact scores of mutations provided by three well known tools were transformed using this baseline tolerance and compared with the performance of the transformed score and the original score in separate sets of variants enriched for driver mutations (positives) and passenger mutations or polymorphisms (negatives). The rationale behind the transformation is that if two mutations with the same FIS affect genes with different germline tolerance to functional SNVs, the impact of the mutation on the least tolerant gene is expected to be greater than its impact on the most tolerant one. Genes with essential cellular functions would appear on the lower end of the functional impact score scale, while genes whose malfunction can be compensated for by diverse mechanisms or does not lead to very deleterious phenotypes are located at he upper end of the FIS scale. The dataset used consisted of all SNVs detected by the 1000 Genomes project within the genomic sequences of 1197 individuals. Ensembl Variant Effect Predictor was then used to detect nsSNVs and to get their SIFT, PolyPhen2.0 scores. Corresponding MutationAssessor FISs were obtained via MA webAPI service. Four systems of functional annotation were used to partition the dataset of SNVs and form pools of functionally related genes, they were the GOBP and GOMF categories, the CP

annotations and Doms. First the functional impact for all germline SNVs detected in the human population (1000 Genomes Project) are computed using SIFT, PolyPhen2.0 and MA. Next a measure of baseline tolerance to germline SNVs is computed for each protein coding gene. This is done by pooling all genes with GOMF terms shared by the gene in question and computing the means and standard deviations of the FISs of the nsSNVs that affect them. Figure15 taken from TransFIC publication [141] shows its work flow.

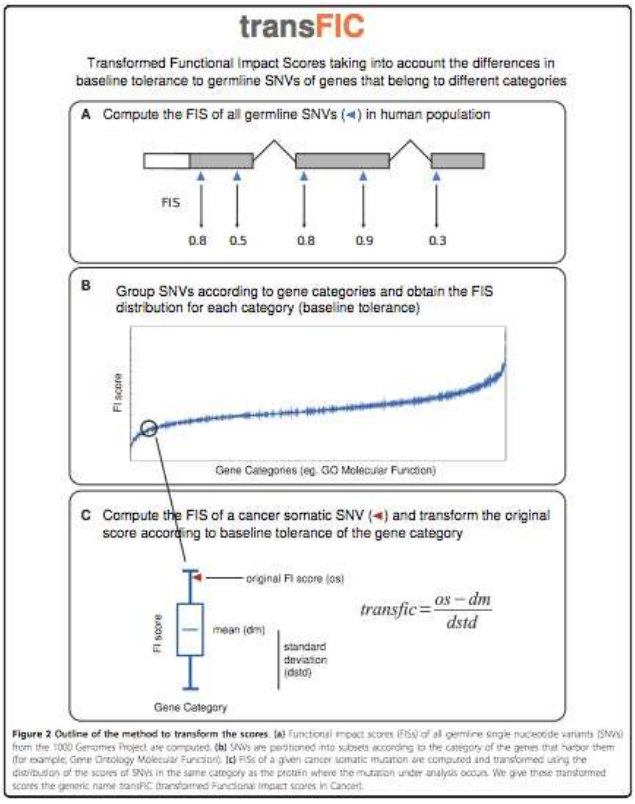


Figure 2 Outline of the method to transform the scores. (a) Functional impact scores (FISs) of all germline single nucleotide variants (SNVs) from the 1000 Genomes Project are computed. (b) SNVs are partitioned into subsets according to the category of the genes that harbor them (for example, Gene Ontology Molecular Function). (c) FISs of a given cancer somatic mutation are computed and transformed using the distribution of the scores of SNVs in the same category as the protein where the mutation under analysis occurs. We give these transformed scores the generic name transFIC (transformed Functional Impact scores in Cancer).

Figure 16: TransFIC Work Flow

Three categories (low, medium and high impact) were devised, into which somatic mutations could be classified based on their transformed FIS. A condition of using at least 20 nsSNVs pooled from genes within the same functional group(s), to compute the baseline tolerance of a gene needs to be fulfilled. Nine proxy datasets were arranged for the evaluation. The transformed FIS of all somatic mutations was computed for all the nine proxy datasets. To assess the performance of each FIS (or transformed FIS) in identifying likely functional somatic mutations, the Matthews correlation coefficient (MCC) and overall accuracy (ACC) yielded by the classification of positive and negative cases in each proxy dataset were computed. It was found that the transformed FIS outperforms the original FIS on all nine proxy validation sets. TransFIC of CHASM also outperformed the original CHASM scores.

TransFIC can be run using transFIC server, which implements the TransFIC of SIFT, Polyphen2 and MutationAssessor. Alternatively, it is possible to download a PERL script that computes transFIC for these same tools. While the webserver takes as input the genomic coordinates (hg19) and nucleotide change of the somatic mutation, the PERL script receives SIFT, PPH2 and MA scores to compute the corresponding TransFIC.

7. MuSiC

MuSiC, Mutational Significance in Cancer, [142] a comprehensive mutational analysis pipeline that uses standardized sequence-based inputs along with

multiple types of clinical data to establish correlations among mutation sites, affected genes and pathways, and to ultimately separate the commonly abundant passenger mutations from the truly significant events. The primary goal of MuSiC is to separate the significant events, which are likely drivers for disease from the passenger mutations present in mutational discovery sets using a variety of statistical methods. The integration of analytical operations in the MuSiC framework is stated to be widely applicable to a broad set of tumor types and offer the benefits of automation as well as standardization. As input MuSiC needs a few basic elements such as the mapped reads in BAM format, predicted or validated SNVs and indels in mutation annotation format (MAF), a set of regions of interest and any relevant numeric and/or categorical clinical data. The user interface allows users to (1) apply statistical methods across the cohort to identify significantly mutated genes and (2) identify significantly altered pathways and gene sets, (3) investigate the proximity of amino acid mutations within the same gene, (4) search for gene- based or site-based relationships and correlations between the mutations themselves, (5) correlate mutations to clinical features, and (6) cross-reference the findings with relevant databases, such as Pfam, COSMIC, and OMIM. MuSiC currently consists of seven analysis modules and an eighth execution module, "MuSiC Play," which runs each analysis module sequentially. MuSiC Play parses the input and output of each of the individual modules and then produces a composite summary of all executed modules. MuSiC includes a number of tools to help identify significant

mutations and relationships in a cancer mutation dataset: SMG for computation of background mutation rate in your dataset and identification of significantly mutated genes, PathScan for Identification of significantly altered gene sets and/or pathways using KEGG or other databases, Proximity to search for mutations physically near one another at the DNA or protein level to identify mutation hotspots, COSMIC-OMIM for comparison of your mutations with those submitted to COSMIC and OMIM databases, Mutation-Relation for Detection of co-occurring or mutually exclusive mutation relationships between genes and Clinical Correlation tool for correlating of gene mutation status with categorical or quantitative clinical data, such as tumor subtype.

For every mutation found in the input MAF file, information related to this mutation is gathered from both the COSMIC and OMIM databases. Relevant information is ascertained by relating the genomic coordinate (COSMIC) or the amino acid change (COSMIC and OMIM) associated with the variant to all database entries. A database entry must lie within the user-specified number of bases (default = 5) or amino acids (default = 2) of the MAF variant to be considered as a "nearby" match. The thresholds for "nearby" matches are viewed as adequate for taking into account the different definitions of reference sequences and gene transcripts that may be used by different contributors to the databases. Alternatively, "exact" matches are direct overlaps in both the location and base/amino acid change of a variant in the MAF and a mutation found in the database. If only the location (the genomic coordinate or amino acid

position) of a variant match a database entry but not the nucleotide or amino acid change, these matches are deemed “position” matches. All discoveries based on these database queries are appended to the input MAF file as extra columns. For each queried database, the tool further prints an output summary which tallies the types of matches found throughout the entire data set using that database. MuSiC can not only analyze mutations at the single gene level i.e. using the SMG ‘significantly mutated genes’ test, it also integrates the pathscan algorithm [143] for mutation analysis at the pathway level. Another module in MuSiC is the mutation relation test (MRT) module, which tests whether mutations in any two genes act concurrently (positive correlation) or exclusively (negative correlation). Finally, the clinical correlation test (CCT) module tests for correlations between mutations and clinical features.

8. SPF-Cancer

SPF-Cancer, [30] is disease specific, SVM machine learning approach to predict cancer-causing missense variants. The input features of the algorithm (SPF-Cancer) include: the amino acid mutation, its local sequence environment, sequence-profile derived features, the output of PANTHER algorithm and a cancer-specific functional-based log-odd score calculated considering the GO slim ontology. The final input vector consists of 51 values: 40 components encoding for the mutation and the local sequence environment

(Seq) - the first 20 out of 40 explicitly define the mutation by setting to –1 the element corresponding to the wild type residue and to 1 the newly introduced residue, the last 20 input values encode for the mutation sequence environment (again the 20 elements represent the 20 residue types); 5 inputs features derived from sequence profile (Prof); 4 elements vector from the PANTHER output; 2 elements encoding for the number of GO slim terms associated to the protein sequence and the GO slim log-odd score (LGO). For each mutation: the frequency of the wild type, the frequency of the mutated residue, the number of totally and locally aligned sequences and a Conservation Index (CI) for the position at hand are derived, the more a residue is functionally important the more is conserved over evolution. The 4 elements vector from PANTHER output is composed by the probability of deleterious mutation, the frequencies of the wild-type and new residues in the PANTHER family alignment and the number of independent counts. The Gene Ontology log-odds score (LGO) is computed to derive information related to the correlation among a given SAPs effect (cancer-causing and neutral) and the protein function. The dataset contains SAPs (Single AminoAcid Polymorphisms) data from different sources. Cancer-causing variants are selected from breast, colorectal, pancreatic tumor resequencing studies [7], [19], [134] and COSMIC database that are provided with CHASM package. Neutral variants are from SwissProt database or generated by CHASM. Other disease-related variants are non “neoplasms” disease-related variants annotated in SwissVar database. The Support Vector Machine (SVM) classifies

SAPs (Single AminoAcid Polymorphisms) in cancer-causing (desired output set to 0) and neutral polymorphism (desired output set to 1). The SVM output is a number between 0 and 1 and the decision threshold has been set to 0.5. The results obtained with our SVM methods are evaluated using a cross-validation procedure. The accuracy measures are calculated using a 2-fold cross validation procedure. A MCC (Matthews correlation coefficient) is defined, the coverage (sensitivity) for each discriminated class is evaluated, the probability of correct predictions P (or positive predictive values) is computed and finally, a reliability score to each prediction is assigned. Other standard scoring measures, such as the area under the ROC curve (AUC) and the true positive rate (TPR = Q(s)) at 10% of False Positive Rate (FPR = 1-P(s)) are also computed.

Some of the other published studies that review different predictive tools are, Ohanian *et al.* [144], Tavtigian *et al.* [44], Thusberg *et al.* [145], Castellana *et al.* [146], Karchin [147], Gnad *et al.*[40], Gong *et al.* [41], Stefl *et al.* [21], Krishnan *et al.* [148]. A list of all these tools and additional tools is presented in Table 1 below.

Table 1: List of functional impact predictive tools

Method		**Custom**	**Decision tree**	**SVM**	**RF**	**HMMs**	**NN**
Features	**Property**						
Sequence based	Custom						
		SIFT		PhD-SNP		PANTHER	SNAP
		MAPP		SNPs&GO			
		Mutation Assessor					
		Align-GVGD					
		Panther					
	Protein Stability			MuStab			
	Using SIFT			SPF-Cancer	CanPredict		SNPdbe
					VEST		
					CHASM		
Sequence and/ or Structure based	Custom						
			PolyPhen	MUpro		topoSNP	PoPMuSiC-2.0
			PolyPhen 2	LS-SNP			PMUT
				SNPs3D			MuPro
				SAPRED			
				SNPs&GO 3D			
				CanDrA			
		HOPE					
	Using SIFT	MutPred			nsSNPAnalyzer		
	Using PolyPhen	FastSNP					
	Protein Stability / Folding Free energy (ΔΔG)			I-Mutant 2.0			
				Scpred			
Structure based	Custom	PoPMuSiC					
		stSNP					
		CC/PBSA					
		Bongo					
	Protein Stability / Folding Free energy (ΔΔG)	SDM			AUTO-MUTE		PoPMuSiC-2.0
		CUPSAT					I-Mutant
		FoldX/FOLDEF					
		ERIS					
		MultiMutate					
		Dmutant					
		Scide					
		Sride					

Comparative Analysis of Predictive methods

Functional prediction of missense mutations can be extremely useful in clinical genetics. Functional predictions can assist in identifying previously unknown causes of Mendelian diseases [149]. There are a lot of cancer diagnostic genetics test being developed to test the presence of certain mutations in the patients' genes or genomes, which can help decision making during diagnosis or treatments to follow. However, there has been a general conclusion that these predictive methods are not accurate enough to be relied on for clinical decision-making but can be very useful when combined with other traditional genetic methods [44], [150].

Predictive methods, their coverage and their scores cannot be compared on the same lines as the predictions and their accuracy vary depending on the type of features selected, test data sets used and scoring methods employed. When methods are originally published the authors provide some performance measures, which cannot be comparable to other methods due to the difference in the approach followed. There cannot be any one measure that can accurately assess and compare the performance of different predictive tools. Comparing multiple methods is problematic because there is no standard classification system used to categorize the predicted functionality of the variants needed to provide a statistical measure of performance of the methods. For example, sequence conservation based methods would have low classification confidence

in cases where the aligned homologs show low sequence diversity at the same time structure based methods would have low confidence where the predictions are based on modeled structures [40]. With an increase in number of predictive tools, some comparative assessment has been done on these predictive tools. We summarize those reviews below.

- Gnad *et al* [40] assessed a group of eight predictive tools namely, SIFT, PolyPhen-2, SNAP, mCluster, log RE, Condel, Mutations Assessor and CHASM. They based their assessment on three test data sets and a negative (neutral) data set. The first data set was a set of likely cancer driver mutations from COSMIC database. Their second data set was created from recurrent somatic mutations in colorectal carcinoma from TCGA, and the third set created from recurrent unique mutations found in breast or colon cancer. The neutral set of mutations was of likely non-deleterious variants from germ-line SNPs found in dbSNP, with a minor allele frequency of atleast 0.25 to avoid rare deleterious mutations and errors. The criteria of selecting these datasets was supported by an initial scoring of all variants using SIFT. Important takeaways from their results are that CHASM, Mutation Assessor, PolyPhen-2, SIFT, Condel and SNAP were able to score most of the variants. CHASM yielded the highest accuracy with the recurrent mutations COSMIC data set but not with other datasets. Mutation Assessor yielded consistently highest results and scores except against CHASM with COSMIC mutations. These differences in performance reflect the training

and testing data set bias. Examining combinations of predictors showed that combinations of SIFT, PolyPhen-2 and Mutation Assessor gave better results compared to other combinations. No combination improved on Mutation Assessor alone. Likely misclassifications were discovered in predictions by all the predictors including Mutation Assessor, so no single predictor can be named the best. It is clear that machine learning- based approaches are essentially affected by this problem. It is perhaps more practical to develop multiple specific algorithms for different classes of mutations, instead of develop a "one-size-fit-all" approach.

- Thusberg *et al.* [81] tested the performance of nine widely used pathogenicity methods namely, MutPred, nsSNPAnalyzer, Panther, PhD-SNP, PolyPhen, PolyPhen2, SIFT, SNAP and SNPs&GO. They used two datasets, one pathogenic dataset and the other a neutral mutations dataset. The pathogenic data set was created from the PhenCode database records, which were also annotated as disease causing in Swiss-Prot database, registries in IDbases and from 18 individual locus specific databases. The neutral data set used was taken generated from dbSNP, containing human non-synonymous coding SNPs with an allele frequency > 0.01 and chromosome sample count >49. These were further divided into two subsets, one which had a 3d structure available in PDB (to test the methods involving structural information) and second subset containing pathogenic mutations not present in Swiss-Prot (to probe the effect of using Swiss-Prot derived data). The

corresponding neutral set was just a subset of the original neutral mutations. All the methods were run on default parameters and outputs converted into binary predictions. They used six different measures to evaluate the performance of the predictors, namely accuracy, precision (or positive predictive value, PPV), specificity, sensitivity, negative predictive value (NPV) and Matthews correlation coefficient (MCC). Correlations between the program outputs were calculated by counting all of the common cases and those predicted correctly and using Spearman's rank correlation coefficient. The performance of all methods was generally worse except for sensitivity. SNPs&GO performed best in terms of accuracy (0.82), precision (0.90), specificity (0.92), and MCC (0.65), but sensitivity was higher in six other methods, and MutPred, Panther, PolyPhen2b, and SNAP performed better in terms of NPV. nsSNPAnalyzer performed worst in terms of MCC (0.19), accuracy (0.60), NPV (0.60), and precision (0.59). The two versions of PolyPhen have very similar overall performance; however, PolyPhen2 is recommended because the quality measures are more balanced. NPs&GO performed best also in the structural subcategory considering accuracy, precision, specificity, and MCC, and MutPred was the best method in terms of sensitivity and NPV [81]. Overall, they found SNPs&GO and MutPred to be clearly the most reliable predictors for their dataset of genetic variants. The two best performing predictors include both protein structural or functional and MSA-derived information in the prediction. They conclude that there is

no single method that could be rated as best by all parameters. Complementary methods could be combined in a metaserver to yield more reliable predictions.

- Hicks *et al.* [151] state that in order to predict missense mutation functionality, the users should consider optimizing both the algorithm and sequence alignment employed. They found that a given algorithm did not necessarily perform best using the alignment provided by the creator of the algorithm. For example, the PolyPhen-2 algorithm reported higher sensitivities in all four genes using alignments other than its own and SIFT had a slightly higher AUC when provided the Align-GVGD alignment containing only orthologs. The three algorithms SIFT, PolyPhen-2, and Xvar all had a high sensitivity, but low specificity implying these algorithms may overcall neutral variants deleterious.
- Castellana *et al.* [146] aimed at measuring the degree of congruency and consensus among a set of tools they tested namely, SIFT, PolyPhen-2, VEP, MutationAssessor, Carol and Condel. Owing to the intrinsic diversity of the numerical results of the predictors (different scales and different algorithms), they opted to compare the categorical outcomes thereby evaluating the degree of agreement of their results. Congruent classifications, namely, those exhibiting low scores were more commonly observed than high scores. Comparing the derived predictions (Carol and Condel), they observed a high

level of uncertainty and that the consensus algorithms may yield results that substantially contrast the predictions that they derive from.

- Chan *et al. [152]* compared computational methods that use evolutionary conservation alone, amino acid (AA) change alone, and a combination of conservation and AA change in predicting the consequences of 254 missense variants in 5 genes (CDKN2A, MLH1, MSH2, MECP2, and TYR). They tested four predictive methods, BLOSUM62, SIFT, PolyPhen and A-GVGD. Their tests gave concordant results for only 63% of the variants. However, when this occurred, the overall predictive value increased to 88%.
- Wei and *et al.* tested six predictive tools, namely, SIFT, PolyPhen, PMut, SNPs3D, PhD-SNP, and nsSNPAnalyzer. They conclude that the top predictive methods are PolyPhen, SIFT, and nsSNPAnalyzer, which have similar performance. They also found that when different combinations of programs were used, the consensus of 5 programs (SNPs3D excluded) gave the best total accuracy (73%).

The overall conclusion of assessing various predictive tools is that predictions obtained using different predictive tools should be interpreted in context with caution, as they can be affected by the extreme sensitivity of specific algorithm settings used, by the 'training' datasets used and by the availability of supplementary information (e.g. orthologous sequences, 3D structures or gene ontology data). Several of the methods are very specific, and might analyze the

same single feature. However, they may have different points of view for analyzing the same property. For example, structural changes may originate from changes in side-chain size, hydropathy, altered contact-forming properties, aggregation, or introduced disorder.

Different groups of methods closely correlate and complement each other and are individually or in combination are suited for different types of functional assessment. It has been suggested in studies that although confidence in a result may be increased if concordant results are obtained with a number of programs, but some pathogenic variants may be missed. At the same time, having less stringent criteria, such as requiring any single program to be stringent towards deleterious mutations, would increase the chances that all the true positives will be detected but may also result in more false-positive results. Also, it needs to be considered that having similar outputs could be a result of similarity of inputs for some combinations of programs and this does not necessarily equate with greater prediction accuracy [144].

Protein Stability Predictors

With more than 80% of disease associated missense mutations affecting the stability of proteins by several kcal/mol [67], there has been a rush in designing predictive methods focusing only on effects of missense mutations on protein stability and thereby its function. Protein stability predictions have been approached by predicting the structural effects of mutations using 1. molecular

mechanics approaches, 2. empirical energy functions, which are fitted to experimental data using weighted terms incorporating physical and statistical factors with structural knowledge and 3. machine learning methods such as support vector machines (SVMs) and neural networks as well as statistical potential energy functions which are derived using statistical analysis of information from different databases. The empirical rules that affect protein stability may include the elimination of hydrogen bonds, diminished hydrophobic interaction, loss of a salt bridge, introduction of the buried charged residue, loss of a disulfide bond, backbone strain caused by substitution of glycine to another residue, or change to proline [153] [122]. Also other damaging effects of missense mutations, such as ligand binding, catalysis, allosteric regulation, and post-translational modification, as the causes of functional disruption can also be closely tied to the stability effect. There are many tools that focus on predicting effects on protein stability such as FOLD-X [70], Site Directed Mutator (SDM) , AUTO-MUTE [68], CUPSAT [71], MultiMutate [72], Dmutant [73], I-Mutant 2.0 [60], PoPMuSiC-2.0 [69], MUpro [59], SCide [154], SCPRED [155], SRide [156]etc.

Such tools can certainly assist in refining and improving prediction accuracies, for example, DMutant improves structure-derived potentials of mean force for structure selection and stability prediction. SCide identifies of stabilization centers in proteins. SCPRED predicts protein structural class for sequences of

twilight-zone similarity. SRide is a server for identifying stabilizing residues in proteins.

1. AUTO-MUTE

AUTO-MUTE [68] uses a four-body, knowledge-based, statistical contact potential. The program calculates an empirical, normalized measure of the environmental perturbation for substitutions. For its models, each feature vector representing a mutant includes input attributes obtained by applying a computational mutagenesis that utilizes a four-body, knowledge-based, statistical contact potential. This amino acid distance potential applies the inverse Boltzmann principle and is derived via Delaunay tessellation of protein structures, a classical computational geometry tiling technique that objectively identifies quadruplets of 3D nearest neighbor residues. This feature vector is then used to estimate the effect of the mutation by considering the spatial perturbation inflicted by the mutation upon its nearest neighbors in the 3D structure.

2. CUPSAT

CUPSAT, Cologne University Protein Stability Analysis Tool, [71] predicts the difference in free energy of unfolding between wild-type and mutant proteins, $\Delta\Delta G$, using structural environment-specific, atomic potentials and torsion-angle potentials derived from nonredundant protein structures obtained from PISCES webserver. For the atom potentials, a radial pair distribution

function with an atom classification system has been used. The atoms□are classified into 40 different types (9) according to their□location, connectivity and chemical nature. Boltzmann's energy values are then calculated from the radial pair distribution of amino acid atoms. The torsion-angle potentials are derived from the distribution of protein backbone Φ and ψ angles for all the amino acids in the dataset. After calculating Boltzmann' energy values, a Gaussian apodization function (11) was applied to assign favourable energy values for the neighbouring orientations of observed Φ- ψ combinations. The secondary structure specificity of mutations and mean-force potentials were implemented and the amino acids were classified into helices, sheets and others, then amino acids belonging to each of these secondary structure elements were further subdivided according to their solvent accessibility thereby constructing the prediction model.

3. SDM

SDM [157] Site Directed Mutator, is a statistical potential energy function developed to predict the effect that SNPs will have on the stability of proteins. It predicts the effect that single point mutations have on protein stability, based on knowledge of observed substitutions that have occurred in homologous proteins and which are encoded in environment-specific substitution tables. SDM calculates a stability score, which is analogous to the free energy difference between a wild-type and mutant protein. SDM uses the local structural

environment of the wild-type and mutant residues to calculate the stability score. The structural parameters that are used to define the local structural environment of amino acid residues are main chain conformation, solvent accessibility and hydrogen bonding class. SDM output therefore includes the local structural environment of the wild-type and amino acid residues. The stability score indicates the predicted effect of the mutation on protein stability. A negative score indicates that the mutation is destabilitzing whereas a positive score indicates that the mutation is stabilizing. $\Delta\Delta G$ less than -2.0 or greater than 2.0 is classed as highly stabilizing and is predicted to be disease-associated.

4. DFIRE-Dmutant

DFIRE-Dmutant [73] uses a statistical potential approach with a distance-dependent, residue-specific, all-atom, and knowledge-based potential for protein structure-based predictions. The distance dependence of the pair probability distribution of the reference state is an averaged distribution over all residue or atomic pairs.

5. FoldX

FoldX [70] [158] is an empirical force field that was developed for the rapid evaluation of the effect of mutations on the stability, folding and dynamics of proteins and nucleic acids. The core functionality of FoldX is the calculation of the free energy of a macromolecule based on its high-resolution 3D structure. It

is based on an empirical potential approach that uses an energy function derived from a weighted combination of physical-energy terms, statistical-energy terms, and structural descriptors calibrated to fit experimental $\Delta\Delta G$ values. An important difference between FoldX and other force fields is the crude entropy estimation that is used to obtain a measure of the free energy. Entropy calculations usually involve large simulations of the conformational freedom of the side chains and the backbone of the protein. In FoldX the entropic penalty for fixing the backbone in a given conformation, is derived from a statistical analysis of the phi–psi distribution of a given amino acid as observed in a set of non-redundant high-resolution crystal structures. This entropy is scaled by (i) the accessibility of the main chain atoms and (ii) energetics of hydrogen bond interactions made by the corresponding residue or its direct neighbors.

FoldX and Dmutant are the only programs that return negative $\Delta\Delta G$ values for stabilizing mutations and positive values for destabilizing mutants.

6. I-Mutant2.0

I-Mutant2.0 [60] is a support vector machine (SVM)-based tools. The services use either a protein structure or a sequence as input. I-Mutant2.0 can be used both as a classifier for predicting the sign of the protein stability change upon mutation and as a regression estimator for predicting the related $\Delta\Delta G$ values. I-Mutant2.0 has been trained to accomplish four different tasks: (i) Prediction of the direction of the protein stability changes upon single point

mutation from the protein tertiary structure (a classification task); (ii) Prediction of the $\Delta\Delta G$ value of the protein stability changes upon single point mutation from the protein tertiary structure (a function approximation task); (iii) Prediction of the direction of the protein stability changes upon single point mutation only from the protein sequence (a classification task); and (iv) Prediction of the $\Delta\Delta G$ value of the protein stability changes upon single point mutation only from the protein sequence (a function approximation task). For the classification task and for assigning the $\Delta\Delta G$ values, it identifyies two labels: one represents the increased protein stability ($\Delta\Delta G > 0$, label is +), the other is associated with the destabilizing mutation ($\Delta\Delta G < 0$, label is -). The input vector consists of 42 values. The first two input values account, respectively, for the temperature and the pH at which the stability of the mutated protein was experimentally determined. The next 20 values (for 20 residue types) explicitly define the mutation (-1 is set to the element corresponding to the deleted residue and 1 to the new residue, all the remaining elements are kept equal to 0). Finally, the last 20 input values encode the residue environment that is a 'spatial environment' when the protein structure is available or the nearest sequence neighbors, when only the protein sequence is available.

7. MultiMutate

MultiMutate [72] uses a four-body scoring function based on Delaunay tessellation of proteins to predict the effects of single- and multiple-residue

mutations on the stability and reactivity of proteins. Each amino acid is represented by a single point located at the centroid of the atoms in its side chain (including the Cα atom). α represents the type of the tetrahedron based on the backbone chain connectivity of the four participating amino acids. There are five tetrahedron types possible, and α takes one of the values 0,1,2,3 or 4 corresponding to these types [159]. The total score of a protein is then defined as the sum of the log-likelihood ratios of all tetrahedra in its Delaunay tessellation. The method calculates the change in how well packed the residues are in the wild-type protein and in the mutant. Score values between 0.5% and -0.5% are classified as negative and it can be interpreted in a relative sense and compare and quantify two otherwise similar conformations based on how well one of them is packed better than other. This method uses the structure of the WT for the mutant as well, changing only the sequence. It uses the side chain center representation compared to the Cα atoms representation and use an increased cutoff on the Delaunay edges of 12 A □ when scoring mutations. With the above settings the change in total score between the WT and the mutant protein is calculated. A positive change is when the mutant score is more than the WT score and indicates that the mutant is more stable than the WT, while a negative change indicates lower stability. The change is score is represented as the fraction of change, given as percentage, to the sum of the scores of the tetrahedra that see any change due to the mutations. If the percentage change is below 0.1% in absolute value, it is assumed there is no change. The authors also

correlate increased (decreased) activity with a negative (positive) change in the total score based on he intuition that well-packed proteins are typically not highly active, and hence the high total score is correlated with less activity. This is however based on a very limited data set.

8. MUpro

MUpro [59],predictions of protein stability changes upon mutations, is a set of machine learning programs to predict how single-site amino acid mutation affects protein stability. It has two machine learning methods: Support Vector Machines and Neural Networks. One advantage of this method is that it does not need tertiary structures to predict protein stability changes. First the methods predict whether a mutation will increase or decrease the stability of protein structure. Second, it uses a machine learning to predict directly the $\Delta\Delta G$ resulting from single site mutations. These methods use different methodologies and input information somewhat differently. For instance, prediction of $\Delta\Delta G$ uses regression methods, while prediction of sign uses classification methods. From the protein sequence it uses a local window centered around the mutated residue as input. To estimate the performance on unseen and nonhomologous proteins, this method removes the mutations associated with the homologous proteins and splits the remaining mutations by individual proteins. It uses the leave-one-out cross validation for the SVM thus empirically estimating how well the method can be generalized to unseen and nonhomologous proteins. The

methods are evaluated on a large dataset containing 1615 mutations using 20 fold cross validation procedure. Under this procedure, the dataset is split evenly into 20 folds. Any one fold is used as test dataset, another remaining 19 folds are used as training dataset. Thus there are 20 pairs of testing and training datasets. For each pair, the SVM and neural network are trained on the training dataset and tested on the testing dataset. The performance on all test datasets are combined and reported as the performance of tested methods.

Other programs which could help the predictive methods are SCide [154], Scpred [155] and SRide [156] which identify stability centers from sequence data. SCide attempts to identify stability centers within experimentally determined protein structures. Stabilizing, cooperative, long-range contacts identified by SCide are formed between regions that are sequentially well separated or that are part of different subunits within a complex. Scpred locates stability-center elements that impart stability via cooperative, long-range interactions. Scpred uses a neural network to predict stabilizing residues in conjunction with sequence information for the protein under study and its homologues. SRide combines several methods to identify residues expected to play key roles in stabilization. It analyzes tertiary structures, rather than primary structures, and the evolutionary conserved residues contained within. A residue is predicted to be stabilizing if it is surrounded by hydrophobic residues, exhibits

long-range order, has a high conservation score, and, if it is part of a stability center.

Computational geometry methods

Given a protein and its amino acid sequence, one can represent it using methods drawn from computational geometry. Each amino acid is considered as a single point in 3D space using numerical coordinates from PDB, with the whole protein then represented by a 3D graph where the nodes are the amino acids and the edges connect to the nearest amino acids. Once a protein is represented numerically/graphically features can be extracted, which can be used for classification of unlabeled mutants. For further processing, amino acids are abstracted in terms of their alpha carbon atomic coordinates. Each protein, that is, an amino acid sequence, is thus a sequence of corresponding alpha carbons.

It can be concluded from the previous sections, that protein stability predictors will be more useful based on the fact that 80% of disease associated nsSNPs are protein destabilizing and sequence and structural information or features together give better predictive values to most of the published predictive methods. This also stands true based on the fact that amino acids residues affect each other and also the whole structure of the protein. One of the widely used form of simplified inter-residue potentials is the contact potential in which amino acids interact if they are spatially located within a certain distance from each other [160], [161]. Packing has an importance equal to that of hydrophobicity in determination of protein stability [162]. The packing of a

protein interior can be closely approximated in most cases as a series of short-range, nearest-neighbor interactions [163]. Contact potentials when calculated for pairs of amino acids reflect structural parameters such as bond lengths, bond angles, and charges. These potentials reflect contact preferences and are thus useful for evaluating sequence-structure compatibility and can yield reliable predictions of folding free energy for datasets of mutations [164]. Pairwise, triplet and quadruple based higher order residue interactions play a crucial role to attain the stable conformation of the protein structures. Higher order residue interactions also contribute to the potential energy landscape of proteins [165]. Potentials based on pair-wise interactions have been successfully applied for predicting protein fold recognition, protein structure predictions, assembly mechanism and stability [166]. Higher order contact potentials, based on 3-residue or 4-residue structural neighbors, provide more interaction details that are not captured in residue pairs. Higher order interactions have be used to improve accuracy of fold recognition and generic structure analysis [95], [159].

Four nearest neighbor residues forming quadruplets can characterize functionally important clusters, such as metal binding motifs, thereby allowing the detection of familial relationships between proteins [167]. These nearest-neighbors have been defined subjectively in based on arbitrary criteria and therefore the affects of the potentials calculated in such methods strongly depend on the criteria chosen. Delaunay tessellation method is applied to objectively define all quadruplets in a protein; this is further discussed in the

next section. The method provides an objective definition of nearest neighbors, thereby removing the dependency of the results on arbitrary cutoffs used to define neighbors. Enhanced performance of four-body potentials based on Delaunay simplices was observed and thought to provide a natural, mathematically rigorous decomposition of the network of fully three-dimensional interactions [168].

Voronoi polydedron and Delaunay tessellation

Voronoi tessellation was introduced to the field of protein structure by Finney and Richards [169]. A 3D structure of a protein is represented by a set of points in 3D space, whose coordinates include Cα carbons of the residues forming the protein. Voronoi tessellation partitions the 3D space into a set of convex polyhedra, each containing a single Cα atom, such that the interior points of each polyhedron are closer to their corresponding Cα atom than any other in the system. Given that each point represents an amino acid (either the position of the carbon alpha or the center of mass of the side chain), the Voronoi polyhedron defines the "influence zone" around a given residue and represents the volume available for that residue. The equivalent tessellation of Delaunay comprises polyhedra whose edges connect the centers of Voronoi polyhedra and meet at a common vertex. This results is an aggregate of space-filling irregular tetrahedra, called Delaunay simplices. Vertices associated with each Delaunay simplex objectively define a set of four nearest-neighbor residues without any explicit dependence on an adjustable distance parameter. Figure 17(a) shows a

Voronoi tessellation, constructed by drawing the bisecting lines between that point and all other points in the set and then finding the smallest polyhedron defined by these lines. Figure (17b) shows Voronoi tessellation obtained by forming polyhedrons for all points in the set. Delaunay tessellations are constructed by adjoining points/residues whose polyhedra have a common edge, shown in Figure 17(c). Delaunay simplices then give ensembles of four nearest neighbor residues. Although Voronoi and Delaunay tessellations represent the same information or are completely determined by each other, Delaunay tessellations have the decisive advantage that 3D Delaunay simplices are always tetrahedral with four residues, whereas Voronoi polyhedra have variable numbers of faces and edges.

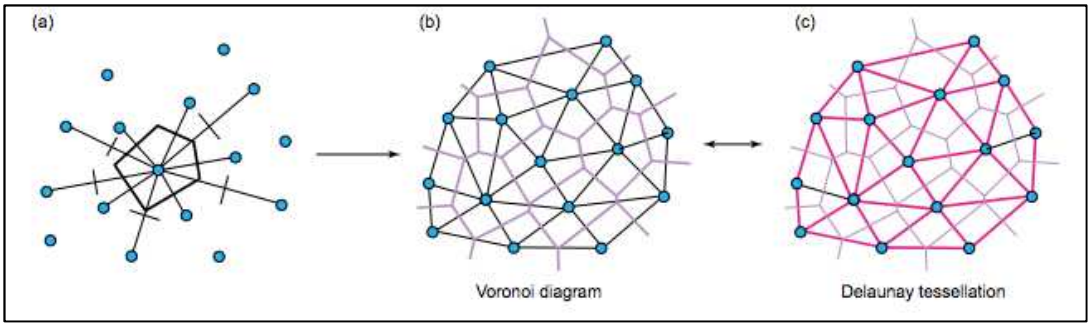


Figure 17: Voronoi diagram and Delaunay tessellation;
Voronoi polyhedron and Delaunay tessellation (a) The bisecting line between a given point and all the other points in the set is drawn, and then the smallest polyhedron defined by these lines is drawn; this is the Voronoi cell of the particular point. (b) When repeated for all points in the set, the Voronoi tessellation is obtained (thick purple lines). (c) The Delaunay tessellation of the set of points (thick red lines) is obtained by drawing all the segments between points that share a common Voronoi face. (Figure modified from Poupon [2004] [170])

Delaunay tessellation of the protein structure can be performed using the Quickhull algorithm [171] given the PDB coordinates file. The Quickhull algorithm produces a Delaunay tessellation by computing the convex-hull i.e. the smallest convex containing specific points, from a set of points.

Voronoi diagrams and their derived variants have been used often for the study of protein structures, protein-protein interactions, packing of protein core, packing at the interface with water, protein cavities, assessing the quality of protein crystal structure [170], [172]. Voronoi tessellations were first performed to estimate the density of globular proteins by calculating volumes around atoms in them. Volumes around atoms were used to analyze the packing geometry of residues inside proteins and to compare proteins [173]. In addition, void volumes may be calculated from Voronoi tessellations and reflect heterogeneous density of proteins. Chakravarty *et al.* [174] found that the creation of voids when mutations are introduced correlated with a decrease in protein stability. More recently Voronoi tessellations have also been used for discriminating thermophilic and mesophilic proteins [175], study the dynamics of the internal protein cavities [176] and also compare the level of intracellular organization between wild-type and mutant populations of cells supporting the notion that centrioles play a role in generating or maintaining global cellular organization [177].

Voronoi and Delaunay tessellations provide a framework for calculating empirical statistical potentials. Singh *et al.* derived five classes of every four

neighboring Cα atoms and developed a four-body potential to evaluate sequence-structure compatibility for solving the inverse protein folding problem [178]. This potential has been successfully tested for inverse protein folding [179], fold recognition [180], decoy structure discrimination [181][159], protein design [182], protein folding on a lattice [183], mutant stability studies [168], computational mutagenesis [184], protein structure similarity comparison [185], and protein structure classification [186].This statistical potential was further used in another study to predict stability changes in proteins caused by mutations [167]. Barenboim *et al.* [187] observed that the four-body statistical potential of polymorphic proteins with disease-associated nsSNPs (daSNPs) was on average significantly lower than the four-body statistical potential of the proteins with neutral SNPs (ntSNPs). topoSNP [121] is a Delaunay tessellation based method predicting disease association of nsSNPs. Multimutate is also a four-body scoring function based on Delaunay tessellation of proteins to predict the effects of single- and multiple-residue mutations on the stability and reactivity of proteins. Both these methods have been detailed in the "Overview of Predictive Tools" section.

Geometrical parameters of a tetrahedral

To analyze the correlations between the structure and sequence of proteins, Singh *et al.* [178] introduced a classification of Delaunay simplices based on the relative positions of vertex residues in the primary sequence. Two residues were defined as distant if they were separated by one or more residues

in the protein primary sequence. Simplices were divided in to five nonredundant classes and the differences between them were investigated by using geometrical paramets of tetrahedral, such as volume and tetrahedrality. class {4}, where all four residues in the simplex are consecutive in the protein primary sequence; class {3,1}, where three residues are consecutive and the fourth is a distant one; class {2,2}, where two pairs of consecutive residues are separated in the sequence; class {2,1,1}, where two residues are consecutive, and the other two are distant both from the first two and from each other; and class {1,1,1,1} where all four residues are distant from each other. They state that any particular protein usually contains all five classes.

Tetrahedrality is a measure of the degree of distortion of the Delaunay simplices in a given protein from the ideal tetrahedron and is calculated as follows:

$$T = \sum_{i>j}(l_i - l_j)^2 / 15\,\bar{l}^{\,2}$$

where l_i is the length of the i-th edge, and $\bar{l}$ is the mean length of the edges of the given simplex. The tetrahedrality of the different simplex classes was calculated for 103 protein chains with high crystallographic resolution, no apparent structure similarity, and low sequence homology. The distribution of the tetrahedrality shows that classes {4} and {2,2} have the lowest tetrahedrality, suggesting that these classes possibly occur in regular protein motifs such as alpha-helices. This hypothesis was verified by comparing the simplex classes with secondary structure assignment (helices, beta-strands, and coils) [178].

Statistical Potential

A *topological score* or *total potential* of a protein can be calculated based on the amino acid composition of the four nearest neighbor residues or quadruplets that are defined by the Delaunay tessellation. The maximum number of all possible quadruplets of natural amino acid residues is 8,855 assuming that the composition of Delaunay simplices is order independent and preference unbiased. Given a four-body potential function and the Delaunay tessellation of a protein structure of interest the *total potential* or *topological score* of the protein is obtained by summing the log-likelihood scores of all the amino acid quadruplets whose representative points form the vertices of tetrahedra in the tessellation. The log- likelihood for each quadruplet is defined as

$$q_{ijkl} = \log(f_{ijkl} / p_{ijkl})$$

where f_{ijkl} represents the frequency of quadruplets containing residues *i,j,k,l* in a representative training set of high-resolution protein structures with low primary sequence identity, and p_{ijkl} is the frequency of random occurrence of the quadruplet.

An *individual residue potential* or *residue environment score* is computed for each amino acid position by summing the log-likelihood scores of only the tetrahedral/simplices that use the representative point of the residue as a vertex, generating a *potential profile* for the protein [178], [188]. The topological scores

have been used for identifying tertiary packing motifs and functional signature motifs common to structures belonging to the same protein family [167].

Assuming minor structural differences and hence similar tessellations between a mutant (generated by amino acid substitutions at one or more positions) and the wild type protein, the total potential and potential profile of the mutant can be derived from the tessellation of the wild type structure. Missense mutations may cause structural rearrangements in proteins, which could give it a slightly different potential score than the wild-type protein. By altering amino acid identities at points representing residue positions, which have been mutated, the alternative quadruplet compositions of tetrahedra utilizing these points as vertices in the tessellation changes their log-likelihood scores and leads to a new total potential and potential profile for the mutant protein.

A scalar *residual score* as well as a vector *residual profile,* defined as the difference between the mutant and wild-type (wt) total potentials and potential profiles, respectively, characterize each protein mutant and can be utilized in predicting effects of mutations on the protein structure and function.

This approach was implemented by Carter *et al.* [168] to predict the stability of mutants. In a study by Vaisman and Masso [184, p. -1] a comprehensive mutational profile for HIV-1 protease was generated. This profile gave the mean potential score differences of all possible 19 substitutions at each residue position. Such profiles provide insight into the hydrophobicity of residues. The distribution of the residue potential scores indicates that low scores are

associated with surface residues, which have less structural neighbors, and high scores with residues in the hydrophobic core, which are structurally important for maintaining the conformation of a protein [168], [184, p. -]

Objectives and Dissertation Format

General areas in need of improvement for the development of better classification approaches seem to be: 1. experimentally validated and unbiased training sets of disease causing or deleterious mutations quantitatively phenotyped at both the organismal and molecular level; 2. identification and characterization of new mutational attributes beyond sequence composition, structure and evolutionary conservation) that improve classification accuracy and 3. new computational approaches need to be developed that improve classification accuracy when similar attributes and training sets are used. The following chapters in this dissertation shed light on all the three issues.

The introduction section describes elaborately the motive and reasoning behind this particular research and its advantages. It details the important role played by missense mutations in causing functional effects leading to different cancers, different efforts undertaken by many global scientific leaders in collecting data from cancer genomes and the variants identified in comprehensive databases and data warehouses and their data structures. It also elaborates the need to predict the functional impact of missense mutations in human cancers, describes the present state of art predictive methods being used to assess the

effects of the missense mutations. A comparative analysis based on various publications is also presented.

Chapter one details the creation of an integrated database of human cancer missense mutations linked to their 3D structures, which has been created with the whole motivation of building a one stop shop of human missense mutations data sets. The database IDHCMM is an integration of COSMIC, TCGA, ICGC, IARC TP53, BIC, and MSKCC sarcoma and prostate data. Huge sets of missense mutations can be downloaded from IDHCMM, which can be used for training or testing purposes in different predictive methods. These mutations have experimental data linked as downloaded from the source databases, which can be used to filtered the mutations from the IDHCMM database. These filters could include sample information from TCGA, Individual information from COSMIC, Tumor stage etc. The database and the related user interface are detailed in chapter one.

The research presented in Chapter two of this dissertation has an aim to systematically implement a structural geometrical approach for analyzing and assessing the effects of missense mutations on protein structure and function in human cancer genes. The approach is based on a potential score of a protein that is calculated based on the amino acid composition of the four nearest neighbor residues (quadruplets) defined by the Delaunay tessellation. The residual score quantifies the relative perturbation in sequence-structure compatibility of a mutant from that of the wild type protein. The potential profiles of wild-type and

mutants would be then analyzed using the residual scores associated with the full complement of single point mutants in order to study their ability to characterize the roles of amino acids in a protein.

A number of proteins, each with more than 100 known mutations from IDHCMM database (mentioned in chapter one) have been selected and tessellated. A critical assessment of the mutant residual scores can be performed based on the biological notion that protein structure determines function. The residual score of a mutant is a measure of the relative change in sequence-structure compatibility from the wild-type protein, therefore it is generally expected that the more negative the residual score, the less active the mutant. The annotated mutant protein systems will be examined for their residual scores, and a comprehensive statistical analysis will be performed to predict the functional effect of the mutants.

Finally the last chapter, Chapter Three presents a pilot study probing into the potential utility of mutated proteins as cancer biomarkers to be used for diagnosis, prognosis, and targeted therapy, on the basis of gene-expression profiles of the genes producing these proteins. Gene-expression technology has been used successfully to show that it can increase the specificity of the molecular classification of breast cancer [189]. In theory, the gene products resulting from somatic mutations are the ultimate protein biomarkers, not just being simply associated with tumors but actually responsible for tumorigenesis. Unlike other protein biomarkers such as carcinoembryonic antigen (CEA) or

prostate-specific antigen (PSA), which are associated with the tumors, the mutant proteins are produced only by tumor cells therefore make the ultimate protein biomarkers. In this pilot study we used machine learning approach to scan through the gene expression data of Prostate cancer in search for cancer biomarkers.

CHAPTER ONE:
Integrated Database Of Human Cancer Missense Mutations Linked To 3D Structures [IDHCMM]

Introduction

In recent years there has been an avalanche of mutational data from scientific literature and an equally enormous or may be even larger contribution from cancer genome sequencing projects. These sequencing projects produce enormous data, which are being analyzed to uncover the large number of genetic variants, which are of substantial interest in methods predicting their effect in oncogenesis. This data is dispersed in different databases at different locations and is specific and relevant to the research being done by the researchers generating this data.

The main aim of cancer research has been to identify all the genes involved in cancer prognosis and to identify the set/sets of mutations that occur in each of these genes. According to the Cancer Gene Census project led by the Cancer Genome Project, based at the Wellcome Trust Sanger Institute, 513 genes and 474 somatic mutations have been implicated in cancer (http://cancer.sanger.ac.uk/cancergenome/projects/census/). Many more have been discovered after that. These somatic variants need to be annotated accurately to relate them to disease susceptibility and get a clinically relevant

interpretation. This would further assist in formulating better diagnostic methods and treatments for cancer patients. Of particular interest are somatic missense mutations, which change the amino acid sequence of proteins thereby possibly affecting protein structure and function.

There are a lot of different comprehensive data sources, such as Online Mendelian Inheritance in Man (OMIM, Wheeler et al, 2004), HGVbase (Fredman et al. 2002) and the Human Gene Mutation Database (HGMD, Stenson *et al.* 2003). These databases are not specific to any one disease and carry information about the genetics and biology of different genes and diseases associated with those genes (OMIM). HGVbase carries information about genome variants and associated genotype – phenotype relationships. HGMD has data related to germline mutations (HGMD). There are some databases, which store somatic mutations in cancer, but most of them are locus-specific, such as the database for p53 (Olivier et al, 2002; Béroud and Soussi, 2003), BIC database for BRCA1 and BRCA2. Some of the largest and most comprehensive efforts towards understanding the molecular basis of cancer, including large-scale genome sequencing projects are TCGA (The Cancer Genome Atlas) and ICGC (The International Cancer Genome Consortium). The Catalogue of Somatic Mutations in Cancer (COSMIC) (http://www.sanger.ac.uk/cosmic) offers a comprehensive view of all previously reported somatic mutations in cancer and associated information.

Here we present an integrated database of cancer mutations, IDHCMM, comprising of only missense or non-synonymous mutations found in humans.

Motivation

The primary motivation of creating this database was to assist in developing models for predicting the effects of missense or non-synonymous mutations on protein structure and thereby the protein function, by making available large number of missense mutations per gene, in one place. Missense mutations have been reported to have direct implications in structure and functional impacts on proteins involved in cancer [22][23][29][36]. It is still a significant challenge and is being seriously pursued to identify driver versus passenger mutations [40]. When it is not clear which of the mutations are driver mutations or passenger, a reasonable and an intuitive approach is to have a set of recurrent or overrepresented mutations in genes, i.e. each gene having at least a 100 mutations or in other words some positions are mutated recurrently i.e. 2 or more times. This kind of a data set has been shown to support prediction assessment studies performed by Gnad *et al.* [40]. This being one of the motivations, creating IDHCMM also represents an effort to centralize different aspects of informational data associated to the missense mutations such as the molecular, histological, clinical, sequencing and analytical data linked with mutations, lodged in all the different databases. It mainly proves valuable in providing 3D structure data from PDB linked to the mutations. All the source databases mentioned above have some mutually exclusive data, which is very

valuable and could be used by many researchers in many different scientific projects. A central repository such as IDHCMM could accelerate, facilitate and ease the research by providing all the data at one place. It would not only be valuable for the profound scientific community but also to novice researchers and translational research members who would not need to understand the complexity of cancer genomics.

Database Construction And Content

The approach we use here to generate models to predict the effects of mutation on proteins is protein structure based and so more the number of mutations mapped on to a gene or protein structure more accurate and reliable the predictive model could be. Our aim was to find proteins, which have 100 or more mutations or in other words proteins where some positions are mutated recurrently i.e. 2 or more times. To achieve this target we integrated open access mutational data from some of the most comprehensive and huge databases namely, TCGA, ICGC, COSMIC. We also included data from other extensively studied and widely used single –gene databases such as IARC TP53 database [1], Breast Cancer Information Core database.

In an attempt to understand and further look into possible inferences from gene data, which include mutational data as well as Gene Expression data, two of the Memorial Sloan-Kettering projects, The Sarcoma Genome Project and Prostate Cancer data sets, were included into our database IDHCMM. Besides integrating

data from six different mutational data sources, IDHCMM also features integration with UniProt and PDB databases.

Data from source databases was filtered for only Missense or Non-synonymous mutations and downloaded. Data from different sources was then mapped in order to integrate into one database. In this processing, duplicate or overlapping entities rooting from different sources were unified based on data definitions obtained from the parent/source databases for each of the data elements in the database. These data elements were then mapped to IDHCMM data elements. Supplementary Document 02 shows the mapping of the data elements from different parent/source databases and IDHCMM. The data flow of IDCHMM is shown in Figure1.

IDHCMM contains more than 1.48 million records, which comprise 215374 distinct Amino Acid Mutations (Missense/Non-Synonymous), 59182 distinct Genes and 684892 distinct 'Gene-Missense Mutation' pairs. Table 1 shows the database statistics by each data source. The database has a total of 25 tables including the source tables and ID mapping tables and some secondary tables created to store filtered data derived from primary tables.

System Design And Implementation

Data sources

Cancer research and studies based on gene mutations have been made possible by a large number of Mutational databases ranging from single gene databases,

single cancer databases, multiple or related gene databases to complex databases which not only include mutational information but also link the mutational data to many other scientifically relevant aspects like genotype-phenotype associations, disease associations, pathways, population statistics etc. However there is scarcity of databases with clinical data linked to the mutations. This requires studies and projects to collect, store, monitor data and continuously follow and record changes at different levels such as bio-specimen and patient data collection, genomic data characterization, sequencing and analysis and then proteomic characterization and analysis. We selected some of the databases, which are huge comprehensive projects and contain a wide spectrum of data namely The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC) and Catalogue of Somatic Mutations in Cancer (COSMIC).

TCGA

The Cancer Genome Atlas is a coordinated project established by NCI and NHGRI completely focusing on understanding the molecular basis of cancer utilizing latest genome sequencing and analysis technologies. All information about TCGA and the TCGA research network can be found at **http://cancergenome.nih.gov/**.

TCGA provides a platform, called the “Data Portal”, for researchers enabling easy download and analysis of data generated by TCGA. All the TCGA data in IDHCMM has been downloaded from this data portal using the “Bulk

Download" process, after confirming with the TCGA helpdesk. The mutational data is present in special files ending with ".MAF". All available '.MAF' files as of date Jan 10th 2013 were downloaded. The data downloaded from TCGA data portal includes data generated by the following institutes:

1. The Broad Institute at MIT
2. The Genome Institute at Washington University
3. The Baylor College of Medicine Human Genome Sequencing Center

The downloaded data pertain to the following cancers:

1. Bladder Urothelial Carcinoma
2. Cervical squamous cell carcinoma and endocervical adenocarcinoma
3. Glioblastoma multiforme
4. Head and Neck squamous cell carcinoma
5. Kidney renal clear cell carcinoma
6. Lung adenocarcinoma
7. Lung squamous cell carcinoma
8. Ovarian serous cystadenocarcinoma
9. Prostate adenocarcinoma
10. Skin Cutaneous Melanoma
11. Thyroid carcinoma
12. Uterine Corpus Endometrial Carcinoma
13. Breast invasive carcinoma
14. Acute Myeloid Leukemia

15. Colon adenocarcinoma
16. Rectum adenocarcinoma

The DNA sequencing technologies used in are:

1. Illumina (Genome Analyzer)
2. ABI SOLiD sequencing

The data from downloaded '.maf' files was then filtered for somatic missense mutations by selecting only those rows from files with a 'Mutation status' as 'Somatic' and 'Variant Classification' as 'Missense'. These filtered files were then converted into '.csv' files to be used to upload data into IDHCMM, an Integrated Database of Human Cancer Missense Mutations.

Along with mutation files, definitions of each of the data elements used in the TCGA database was acquired from their websites and via communicating with the Database support/helpdesk when ever required. These were needed for further mapping of these data elements with the data elements from other data sources like COSMIC and ICGC etc.

Problems encountered with TCGA data:

1. The data shows more data elements/fields or columns in the files other than those listed and defined in the database documentation. For ex: TCGA has 34 data elements listed on their wiki at https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+%28MAF%29+Specification, however, the .maf data files downloaded have about 100 more fields of data. The TCGA helpdesk was contacted and requested to

provide data definitions. After a couple of emails and information exchange, data definitions for some more data elements, identified to be from 'Oncotator' (http://www.broadinstitute.org/oncotator/) were provided. However these did not necessarily cover all the data elements listed in the '.maf' files. Some of them are still undefined, and have not been used in IDHCMM.

2. Not all mutational data could be downloaded from TCGA download portal.
    a. Only 31 files with mutational data could be downloaded via the search engine.
    b. There are about 76 more mutational data files, which can be accessed programmatically via their web-services from their file system. However, accessing these files from file system has some limitations:
        i. The web-service lists some filters, which need to be used to download data, however it doesn't list filters appropriate to download only mutational data.
        ii. Huge data files will need to be downloaded first and the mutational data files will have to be searched from this set.
        iii. The web-service also limits the searches based on cancer type, center, level and platform used, which will require as many requests as there could be values of the above mandatory filters.

Data from only the 31 files, which could be downloaded from the TCGA data portal, open access, were used in this project. Data elements from each of the mutation files were mapped to data elements in IDHCMM and the data uploaded to the database. This is discussed further in the 'Data Mapping' section.

ICGC

ICGC, *The International Cancer Genome Consortium*, coordinates a large number of projects with a common goal of unraveling and generating comprehensive catalogues of genomic abnormalities, which include somatic mutations, abnormal expression of genes, epigenetic modifications, in different forms of cancer. All information about ICGC and the ICGC Cancer Genome Projects can be found at http://icgc.org/icgc. ICGC has both open access data and controlled data. To access and download controlled data researchers need to apply and get permission. Here we have used open access data only. ICGC provides access to its open access data via a data portal, which provides tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.

Data from ICGC projects was downloaded in November of 2012, of Release 10. The FTP site of the ICGC data portal was used to download all the mutational data files. This step was performed for each of the cancer type provided. There were 25 cancer types, which had mutation files available on the FTP site.

Following is the list of the cancer datasets for which mutation files were downloaded:

1. Acute Myeloid Leukemia TCGA US
2. Breast Carcinoma-WTSI UK
3. Breast Invasive Carcinoma-TCGA US
4. Breast Cancer-JHU US
5. Chronic Lymphocytic Leukemia-ISC-MICINN ES
6. Colon Adenocarcinoma-TCGA US
7. Colorectal Cancer-JHU US
8. Gastric Cancer-CCGC CN
9. Glioblastoma Multiforme-JHU US
10. Glioblastoma Multiforme-TCGA US
11. Liver Cancer-INCA FR
12. Liver Cancer-NCC JP
13. Liver Cancer-RIKEN JP
14. Lung Adenocarcinoma-TSP US
15. Lung Squamous Cell Carcinoma-TCGA US
16. Malignant Melanoma-WTSI UK
17. Myeloproliferative Disorders-WTSI UK
18. Ovarian Serous Cystadenocarcinoma-TCGA US
19. Pancreatic Cancer-JHU US
20. Pancreatic Cancer-OICR CA

21. Pancreatic Cancer-QCMG AU
22. Pediatric Brain Tumors-DKFZ DE
23. Rectum Adenocarcinoma-TCGA US
24. Small Cell Lung Carcinoma-WTSI UK
25. Uterine Corpus Endometrioid Carcinoma-TCGA US

Each dataset name listed above includes the name of the cancer type, followed by the name of the Institute where the data was generated and then the country where the institute is located. A list of the institutes where the data is coming from is as follows:

1. TCGA: *The Cancer Genome Atlas, USA*
2. WTSI: *The Wellcome Trust Sanger Institute, UK*
3. DKFZ: *The German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ), Germany*
4. QCMG: *Queensland Centre for Medical Genomics, Australia*
5. *OICR: The Ontario Institute for Cancer Research, Canada*
6. JHU US: *John Hopkins University, USA*
7. TSP: *Tumor Sequencing Project, USA*
8. RIKEN: *National Institute of* Biomedical Innovation, *Japan*
9. NCC: *National Cancer Center, Japan*
10. INCA: *The French National Cancer Institute (Institut National du Cancer), France*
11. CCGC: *The Chinese Cancer Genome Consortium, China*

12. ISC-MICINN: *Spanish consortium, Spain*

At the time of download there were some cancer datasets which did not have the mutations data available, these are listed below:

1. Kidney Renal Clear Cell Carcinoma-TCGA, US
2. Lung Adenocarcinoma-TCGA, US
3. Cervical Squamous Cell Carcinoma-TCGA, US
4. Bladder Urothelial Carcinoma-TCGA, US

Mutational data from these data sets and more newly sequenced tumors could be added as an update to IDHCMM as future perspective.

The ftp site provided all the mutations in files with file names prefixed with 'ssm', which means Simple somatic mutations. Once downloaded these files were filtered for data with a 'Consequence type' as 'Non-synonymous coding'. The option of 'Missense Mutations' was not provided in that release of ICGC data. Then they were converted to '.csv' files for upload into IDHCMM. Data definitions of all the data elements/fields used in ICGC was acquired from the ICGC data submission manual. Data elements from each of the mutation files were mapped to data elements in IDHCMM based on the data definitions from the manual and the data was uploaded to the database. This is discussed further in the 'Data Mapping' section.

COSMIC

COSMIC, *Catalogue of Somatic Mutations in Cancer*, stores and displays information pertaining to somatic mutations and related details, especially human cancers. It also contains information about publications and samples. The mutation data and associated data are extracted from literature. COSMIC provides 'BioMart', to help users download data. It provides a set of filters and attributes to select the kind of data a user needs to download.

Most recent download of data from COSMIC was made on May 12th 2013 to IDHCMM and was from version 61. All the downloaded mutations were verified, confirmed somatic variants and were all missense mutations. At the time of download, COSMIC did not provide definitions for all the data elements. Cosmic helpdesk was contacted and requested to provide the definitions, which would be needed for mapping the data elements into IDHCMM. The definitions were provided by email upon request.

BIC

Breast Cancer Information Core, is an open access on-line breast cancer mutation database. It is an international collaborative effort hosted by NHGRI. BIC catalogues all the mutations and polymorphisms in breast cancer susceptibility genes, BRCA1 and BRCA2. In addition to mutation information the database contains a collection of mutation detection protocols, lists of gene specific DNA primers and published protocols. Mutation data is entered in this

database by individual investigators, hospital-based labs and a commercial lab performing the bulk of BRCA1/BRCA2 tests in North America.

The latest download of data was on May 10th 2013 and filtered for the 'Mutation type' 'M' (for missense) as described in the BIC database glossary. As of the date of download there were 4506 missense mutations in the dataset. Definitions for data elements were retrieved from the BIC database glossary (http://research.nhgri.nih.gov/projects/bic/Member/glossary.shtml).

IARC TP53

The IARC TP53 mutation database compiles all TP53 mutations that have been reported in the published literature since 1989. This database is updated every year. IARC TP53 database includes various annotations on the predicted or experimentally assessed functional impact of mutations, clinicopathologic characteristics of tumors and demographic and life-style information on patients. It provides the following datasets:

1. TP53 somatic mutations in sporadic cancers
2. TP53 germline mutation in familial cancers
3. Common TP53 polymorphisms identified in human populations
4. Functional and structural properties of p53 mutant proteins
5. TP53 gene status in human cell-lines
6. Mouse-models with engineered TP53

The IARC TP53 Database is a free service offered to the scientific community. Data pertaining only to TP53 somatic mutations, was downloaded for use in IDHCMM. The latest download was done on May 12 2013 from the release 'R16'. It was then filtered (by column 'Effect') to get only missense mutations into IDHCMM. At the time of download it had 21,614 missense mutations. Data definitions of all the data elements used in the IARC TP53 database were retrieved from the user's manual provided on the database website (http://p53.iarc.fr/Manual.aspx).

MSKCC

A major focus at the Computational Biology Center at Memorial Sloan- Kettering Cancer Center is the integrative analysis of cancer genomics data sets. The methods developed span analysis of mutations, identification of recurrent DNA copy-number alterations, and the identification of altered signaling pathways. Two of the major projects undertaken by MSKCC for these studies are:

1. Integrative genomic profiling of human prostate cancer
    a. Includes integrated genomic profiling of 218 prostate tumors.
    b. Eighty tumors were examined for somatic mutations in 138 genes by exon sequencing, these and an additional 76 tumors were also profiled for well-known oncogenic mutations in 22 genes by mass spectrometry using the iPLEX Sequenom assay. In total, 84 confirmed somatic mutations were detected in 57

different genes. Thirty-seven percent of the missense mutations detected were predicted to affect protein function based on an algorithm, Mutation Assessor, which uses a combination of evolutionary information from protein-family sequence alignments and residue placement in known or homology-deduced three-dimensional protein and complex structures.

2. The Sarcoma Genome Project: A collaboration between Memorial Sloan-Kettering Cancer Center and The Broad Institute.
    a. It is an integrative analysis of DNA sequence, copy number and mRNA expression in 207 samples encompassing seven major subtypes.
    b. It includes detailed map of molecular alterations across diverse sarcoma subtypes.

Data from MSKCC projects was included with an intention to study gene expression along with the mutational data as these data sets have both mutational data s well as gene expression data for all the genes included. The mutations were available for download for Prostate cancer and were downloaded on November 3rd 2012 and then filtered for 'Mutation type', 'Missense'. The Prostate cancer data had 254 missense mutations as of the date of download. Mutational data for the Sarcoma data set were provided by the MSKCC helpdesk upon request made after being not able to download using the CBio data portal. These mutations were also filtered for missense mutations. This dataset had 28

missense mutations. Data definitions for the data elements provided in the Sarcoma data set were all the same as the TCGA data set. Definitions for data elements for the prostate cancer were acquired from http://mutationassessor.org/howitworks.php

Table 3: Data sources and statistics

	Data Sources							
	TCGA	ICGC	COSMIC	BIC		IARC TP53	MSKCC	
				BRCA1	BRCA2		Prostate	Sarcoma
Total Number of records	625,871	690,044	136,258	4,505	7,135	21,613	254	28
Total no of Distinct AA Mutations	135,691	12,021	65,732	135	141	1,435	193	26
Total no of Distinct Genes	21,450	18,889	18,711	1	1	1	112	17
Total no of Distinct Gene and AA-Mutation pairs	385,633	181,885	115,443	135	141	1,435	194	26
Date of source data download	Oct' 2012	Nov'2012	May 2013	May 2013	May 2013	May 2013	Jan 2013	Jan 2013

Data Integration/Data Mapping

The computational mutagenesis method, which we intend to use, is based on the application of a Delaunay tessellation-derived four-body statistical potential function. Since the potential is derived via an approach that utilizes the atomic coordinates of non-homologous, high-resolution protein structures, the computational mutagenesis incorporates information about both sequence and structure. Using this methodology, every single or multiple mutant of a protein can be characterized by a scalar residual score, which measures the relative change in overall sequence-structure compatibility from wild-type, as well as a vector residual profile, which quantifies environmental perturbations from wild-type at every amino acid position. The main objective in creating an integrated database was to pool as many mutations per gene as we could, to study these effects of mutations on the structure of each protein and eventually the protein's function.

Each of the missense mutations and all associated details obtained from the six different data sources needed to be pooled together to form one central repository. In order to achieve this goal, all the data elements from all the data sources needed to be compared and mapped to each other. The data from main sources was filtered, cleaned, transformed when and wherever needed and then catalogued such that it can be searched and used for analytical processing further. Searching and extracting data from this integrated database/data

warehouse is being enabled via a webpage and is explained in detail in the 'User interface' section. The biggest advantage of integrating mutational data from all these different databases it to be able to get a large number of amino acid mutations for each gene, which would not be the same with any of the individual databases. More number of mutations for each gene would definitely provide a greater perspective in structure-based studies like ours where the effect of the mutations on the structure and the functional impact on the protein is being studied. Secondly, large sets of mutations per gene can eliminate the need of computational mutagenesis also adding the fact that all the mutations are real and have been collected from literature or directly from experiments. Apart from these, data from our database also includes associated data from all data sources for each of the missense mutations. Associated data could be like data from:

- Source databases, such as the source database mutation id ex: COSMIC ID, source database version, Source database sample ID etc.
- The gene ex: Gene IDs (including Entrez, Ensemble, RefSeq, Transcript ID), Gene Name etc.
- Mutation specifics such as CDS mutation Position, AA-mutation position, Mutation Validation status etc.
- Predictions from SIFT and PolyPhen
- UniProt Ids
- PDB structure data including Resolution, chain, missing residue information etc.

- Clinical data from source databases wherever available, such as Tumor grade, Tumor stage, Histology.
- Experimental details such as Sample and Specimen data
- It also incorporates open access individual data available from source databases like gender, age at diagnosis, ethnicity, Family history, Country, Exposures (Tobacco, Alcohol) etc.

A list of all the data elements used in IDHCMM, and their definitions are listed in Supplementary Document 02.

However, it is important to mention here that because IDHCMM is an integrated database, not all the fields would be populated, that is each mutation would not have information against each of the data element in IDHCMM. This is mainly because IDHCMM is designed to represent a data warehouse, storing all data from all the selected source databases. There are a lot of data elements in each of the data sources, which do not map to any of the data elements from other sources. Such data elements have been retained in order to keep the data but cannot be merged with any other data from any other data source. In such cases these data elements in IDHCMM will carry data only from that particular data source. Also, the availability of data values in IDHCMM is solely based on their presence in the source databases.

Data Mapping: Level1

Data mapping here involved creating a mapping between the six different data sources. This was the first step towards data integration. Each source may

contain incorrect data values and the data in the sources may be represented differently, overlap or contradict. This is because the sources are developed and maintained independently to serve specific needs. This results in a large degree of heterogeneity in data management systems, data models, schema designs and the actual data. Mapping was required to identify and consolidate redundant columns of data/data elements and help in creating a distinct list of columns or data elements for the integrated database IDHCMM. The mapping approach was data-driven and included both schema matching as well as data transformations wherever needed.

At the schema level, differences in schema design were addressed by schema translation and schema integration. The main problem with schema design was naming conflict. Naming conflicts arise when the same name is used for different data elements (homonyms) or different names are used for the same object (synonyms). All the six data sources had different representations or definitions for their data elements, which actually could be representing the same information/data. Presence of this kind of heterogeneity called for an ETL process (Extract, Transform and Load) involving Schema mapping and Data Transformations. This was needed to identify if the data elements being compared are semantically or structurally related or may be even are the same. For example, TCGA data and MSKCC-Sarcoma data had the gene name represented by the data element 'Hugo Symbol', and other databases used 'Gene Name'. In some cases this schema mapping was complex and required one-to-

many mappings. For example, different TCGA data files have the column names 'Protein_Change', 'amino_acid_change', and 'AAChange', all for the same information, the amino acid mutation. Schema mapping between data elements from the six data sources was done manually and is detailed in the Supplementary Document 02.

In addition to schema level conflicts, there were conflicts at the instance level where even when there are same names for data elements, there were different value representations. Data transformations were performed where a set of values from the source databases had to be converted into a unified format followed in IDHCMM. For example, the amino acid mutation format was found to be different in different data sources. COSMIC represents an amino acid mutation as 'p.G12D' where as some ICGC data files represent it as 'D>G' and data from MSKCC represent it as 'A212T'. List of all the data elements in IDHCMM and their definitions can be found in Supplementary Document 03. Once the mapping was done, it was implemented into the database system using sql scripts to generate one huge central repository of all the mutations from all the source databases. The software used is mentioned the 'Technology Stack' section.

Integrated Database/Data warehouse Design

All the data from all source databases was initially loaded into IDHCMM as source tables. Each entry from each source was given an IDHCMM identifier.

These source tables also included the source database identifiers and the source version/release numbers such that the data values could anytime be referred back to the source databases. With an aim to bring all the data together into a standard homogenous format, data from all the sources tables was pooled into one big table. This table is the first and the main IDHCMM table 'All Mutations' containing mutations from all the sources and can be regarded as the parent table. This table has the data elements designed for IDHCMM, to form a comprehensive list, which could accommodate most of the data elements from all the six sources. The most important task while pooling the data was to establish data compatibility. Data mapping was done prior to this step and was followed to map and load data from different sources into this parent table. The parent 'All Mutations' table retained the IDHCMM identifiers allotted to each entry as well as the source database identifiers along with all the associated data.

The 'All Mutations' table at the time of writing this manuscript contains 1,485,708 records. This table was then cleaned and filtered to retain useful and meaningful data. Once the cleaning was done, a number of other tables were derived/generated from 'All Mutations' in order to get to only the transformed, valid and complete mutation records. After this, level 2 mapping (Section Data Mapping: Level2) was performed. This involved mapping the mutation records to UniProt Ids and then using UniProt Ids to map them to the protein structure i.e. the PDB ids. Figure 1 illustrates the data workflow overview.

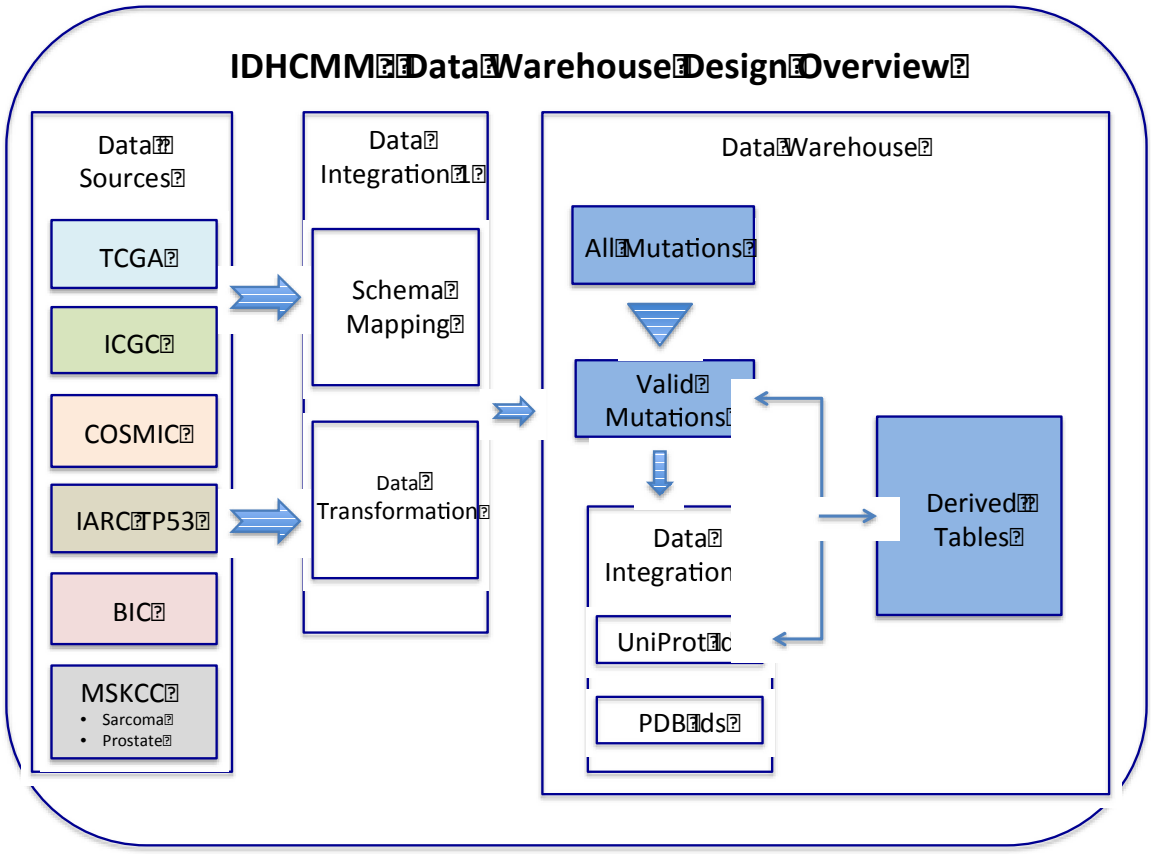


Figure 18 Data Work-Flow Overview

Data Mapping: Level2

With an aim to link the mutations to their protein structures, these mutations were mapped/ matched over a couple of different databases establishing a link via their Ids. All the valid mutations which had associated database ids such as Entrez Gene ID, Ensembl Gene ID, Refseq ID, Transcript ID, Swissprot AC ID or Swissprot Entry ID were extracted and Uniprot Ids for all these mutation records

were acquired using the ID mapping programmatic access service available at the Uniprot website (http://www.uniprot.org/faq/28).

This mapping to Uniprot ids was saved in a different table in IDHCMM and then used to acquire PDB Ids for each of the Uniprot ids also using the ID mapping programmatic access service available at the Uniprot website. This procedure acquired single or multiple PDB ids for each of the mutation records. A new table 'GeneTOStructure' was created which connects the mutation records via their PDB id to their structure information which includes the PDB ID, Structure Resolution, Protein Chain and Structure start position and end position (fragments).

As a next step, information regarding missing residues from the PDB database for each of these PDB ids was also downloaded and then the Mutation position was checked against the missing residues for each of the mutation against each of the structure associated with this mutation. This was done to make sure the mutation falls with in the protein structure. All the genes, which had the mutation with in the structure fragments were filtered and put into a new table 'MutInStructure' for ease of access to this data.

Material And Methods

Technology Stack

An overview of the multi-layer architecture of IDHCMM is represented in Figure2. The software used in creating IDHCMM is detailed below.

1. Database Components: The database was built on MySQL server, version 14.14_5.5.9. MySQL Workbench was used to administer (design, generate and manage) the IDHCMM database.
2. Java (version 1.6.0_65) was used as the programming environment.
3. Tomcat Application Server 6.0.37 was used as web application deployment server.
4. Eclipse IDE (Kepler Release 4.3) was used as development tool. Maven (m2e-wtp) Plugin for Eclipse was used for compile, build and deployment of web application.
5. Web Application Components:
    - Java class files
    - JSP/HTML web pages
    - BlueTrip CSS Framework (Style Sheets)
    - DisplayTag Library 1.2 (Table display with sorting, paging and export features)
    - JQuery 1.8.3
    - JQuery UI 1.9.2
    - MySQL/Java Connector 5.1.24
    - JUnit 3.8.1 (Unit testing)
    - Apache Log4J 1.6.5 (Logging)
    - Apache Commons DBUtils (JDBC Helper Library)
    - Apache Commons DBCP (Database Connection Pooling Services)

- Apache Commons Configuration (Reading of configuration files)

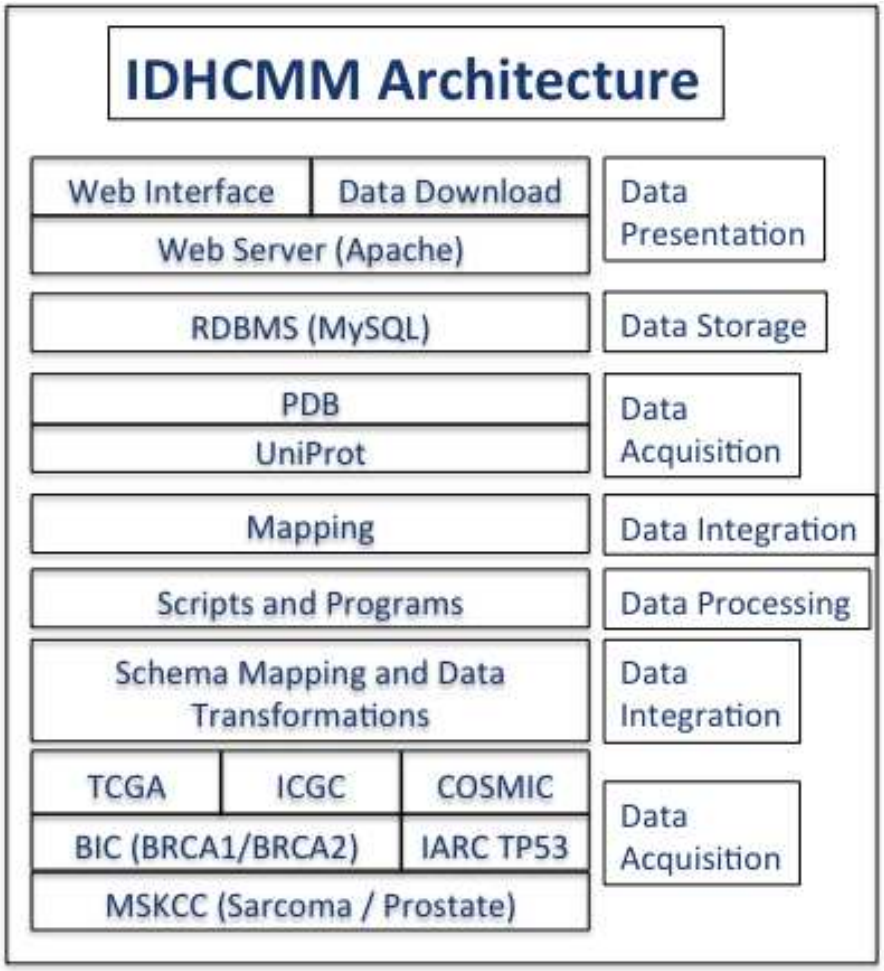


Figure 19: Multi-Layer Architecture of IDHCMM

User Interface

To facilitate the access and use of IDHCMM resource we implemented a simple web interface for the users. The website was implemented in Java and provides diverse query options. Users can query, retrieve and download data from

IDHCMM. Java Scripts were linked to the MySQL server to query and retrieve the data for webpage queries.

The search page provides an interface for querying the IDHCMM database with varied options, which include aspects of ‘Gene’, ‘Mutation’, ‘Protein Structure’, and ‘IDs’. ‘ Functional Impact Predictions’ and ‘Clinical Information’.

User Interface (UI) can be divided in the four main components:

1. Login: Login UI provides a web page to enter user credentials. If the user credentials are not verified, then the user is not allowed to perform search. Upon successfully validating user credentials, user is redirected to the home page where the search criteria can be entered.
2. Search Criteria: Search Criteria UI provides different search options along with options to select the required output columns.
    - Input Criteria: The main input options users can perform a search based on, are Gene Names, PDB Ids, Amino Acid Mutations and UniProt Ids. The search input text can be comma separated or in the form of a list with each input in new line.
    - Output Criteria: There are six categories of output criteria for users to selects options from. The categories and the options are listed below.
        - Gene Information: Gene Name, Entrez Gene ID, Ensembl Gene ID, Chromosome

- Mutation Information: CDS Mutation, CDS Mutation Type, CDS Mutation Start, CDS Mutation Stop, AA Mutation, AA Mutation Type, AA Mutation Start, AA Mutation Stop, Mutation Validation Status, Mutation Detection Platform, Mutation Validation Platform.
- Structure Information: PDB ID, PDB Fragment, Resolution, Chains, Missing residues, Domain Affected.
- Database IDs: Refseq ID, Transcript ID, UniProtKB AC, PubMed ID, Pfam Accession ID, neXtProt ID, PharmaGKB.
- Functional Impact predictions: Mutation Assessor Prediction, Polyphen Prediction, SIFT Prediction.
- Clinical Information has four sub-categories:
    - Sample: Sample Name, Sample Type, Sample ID, Sample Source, Sample Source, Specimen ID, Specimen Type, and Depositor.
    - Tumor: Tumor Source, Tumor Depth, Tumor Confirmed, Tumor Grade, Control Genotype, Tumor Genotype, Tumor Stage, Tumor Origin, ICD-10, TNM, p53 IHC, Morphology, Short Topology, Primary Site, Primary Histology, Site Subtype 1, Site Subtype 2, Site Subtype 3, Histology Subtype 1, Histology Subtype 2, Histology Subtype 3

- Individual: Ethnicity, Gender, Age At Diagnosis, Family History, Tobacco, Alcohol, Exposure, Geo Area, Country, Population.
- Other: Drug Target, Specimen Donor Treatment Type, and Contact Person.

3. Search Results: Search Results UI provides the search results in columns along with the source database name and input criteria provided by the user (gene, pdb_id etc.). The UI provides options to sort the results by any column and export the results in excel, pdf and xml formats.
4. Help/Contact pages: Help page provides help, other supplementary data files and other required resources or reference links. The contact page provides contact information.

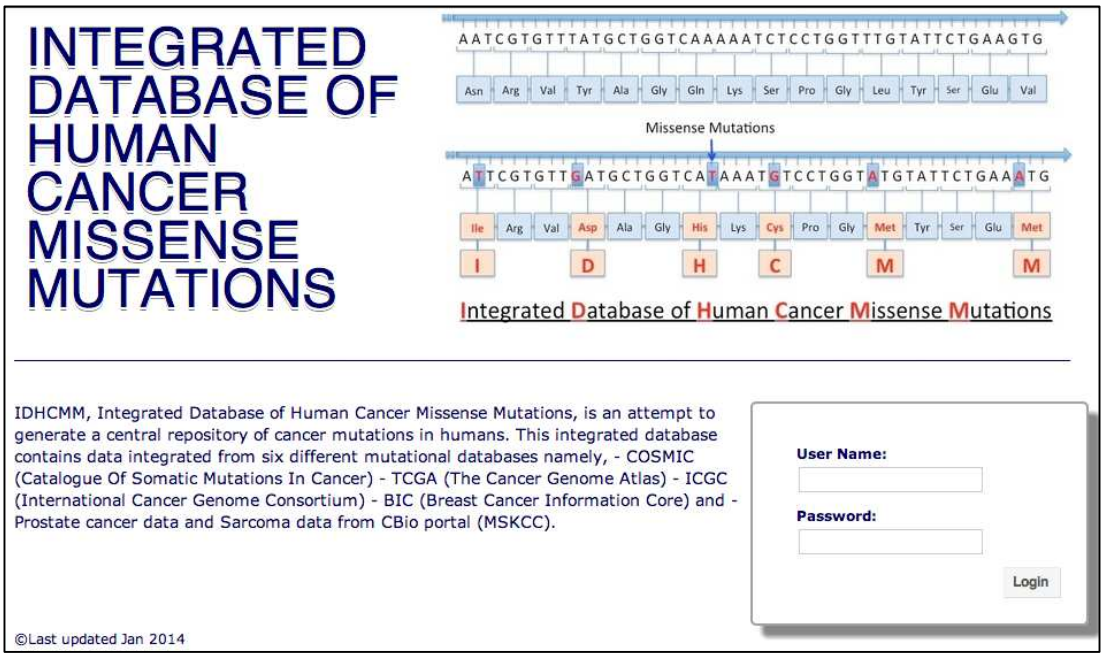


Figure 20 IDHCMM Login Page

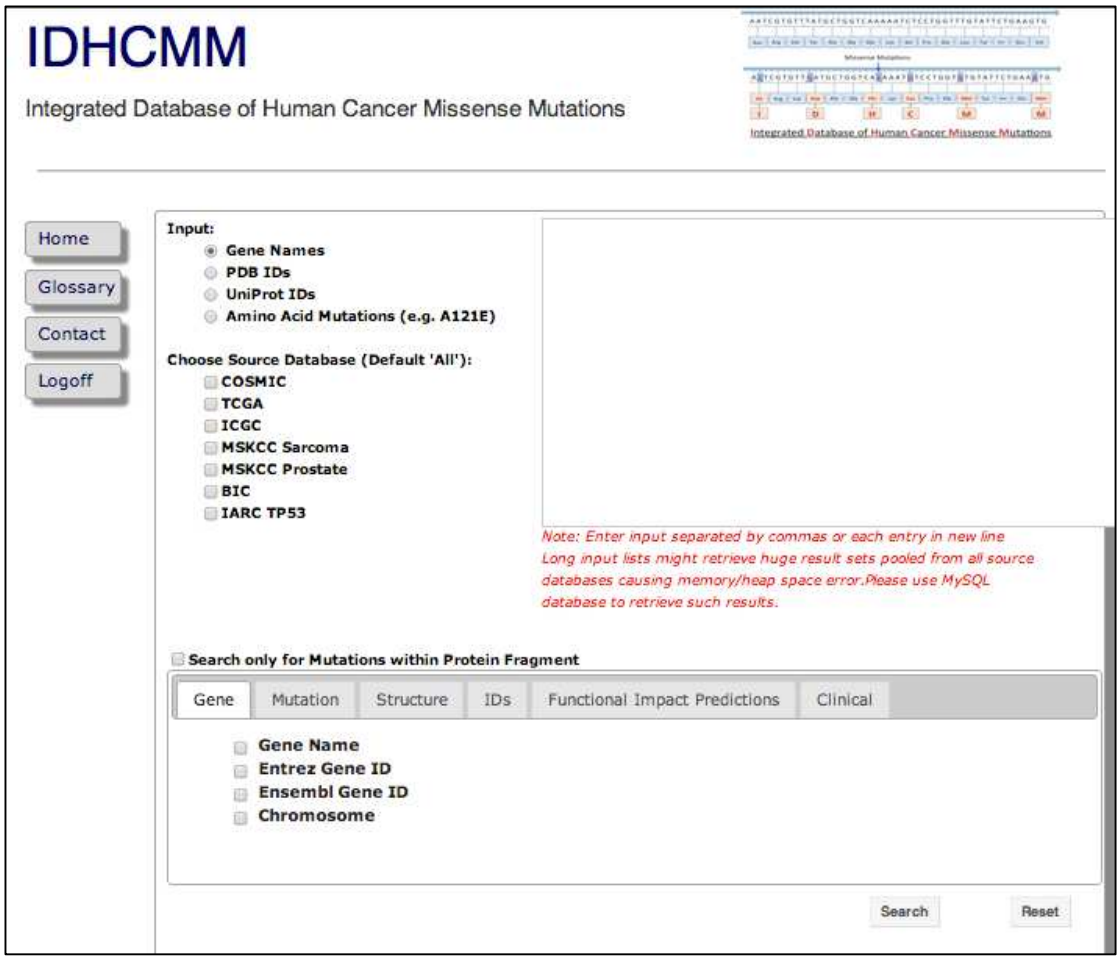


Figure 21 IDHCMM Search Page

Availability and requirements

IDHCMM can be easily ported to different systems. The database is available as a MySQL dump and can be requested. The webpage can also be ported along with the MySQL dump. All of the database and webpage development has been done on Mac OSX 10.7.5 with 4GB memory, however it is preferred to run the database and the web service on a machine with a higher memory.

Future Improvements

IDHCMM is a result of a basic integration of different comprehensive and important mutation databases. There is a wide scope of improvement and enhancement from this point. This project was done as a part of a doctoral dissertation and out of a need for a big set of mutational data as such hence has the minimum needed integration done. Future improvements can make this database more useful and fortify it multi dimensionally. Some of the suggested future enhancements could be as follows:

- IDHCMM presently contains only public data from different listed data sources. Protected or restricted data can be requested and added to IDHCMM. The user interface has a user validation feature which will secure the private data.
- Records in IDHCMM can be linked to their source database pages, for accessing the information at the source website.
- More data sources can be added, such as SNP500Cancer.
- More clinical data other than that provided by just the data sources can be included into IDHCMM.
- More data from PDB files can be extracted such as functionally relevant data and can be used in grouping or classifying mutations accordingly.
- More dimensions like Gene Expression, can be linked to the mutational data in IDHCMM. A preliminary effort in this direction has been performed and is presented as a part of this dissertation.

- The mutations from different source databases have been mapped to structure based on the presence of identification such as Entrez Gene Id, Ensembl Gene ID or RefSeq Id. Not all records essentially have an identification associated with them. If each record had an identification attached, the final draft of the database would be much more populated. This can be checked periodically. As and when the source databases are updated IDHCMM can be updated and mapping done to integrate data for these 'orphan' records. Also, as and when UniProt and PDB get updated, more and more genes can be linked to their structures.
- There are a huge number of records, which have important data missing, for example the mutation position in case of Amino acid mutation information ex, from the source database ICGC. This missing data lead to filtering out of a lot of mutational records, which actually have very valuable clinical information associated to them. In some of such cases, the position of the mutation is specified at the DNA level, which can be converted into an Amino acid position. This will also help in adding to the final draft of the database.
- Fucntional impact predictions are retrieved from source MSKCC data. PolyPhen and SIFT prediction are retrieved for TCGA data. These predictions can be got for all other mutations in the database by running the predictors.
- Search functionality can be altered such that users can retrieve just by specifying features and not giving any specific input lists such that they can get all records with data in the specified fields from IDHCMM .

CHAPTER TWO:
Delaunay Tessellation based models predicting effects of missense mutations in cancer proteins

Introduction

Scientific developments in gene sequencing and analysis technologies are at the next step of the sequence to function cascade where it is time now to study the effects of mutations in the sequence on structure and function of a protein and then apply it to clinical implications. The rapid progress of sequencing technologies has generated a torrent of mutational information along with all other kinds of data, from not only normal human genomes but also specifically from disease-associated genomes, cancer genomes being the foremost. As described in the introductory chapter of this dissertation missense mutations have taken the center stage in predicting the functional impact on proteins involved in cancer. A lot of predictive methods have already been developed which have undertaken the task of predicting the effects of missense mutations on protein functions (described in detail in the Introduction chapter). Many of these approaches rely on the knowledge derived from the analysis of significant spatial and compositional patterns in known protein sequences and structures and understanding of the role these patterns play in the extremely complex processes, like protein folding or protein function. Especially for structural

patters such an analysis requires an objective definition of nearest neighbor residues. Statistical geometry methods can define the nearest neighbor atoms or groups of atoms and identify them by statistical analysis of irregular polyhedra obtained as a result of a specific tessellation in three-dimensional space.

Materials and Methods

The objective of this study is to model the disease potential of human cancer missense mutations and classify the effect as either high or low and assign it to exclusively well-studied missense mutations taken from major cancer databases thereby characterizing how the corresponding single residue substitutions impact protein function. The research presented in this Chapter has an aim to systematically implement a structural geometrical approach for analyzing and assessing the effects of missense mutations on protein structure and function in human cancer genes. This approach is based on a potential score of a protein that is calculated based on the amino acid composition of the four nearest neighbor residues (quadruplets) defined by the Delaunay tessellation.

Mutational Data Sets

Cancer associated missense mutations were collected from six different comprehensive cancer-sequencing projects, namely TCGA, ICGC, COSMIC, BIC, IARC TP53 and MSKCC projects (Sarcoma and Prostate cancers). An integrated database, IDHCMM Integrated Database of Human Cancer Missense Mutations,

was generated with data from all these sources. Only somatic missense mutations or non-synonymous coding mutations were downloaded from these source databases. The mutations were then mapped to their PDB structures (x-ray crystal structures) based on their associated unique Ensembl, Entrez, RefSeq and SwissProt IDs. The construction, content and working functionalities of the database have been explained in Chapter one.

IDHCMM contains more than 1.48 million records, which comprise 215374 distinct Amino Acid Mutations (Missense/Non-Synonymous), 59182 distinct Genes and 684892 distinct 'Gene-Missense Mutation' pairs.

Each gene in IDHCMM could retrieve a huge list of all the missense or non-synonymous mutations associated with the gene, pooled from all the source databases. The associated mutational data includes CDS Mutation, CDS Mutation Type, CDS Mutation Start, CDS Mutation Stop, AA Mutation, AA Mutation Type, AA Mutation position, Mutation Validation Status, Mutation Detection Platform, Mutation Validation Platform. Each of the genes is also associated with their protein 3D structure from PDB. The 3D structure information includes the PDB ID, structure resolution, the chain information, the amino acid residue positions in the structure and the domain affected. Each of these PDB entries is also linked to its missing residues information. Search can be performed based on Gene names, actual amino acid mutation, PDB ids and also UniProt ids.

There are 3 genes, TP53, PTEN and PIK3CA, each with more than 200 missense mutations listed. Nine genes, TP53, PTEN, PIK3CA, VHL, CDKN2A,

EGFR, ANK2, F8 and SI, in IDHCMM database have more than 100 known cancer missense mutations (including the above three) and 38 genes with mutations count between 50 and 100. All these genes have PDB structures and other related structural information associated. Nine genes with more than 100 mutations each were considered for tessellation. However, TP53 and VHL were not considered in this study as they have already been studied elsewhere using the same approach as ours. Mathe *et al.* applied the presented computational geometry approach to predict the functional impact (transactivation activity) of missense mutations in the DBD of the tumor suppressor *TP53* [190]. The method was found to predict transactivation with an accuracy varying between 64.2 and 78.5%, depending on the promoter. Another study observed that the structure based models generated using our approach to predict functional impact of mutations in VHL tumor suppressor protein gave high AUC values and accuracy showing that these models can be used further to make better predictions of functional impacts on proteins.

List of missense mutations, 3D protein structures, and associated information was collected from IDHCMM for the remaining 7 proteins PTEN, PIK3CA, CDKN2A, EGFR, ANK2, F8 and SI. F8 was excluded from the study for having a resolution < 3. ANK2 was also eliminated from the study as very few of the positions of the missense mutations overlapped with the protein 3D structure positions. This left us with 5 proteins to proceed - PTEN, PIK3CA, CDKN2A, EGFR and SI.

Each of these candidate proteins had multiple protein structures from PDB with different resolutions, different lengths and a number of missing residues. For example, EGFR had more than 55 3D protein structures associated, each with more than 100 missense mutations aligning within the structure positions. There are conditions that preclude a reliable tessellation of approximately one-third of solved X-ray protein structures in the PDB, including non-consecutive coordinate file residue numbering (e.g., unresolved portions of a structure) resulting in some missing residue C-alpha coordinates, as well as multiple C-alpha coordinates (i.e., parallel occupancies) for one or more residues [191]. So before PDB files can be tessellated, they must be checked for consecutive residue numbering and the missing residues in the sequence. All the structures for each protein were scanned for missing residues and protein structures with a higher resolution, longer fragments with high number of overlapping missense mutations and less number of missing residues, were selected. These are listed in Table 3.

After all the filtering we obtained 6 different PDB protein structures pertaining to 5 different proteins. The PDB files for each of the proteins were then tailored for obtaining the longest fragments in the protein structure, essentially containing coordinates of only the consecutive atoms identified in table 3. A separate PDB file was generated for each fragment of each protein. In order to determine the statistical potential of the selected reference set of proteins, they were tessellated as described by Singh *et al.* [178] and Vaisman *et al.* [188].

Table 4: List of Proteins Selected for Tessellation

Gene name	PDB Id	Chain	Resolution	Total no. of mutations	Structure positions	Missing residues	Longest fragment	Fragment length	No. of mutations in fragment
EGFR	3POZ	A	1.5	125	696-1022	696-700, 734-737, 748-754, 868-874, 1004-1009, 1018-1022	755-867	113	79
	3W32	A	1.8	141	696-1022	693-700, 1018-1022	701-1017	317	141
PTEN	1D5R	A	2.1	280	8-353	7-13, 282-285 309-312, 352	14-281	268	261
SI	3LPP	A	2.15	110	29-898	1-27 *	29-898	870	109
PIK3CA	3HHM	A	2.8	206	1-1068	-28**, -10-0**, 1-4, 310-320, 414-420, 517-523, 941-952, 1063-1068	5-309	305	50
							524-940	417	50
							953-1062	110	68
CDKN2A	1BI7	B	3.4	172	1-156	1-9, 135-156	10-134	125	172

* expression tag as per the 3LPP PDB file

**Residues prior to the beginning of the reference sequence have been numbered negative due to the presence of an expression tag (as clarified by RCSB helpdesk)

Mutant Proteins

PTEN

PTEN, also known as MMAC-1 or TEP-1, is one of the most frequently mutated tumor suppressors in human cancer. It is also essential for embryonic development, cell migration and apoptosis. PTEN functions primarily as a lipid phosphatase to regulate crucial signal transduction pathways; a key target is phosphatidylinositol 3,4,5-trisphosphate. Functions for PTEN have been identified in the regulation of many normal cell processes, including growth, adhesion, migration, invasion and apoptosis. PTEN plays particularly important roles in regulating anoikis (apoptosis of cells after loss of contact with extracellular matrix) and cell migration. Mutations in both alleles of the PTEN gene arise during cancer progression in a remarkable variety of cancers, including brain, prostate, breast and endometrial cancers, plus melanoma; frequencies of mutations in both alleles reach 50% for certain cancers in some studies. The PTEN gene is also mutated in inherited cancer syndromes such as Cowden syndrome. PTEN gene can be found mutated both very early in tumorigenesis (as in hereditary cancer syndromes) and also much later in advanced cancers. Defects in PTEN, whether they are inherited mutations, or arise through later somatic mutation or epigenetic reduction, can cooperate at multiple stages with the loss of other tumor suppressors and/or activation of oncogenes to promote malignancy. In fact, cooperation between mutations in PTEN and in p27KIP1 or

Wnt-1 has recently been shown to promote oncogenesis (Di Cristofano et al., 2001; Li et al., 2001).

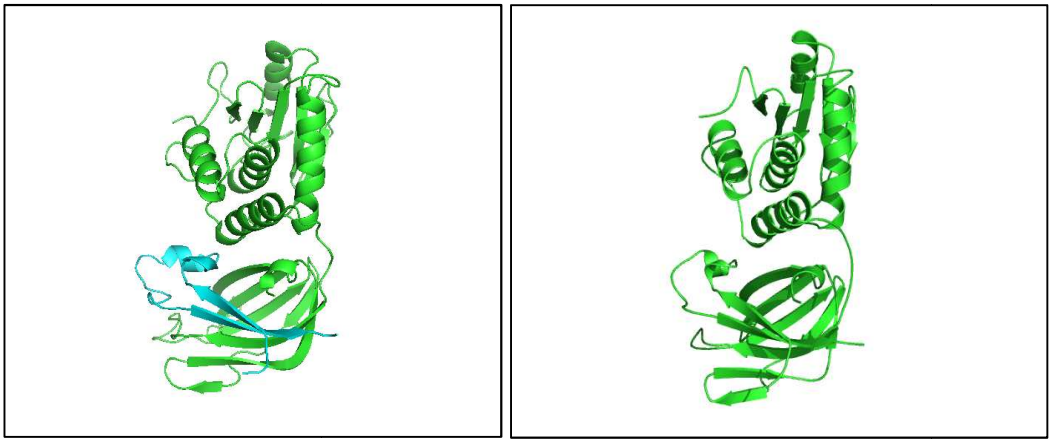


Figure 22 (a) (b)
(a) shows the PDB structure of 1D5R, the fragment in blue is the one which has missing residues in between. Figure 22 (b) shows the structure of 1D5R longest fragment selected in this study, which is a continuous chain and does not have any missing residues.

EGFR

EGFR, Epidermal Growth Factor Receptor, an oncogene, also called ErbB-1 or HER1 is a cell-surface receptor that binds epidermal growth factors and so is a major regulator of several distinct and diverse signaling pathways. These signalling pathways can have various effects, including cell growth, proliferation and migration. It is frequently overexpressed in many malignancies including non-small cell lung cancer (NSCLC), and overexpression may be associated with a negative prognosis. Mutations that lead to EGFR upregulation or overactivity have been associated with a number of cancers, including lung cancer, anal cancers and glioblastoma multiforme. Both EGFR overexpression and activating

mutations in the tyrosine kinase domain of the *EGFR* gene lead to tumor growth and progression. A large body of experimental and clinical work supports the view that EGFR is a relevant target for cancer therapy. Consequently, EGFR has become a target for anti-cancer drug therapy. Mutations, amplifications or misregulations of EGFR or family members are implicated in about 30% of all epithelial cancers.

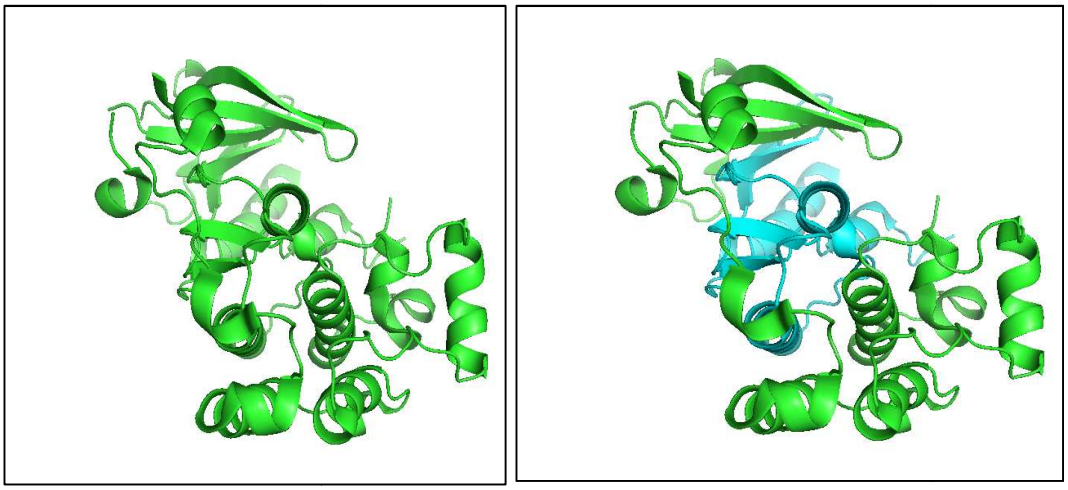
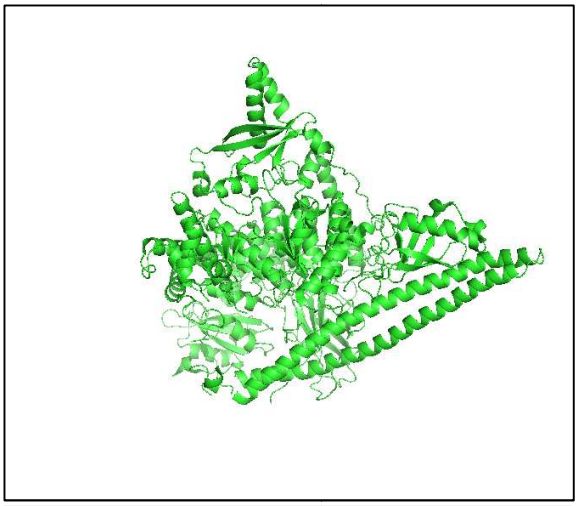


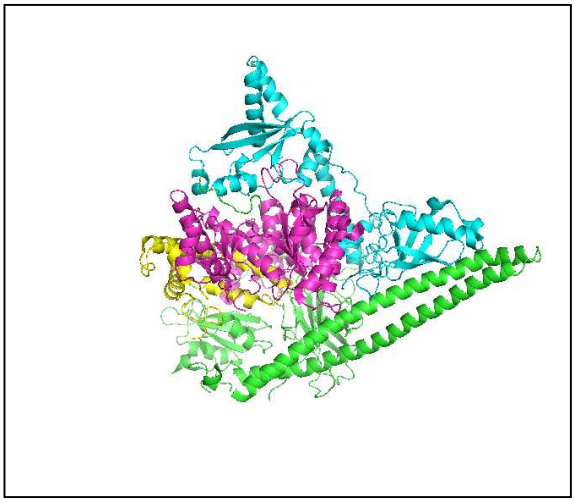
Figure 23 (a) (b)
(a) shows the PDB structure of 3POZ, Figure 23 (b) shows the structure of 3POZ longest continuous fragment, in blue, selected in this study, which does not have any missing residues.

PIK3CA

The phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha, also called p110α protein, is a class I PI 3-kinase catalytic subunit. The human p110α protein is encoded by the *PIK3CA* gene. The involvement of the PIK3CA gene product p110a, in human cancer has been suggested for over 15 years, and support for this proposal had been provided by both genetic and functional

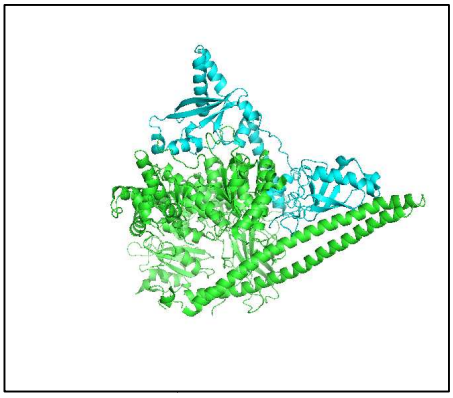
studies, including most recently the discovery of common activating missense mutations of PIK3CA in a wide variety of common human tumor types. Somatic mutations at the PIK3CA gene have been found in tumors and thus, it can be considered a bona fide oncogene (Samuels et al., 2004). Most of the mutations cluster in hotspots within the helical or the catalytic domains. Mutations in the *PIK3CA* gene have been identified in carcinomas arising from colon, breast, ovary, liver, stomach, and lung as well as in glioblastomas. Evidence suggests that such mutations lead to constitutive activation of the PI3K pathway. Mutations in *PIK3CA* are clustered and occur mainly in the helical (exon 9) and kinase (exon 20) domains of the protein. *PIK3CA* mutations frequently occur in diverse cancers and are associated with constitutive activation of the PI3K/AKT/mTOR pathway. In addition, *PIK3CA* mutations predicted sensitivity to PI3K/AKT/mTOR inhibitors in multiple tumor types in preclinical and early clinical experiments. The PDB structures chosen based on the selection criteria set mentioned in the prior section, for the protein PIK3CA is 3HHM chain A.



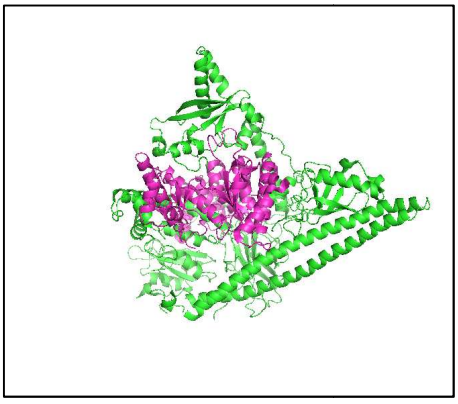


(b)

(a)



(c)



(d)

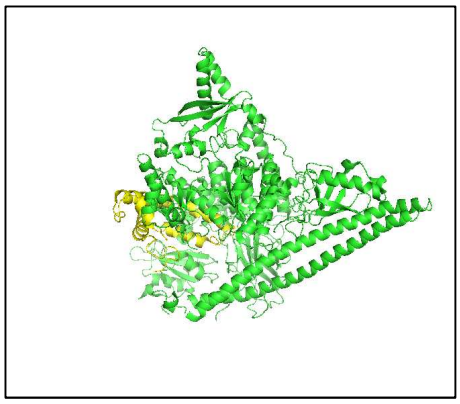


Figure 24 (e)
(a) shows the PDB structure of 3HHM, Figure 24 (b) shows the structure of 3HHM with the three longest continuous fragment, in blue, pink and yellow, selected in this study, which does not have any missing residues. Figure 24 (c) shows the first fragment, in blue, which spans the protein from residue 5-309; Figure 24 (d) shows the second fragment, in pink, which spans the protein from residue 524-940; Figure 24 (e) shows the third fragment, in yellow, which spans the protein from residue 953-1062;

CDKN2A

CDKN2A, cyclin-dependent kinase inhibitor 2A or multiple tumor suppressor 1, gene encodes a tumor suppressor protein p16 in humans. p16 plays an important role in cell cycle regulation by decelerating cells progression from G1 phase to S phase, and therefore acts as a tumor suppressor that is implicated in the prevention of cancers, notably melanoma, oropharyngeal squamous cell carcinoma, and esophageal cancer. The CDKN2A gene is frequently mutated or deleted in a wide variety of tumors. Mutations in the CDKN2A gene are associated with increased risk of a wide range of cancers and alterations of the gene are frequently seen in cancer cell lines. Examples include Pancreatic adenocarcinoma, esophageal cancer and gastric cancer cell lines. Furthermore, p16 is now being explored as a prognostic biomarker for a number of cancers. For patients with oropharyngeal squamous cell carcinoma, using

immunohistochemistry to detect the presence of the p16 biomarker has been shown to be the strongest indicator of disease course. Presence of the biomarker is associated with a more favorable prognosis as measured by cancer-specific survival (CSS), recurrence-free survival (RFS), locoregional control (LRC), as well as other measurements. The appearance of hyper methylation of p16 is also being evaluated as a potential prognostic biomarker for prostate cancer. Somatic mutations of *CDKN2A* are present in up to 95% of pancreatic tumors. The CDKN2A locus is a valuable model for assessing relationships among variation, structure, function, and disease because variants of this gene are associated with hereditary cancer, somatic alterations play a role in carcinogenesis, allelic variants occur whose functional consequences are unknown and the crystal structure is known. The PDB structures chosen based on the selection criteria set mentioned in the prior section, for the protein CDKN2A is 1BI7 chain B.

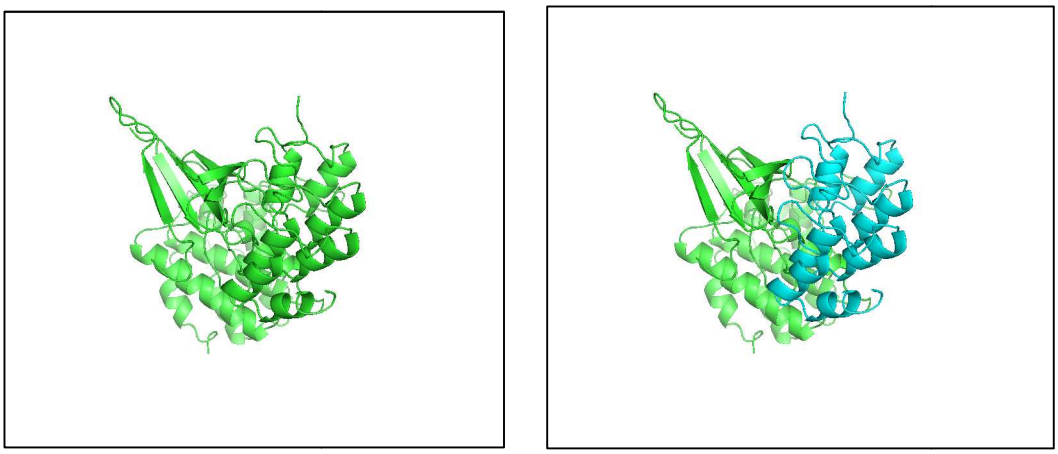


Figure 25 (a) (b)
(a) shows the PDB structure of 1BI7, Figure 25 (b) shows the structure of 1BI7 longest continuous fragment, in blue, selected in this study, which does not have any missing residues.

SI

SI, Sucrase-isomaltase, is a type II transmembrane glycoprotein with preferential expression in the apical membranes of the polarized enterocytes of the intestinal brush border membrane, where it is essential for the processing of dietary carbohydrates. SI mutations result in loss of enzyme function by preventing the biosynthesis of catalytically competent SI at the cell surface. These mutations disrupt the folding and processing of the sucrose-isomaltase enzyme, transportation of the enzyme within the intestinal epithelial cells, the orientation of the enzyme to the cell surface, or its normal functioning. An impairment in any of these cell processes results in a sucrase-isomaltase enzyme that cannot effectively break down sucrose, maltose, or other compounds made from these sugar molecules (carbohydrates). Sucrase-isomaltase is a tissue-based phenotypic marker that is an independent prognostic factor in colorectal cancer. A study suggests that SI expression correlates with the progression of dysplastic adeno

carcinoma therefore sucrase-isomaltase expression may be useful as a clinical marker to improve our prognostic capabilities in patients with dysplastic lesions of the colon, that is, inflammatory bowel disease. Interestingly SI, though not implemented in many cancers, has listed a huge number of missense mutations in different cancer databases. The PDB structures chosen based on the selection criteria set mentioned in the prior section, for the protein SI is 3LPP chain B.

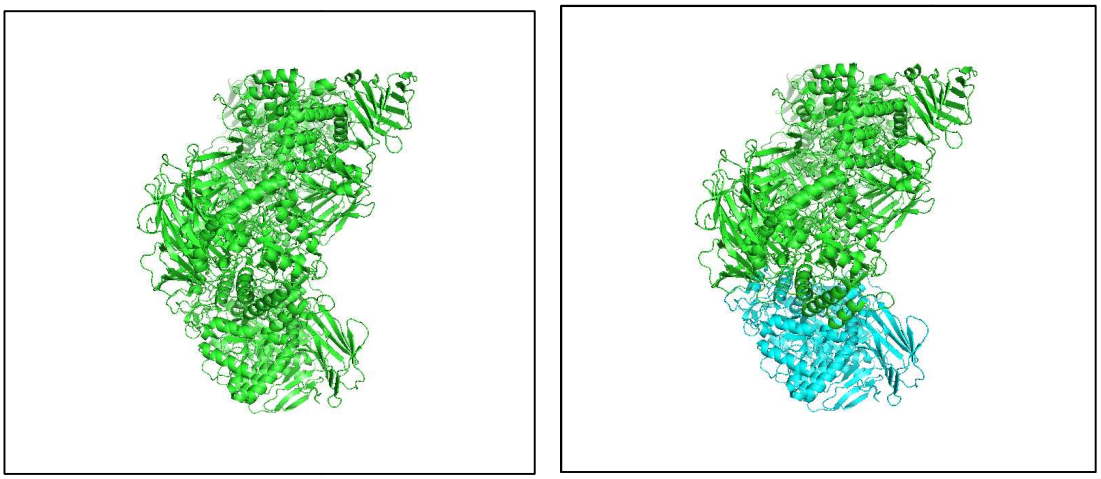


Figure 26 (a) (b)
(a) shows the PDB structure of 3LPP, Figure 26 (b) shows the structure of 3LPP longest continuous fragment, in blue, selected in this study, which does not have any missing residues.

The mutant dataset, as described in the latter section of this chapter, is used to train a model, based on the machine learning supervised classification algorithms, useful for predicting the disease potential of human cancer missense mutation mapped to solved protein structures. Our models, trained using a relatively large number of missense mutations per protein, are shown to perform at least as well as other methods.

Choosing an appropriate control data set for any analysis of disease mutations is an important step[192]. The control dataset was generated by random shuffling of the two classes, with in each of the mutant dataset.

Delaunay tessellation

Delaunay tessellation of a given protein structure yields an aggregate of non-overlapping, space-filling, irregular tetrahedral, referred to as Delaunay simplices, whose vertices are the amino acid point representations. Each simplex in a protein structure tessellation objectively defines a quadruplet of nearest-neighbor residues in the protein based on the identity of the four amino acids represented by the vertices of the simplex. The Quickhull algorithm [171] is used to perform the Delaunay tessellations, and an in-house suite of Java programs is used to perform pre- and post-processing as well as the subsequent calculations and analyses. Figure27 shows the Delaunay tessellation diagram of one of the fragments of a protein from our dataset, 3HHM (953-1062), a 110 residue fragment of the protein PIK3CA, chain A.

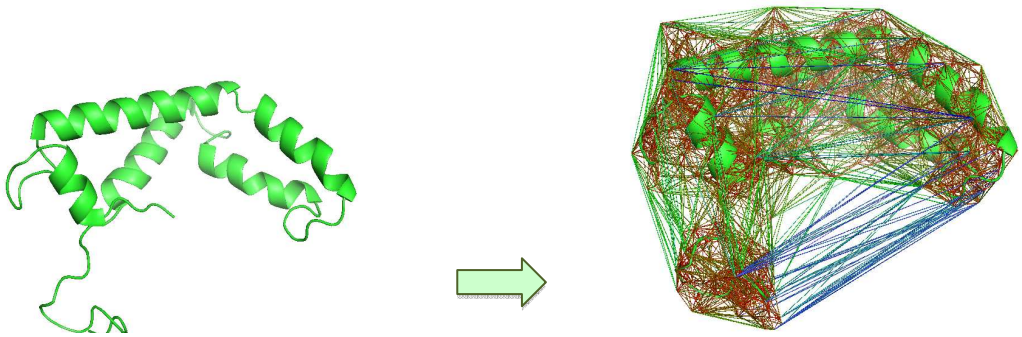
Potential Score

Taking the PDB coordinates of the wild-type protein structures in the training set, the total potential or topological score of the protein is calculated as the sum of the log-likelihood scores of all the simplices that form the Delaunay tessellation of the structure.

The log likelihood of each of the quadruplets is calculated as

$$\frac{\qquad}{\qquad}$$

where f_{ijkl} represents the frequency of quadruplets containing residues i,j,k,l in a representative training set of high-resolution protein structures with low primary sequence identity.



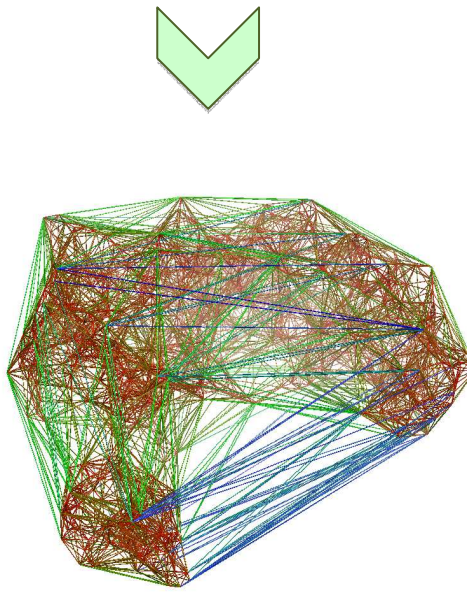


Figure 27: Representation of Delaunay Tessellation of a Protein Structure (3HHM_A)
Each point represents the center of mass of an amino acid side chain. The tessellation of a protein gives the list of all four nearest neighbor residues that constitute this protein.

And p_{ijkl} is the expected frequency of the quadruplet, calculated with the following equation:

where a_i, a_j, a_k, and a_l represent the frequencies of amino acids i, j, k, and 1 in the training set, and C is a permutation factor defined as

---

where n is the number of distinct residue types in a quadruplet, and t_i is the number of amino acids of type i. A *total potential* (tp) or topological score for the protein is calculated by globally adding up the log-likelihood scores of all tetrahedral simplices in the tessellation.

An individual *residue environment score* is also calculated for each position by first locally identifying the simplices that specifically share the corresponding C-alpha coordinate as a vertex and then adding up the log-likelihood scores of only the amino acid quadruplets represented by these simplices. Collectively, the vector of residue environment scores for all of the amino acid positions in a protein is referred to as the potential profile $Q = <q_1, q_2, q_3, \ldots. q_N>$ for the protein, where q_i = *residue environment score* for the amino acid at position i and N = primary sequence length of the solved protein structure.

Wild-type Protein and the Mutant Protein

Since protein structure dictates function, it follows that the relative structural differences between variant proteins and their wild type counterparts also correlate with the corresponding relative functional changes. A potential profile can be easily calculated for both wild type and mutant proteins, assuming that the structural differences between them are small and that their tessellation results are similar. In this case the difference between the profiles is defined only by the change in composition of the simplices involving the mutational sites. The resulting difference in profiles provides important insights into the changes in protein energetics due to the mutation.

The topological profile or total potential of the same protein with mutations is obtained by utilizing the identical tessellation while substituting

only the amino acid residue at the point of mutation. This causes a change in the log-likelihood scores of all simplices that use the point as a vertex, since member/members of their respective quadruplets is mutated. Finally, the *residual score* (RS) defined as the difference between the total potential of the mutant and the total potential of the wild-type protein, which reflects the relative change of the mutant protein sequence-structure compatibility from wild-type, is calculated. The *Residual Profiles* for mutants are vectors of N (number of amino acids in a given protein) elements, each representing the residual score for each residue [193], [194].

The potential profiles of wild-type and mutants were then analyzed in order to study their ability to characterize the effect of mutations in a protein. The residual score of a mutant is a measure of the relative change in sequence-structure compatibility from the wild-type protein, therefore it is generally expected that the more negative the residual score, the less active the mutant. The annotated mutant protein systems were examined for their residual profiles, and a comprehensive statistical analysis using machine learning algorithms was performed to predict the functional effect of the mutants.

Comparing with other Predictive methods

In the "Introduction" section of this dissertation, we had elaborately looked at different predictive methods published, for predicting functional impact on proteins. Different functions of proteins could be affected by different changes in different parts of the gene coding for a particular protein. Based on these

changes, many different predictors have been designed which focus on a particular functionality of the protein, like activity, stability, binding etc. The comparative study showed that most of the approaches are either specific to only some of the protein aspects or are data dependent and do not produce good accuracies. SIFT, PolyPhen and MutationaAssessor have stood out to be widely accepted predictors covering a wide set of protein features and producing the consistently good prediction accuracies. MutationAssessor has provided best prediction accuracies as of now, as observed in many comparative studies [40], [146] . Gnad and et al. also compared single as well as combinations of predictors and found that combinations of SIFT, PolyPhen-2 and Mutation Assessor gave better results compared to other combinations. No combination improved on Mutation Assessor alone [40]. It has been trained and tested on cancer data and the approach is based on sequence evolutionary conservation information. The novelty of the approach, as stated by the authors, is in exploiting the evolutionary conservation in protein subfamilies, up to the level of protein subfamilies. However, MutationAssessor does not take into account any structural features in its feature set. This, along with the fact that it has produced the best predictions among the state-of-the-art predictors available, makes it a good candidate to compare models from our approach.

Unfortunately there is no enough experimentally validated cancer phenotype data available on impact of variants, which can ascertain the effects of different missense mutations on structure and function of different kinds of proteins.

There exist only models generated by different predictive methods, which can be relied upon based on statistical evaluations of the methods and their accuracies. In the absence of the experimentally validated data, replying upon the best of the models is what seems to be the most appropriate and logical approach.

To validate our models in this study, and to see if our approach can model the functional impact of protein missense mutations from cancer phenotypes, we compared our functional profiles with the feature vectors of MutationAssessor. The functional impact variable of all mutants for each protein in our data set was assigned into two classes: High and Low. The class labels were obtained and assigned by running MutationAssessor predictions for the same datasets. MutationAssessor was accessed via their webserver at http://mutationassessor.org/ . MutationAssessor calculates a combined score for the functional impact of changing an amino acid residue with one particular set of interactors of the mutated protein, to another set of interactors, which consequently leads to an altered biological function. It quantifies the entropy differences resulting from a mutation that affects conserved residue patterns in protein subfamilies and defines a conservation score and a specificity score. These two scores are combined by simple averaging, to obtain the Functional impact score (FIS). The functional impact obtained as this score is also given labels based on its range as shown in table below:

Table 5: Fucntional Impact Classes from MutationAssessor

Impact Class Label	**FIS Score range**
Neutral□	FIS ≤ 0.8
Low	0.8 < FIS ≤1.9
Medium	1.9 < FIS ≤ 3.5
High	FIS > 3.5

In our study here, we assign two class labels for impact, High and Low assigned by running the MutationAssessor for all the protein mutants and considering only two classes High and Low. The four classes labeled by the MutationAssessor approach were drilled down to two classes by splitting the top half of the Medium Impact class into High Impact and the lower half into the Low Impact class. Neutral class was not used in order to focus only on the two classes of impact and to avoid the diluting effect of the neutrals on the dataset. We tried to obtain better accuracies of the mutants based on 2 class classification and looking at results observed from preliminary tests of lower accuracies when using 3 or four class distinction, the Neutral class was not used. The mutants were grouped into only 2 classes High and Low.

In order to examine if accuracies change with difference in of functional impact, the datasets were broken down into smaller sets with each subsequent dataset having 5 more mutants from the Higher Impact class and 5 from the Low Impact class, such that the smallest dataset had top 10 with the highest Impact score and lower 10 with the Lowest impact scores. The subsequent data set would contain

15 of each and so on. This represents the distance and separation between functional impact of the mutants. These datasets were then submitted to Machine learning algorithms.

Machine Learning

Machine learning algorithms implemented in this method were support vector machines (SMO), decision trees (J48) and Random Forest (RF). These algorithms implemented are available as part of an extensive suite of machine learning tools referred to as Weka (Waikato Environment for Knowledge Analysis; http://www.cs.waikato.ac.nz/ml/weka/index.html) [195].

Supervised classification algorithms require that the mutants of an enzyme be represented as vectors of the same dimension, with each vector component describing a particular attribute of the mutants. The attributes explored in the cited literature include information readily available from sequence data (e.g., physicochemical classes of wt and mutant residues, hydrophobicity difference, and conservation score at mutated residue position), and information directly predicted from protein structure (e.g., secondary structure, buried charge, and solvent accessibility).

Here we have used residual profiles of the mutants as feature vectors for machine learning. Each residual profile vector of a mutant typically contains sparse numbers of non-zero components, reflecting the mutated residue positions and all of its nearest neighbors (i.e., positions with which it forms Delaunay

simplices). The non-zero components of a mutant vector correspond to all positions in the protein structure that participate in nearest-neighbor topological contacts with the mutated residue position. Additionally, the values at these non-zero components are a unique reflection of the type of residue replacement occurring at the mutated position. For every set of protein mutants that reflect all mutations at a specific residue position in the dataset, the corresponding residual profile vectors share the property that the zero and non-zero components are located in the same vector. Finally, we develop models by using a two-class labeling of the mutants. The approach described above is applied to the all the proteins in the mutational data set mentioned in earlier section.

Support Vector Machines

Support vector machine (SVM), a learning algorithm that is only applicable to two-class problems, uses a kernel function to map the training instances (the residual profile vectors of the mutants) nonlinearly into a higher-dimensional feature space. An optimal separating hyperplane, one that provides a maximal margin of separation between instances from the two classes, is subsequently constructed in the feature space and corresponds to a nonlinear decision boundary in the original space. We have used the support vector machine implementation available with Weka [195] which is based on a sequential minimal optimization algorithm developed by John Platt. All of the default parameters have been used here.

Decision Trees

Decision tree [196] learning yields a classifier in the form of a rooted tree, such that a mutant is sorted down the tree by performing tests at each of the internal nodes. A decision is made at each internal node based on the value of a specific attribute (a component of the mutant residual profile vector), leading the mutant down a particular branch to the next node. Since the mutant attribute values in this dataset are all numeric, the decisions that branch from an internal node are binary in nature and take the form D25 $<a$ or D25 $> a$, where D25 is the attribute name or label and a is a real number. The recursive process terminates once the mutant reaches a leaf node, where the mutant class is provided. A divide-and-conquer approach is employed during training, whereby at each stage starting from the root, an attribute is selected that best separates the classes. In order to avoid overfitting of the training data, which generally leads to poor model performance on independent test sets, the learned trees are typically pruned. The Weka [195] decision tree implementation, J48 and Random Forest were used for our research. All of the default parameters have been used here.

Java programs were written to process the data files, format conversion and running the Weka attribute selection and Classifiers. Weka was run on command line to automate the process and doing the attribute selection and classification in batch mode rather than one file at a time on the GUI based version.

Model testing is performed using stratified tenfold cross-validation for the purpose of generating receiver operating characteristic (ROC) curves. Tenfold

cross-validation entails a randomization of the training set mutants into 10 equally sized subsets, and each subset is subsequently used as a test set after a decision tree classifier is trained with the remaining nine mutant subsets combined. In this way, a class prediction is obtained for every mutant in the original training set, and stratification ensures that each class is properly represented in the training and test sets. A comparison of the actual and predicted classes for each of the mutants based on the outcome of 10 CV provides a simple accuracy measure for the model.

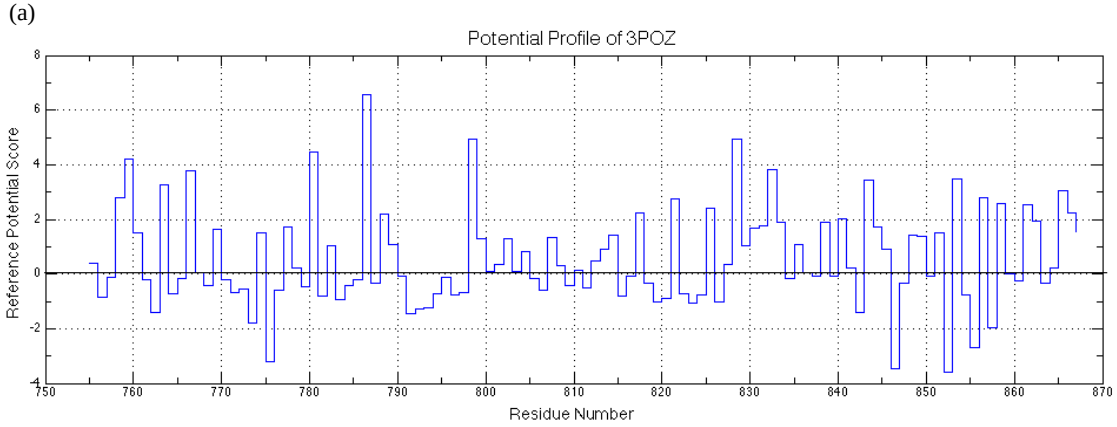
In the two-class model, considering the two class labels as high and low, accuracy is given as the sum of the true positive and true negative values divided by the sum of true positive, true negative and false positive and false negative values.

The number of correct classifications (True Positive and True Negative) and the misclassifications (False Positive and False Negative) are tabulated into a 2x2 confusion matrix and the True Positive Rate (TPR) and False Positive Rate (FPR) are calculated. The TPR and FPR form the coordinates of the ROC curves defining a single point in the unit square. Default costs, associated with the entries of the confusion matrix, are 0 for the correct predictions and 1 for the misclassifications.

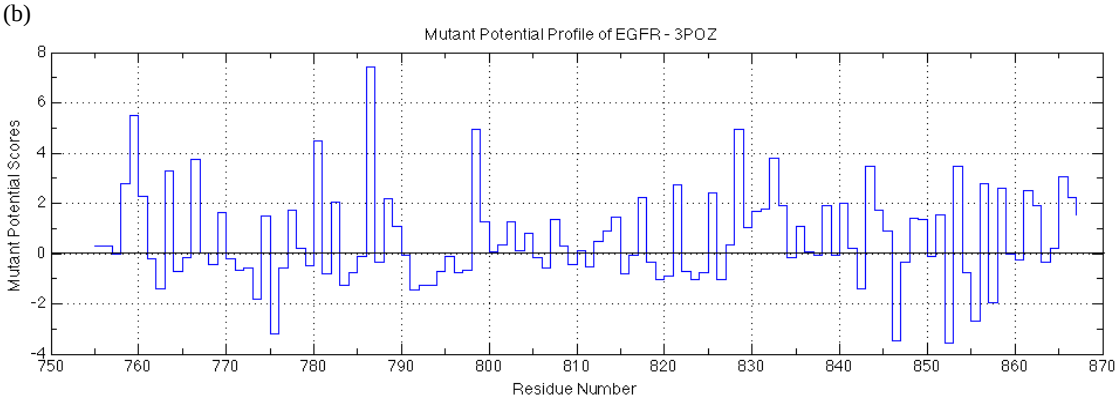
Results

A computational geometry approach using Delaunay tessellations of proteins is proposed to study the structural and functional effects of missense mutations. To validate the utility of statistical scores derived from the Delaunay tessellation methodology, we calculated such scores on six different protein structures each with more than 100 mutations, listed in Table 3. Potential scores for all the reference proteins, all the mutants of each protein and the residual scores (difference between the potential scores of reference and mutant protein) for each of the mutants were calculated using in house Java programs. Potential score differences for mutant proteins with respect to the reference PDB structure, called residual scores, were calculated and reflect the differences in log likelihoods of quadruplets. The *Residual Profiles* for mutants are vectors of elements equal to the number of amino acids in the given protein, each representing the residual score for each residue. Graphical representation of the Potential profiles, and Residual profile of the reference protein EGFR, 3D structure 3POZ, is shown in Figure 28.

(a)



(b)



(c)

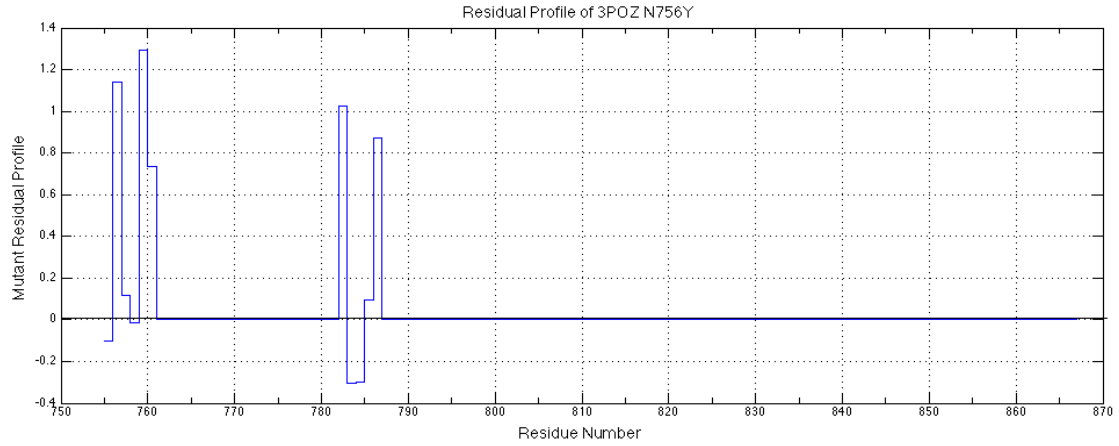


Figure 28 (a) shows the Potential profile of reference EGFR wild-type protein, comprising of potential scores of all 113 residues of the 3POZ fragment. (b) shows the potential profile of the mutant protein and (c) shows the difference between the reference and the mutant profiles, called the Residual profile. The graphs were produced using Matlab software. Profile graphs for other proteins from the dataset are present in the appendix.

Residual profiles of each mutant were then used in machine learning methods as feature vector. Along with the residual profiles, some more features were added to the feature vector (1) the actual mutation in wildtype aa-Position-Mutant aa format for example, R14M, (2) wild-type amino acid, (3) the mutation position and (4) the mutant residue were also added to the feature vector. While running the classifiers, these additional features were introduced into the feature vector

one at a time and also in different combinations, which make 16 different sets of vectors, to see if the accuracy of the models change. These combinations are elaborated in the Table 5 and discussed further in this section.

The mutant data sets, each containing 10 more mutants than previous set, with each mutant assigned to one of the two functional impact classes: High and Low and files labeled with the additional feature combination, were submitted to Support vector machine, J48 and Random Forest classifiers, in Weka.

Feature Selection: Using Weka, selective features from the feature vectors that performed best with a classification algorithm were extracted for all the data files. All combinations of possible attribute selection methods were run for all the data files. The classification results were analyzed for both combination of additional features as well as for step wise introduction of closer functional impact of mutants into datasets. The results of the run are shown for the mutants of PTEN protein structure 1D5R fragment in Table 6 and 7.

Table 6: Additional Features added to feature vectors

Additional Features	**Notation used**	**Mutation (**R14M)	**Wild-Type AA** (R)	**Position** (14)	**Mutant AA** (M)
All Features	AF	☑	☑	☑	☑
Take Out Mutation	noMut		☑	☑	☑
Take Out WT AA	noWAA	☑		☑	☑
Take Out Position	noPos	☑	☑		☑
Take out Mutant AA	noMAA	☑	☑	☑	

Take out Mutation and WT AA	noMut_WAA			☑	☑
Take out Mutation and Position	noMut_Pos		☑		☑
Take out Mutation and Mutant AA	noMut_MAA		☑	☑	
Take out WT AA and Position	noWAA_Pos	☑			☑
Take out WT AA and Mutant AA	noWAA_MAA	☑		☑	
Take out Position and Mutant AA	noPOS_MAA	☑	☑		
Take out Mutation, WT AA and Position	noMUT_WAA_POS				☑
Take out WT AA, Position and Mutant AA	noWAA_Pos_MAA	☑			
Take out Mutation, Position and Mutant AA	noMut_Pos_MAA		☑		
Take out Mutation, WT AA and Mutant AA	noMut_WAA_MAA			☑	
Take out Mutation, WT AA, Position and Mutant AA	noMUT_WAA_POS_MAA				

Table 7: Classification analysis for PTEN 1D5R mutants based on number of mutants, shows the classification accuracies achieved by three different classifiers for PTEN 1D5R fragment, for all the datasets with different number of mutants from each Impact class, representing the functional distance between the mutants.

No. of Mutants from each Impact Class	Combination of Additional Features	Classification accuracies		
		Support Vector Machine	J48	Random Forest
10 Top High Impact and 10 Bottom Low Impact	All Features	65%	70%	95%
15 Top High Impact and 15 Bottom Low Impact	All Features	66.6%	80%	90%
20 Top High Impact and 20 Bottom Low Impact	All Features	65.0%	77.5%	80%
25 Top High Impact and 25 Bottom Low Impact	All Features	68%	76%	84%
⋮				
115 Top High Impact and 115 Bottom Low Impact	All Features	68.2%	73%	79.1%
120 Top High Impact and 120 Bottom Low Impact	All Features	64.1%	70.8%	75%

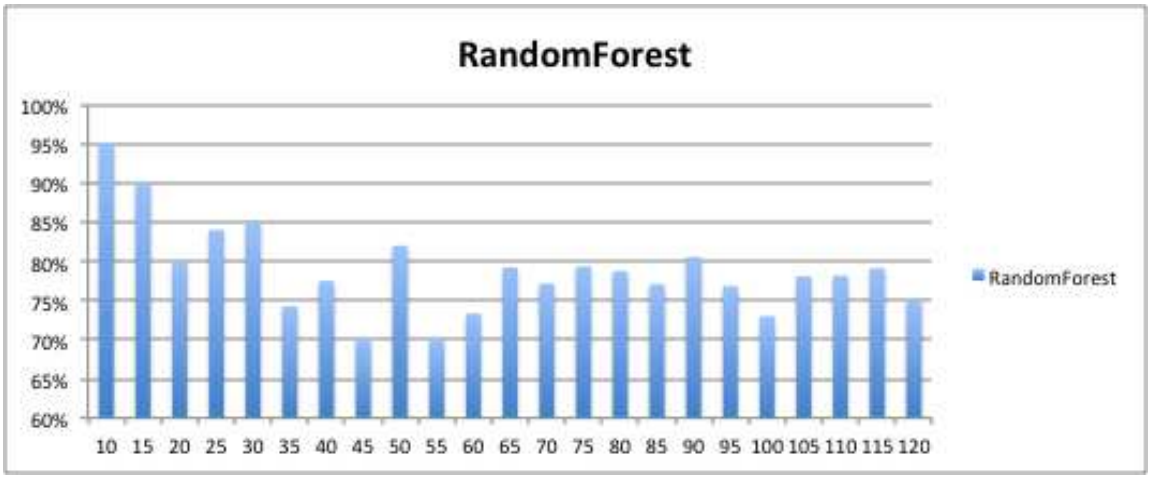


Figure 28: Random Forest accuracies across increasing number of mutants from each Impact class; shows the accuracy of the Random Forest classifier for PTEN protein 1D5R structure, across increasing number of mutants from each functional Impact class, representing the functional distance between the mutants.

Table 8: Classification analysis for PTEN 1D5R mutants based on Additional Features; shows the classification accuracies achieved by three different classifiers for PTEN 1D5R fragment, for all the datasets with a sample results for 10 mutants each from Impact class data set, and different combinations of additional features.

No. of Mutants from each Impact Class	Combination of Additional Features	Classification accuracies		
		Support Vector Machine	J48	Random Forest
10 Top High Impact and 10 Bottom Low Impact	AF	65%	70%	95%
10 Top High Impact and 10 Bottom Low Impact	noMut	60.0%	70%	95%
10 Top High Impact and 10 Bottom Low Impact	noWAA	60.0%	70%	95%
10 Top High Impact and 10 Bottom Low Impact	noPos	55.0%	80%	80%
: : : :				
10 Top High Impact and 10 Bottom Low Impact	noMut_WAA_MAA	55.0%	70%	90%
10 Top High Impact and 10 Bottom Low Impact	noMUT_WAA_POS_MAA	55.0%	80%	95%

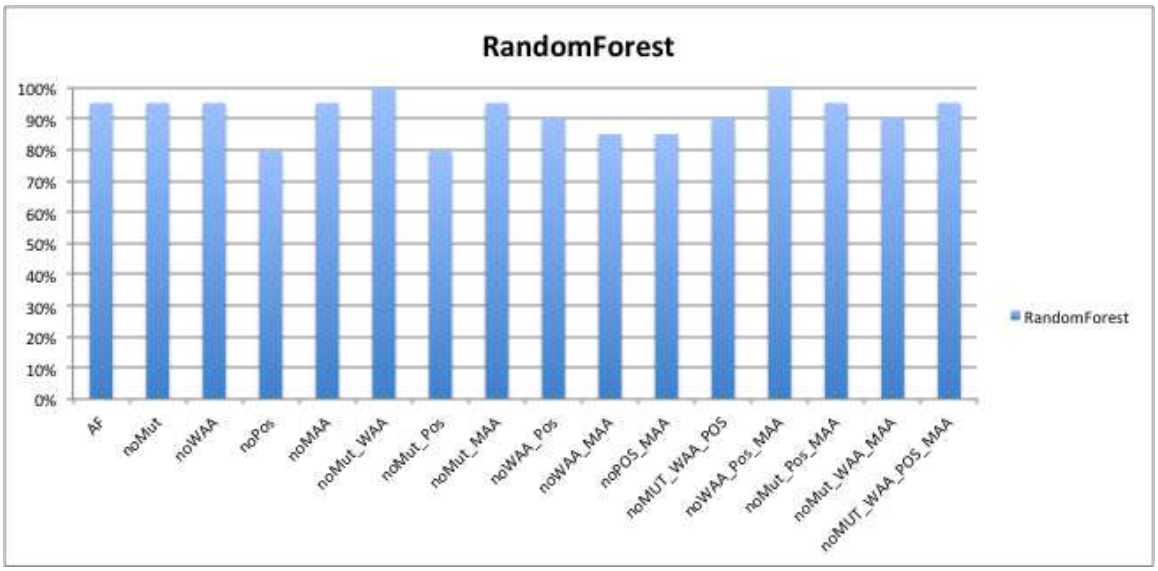


Figure 29: Random Forest accuracies across different additional feature combinations for 10 mutant each from Impact Class data set of PTEN 1D5R fragment.

From all the above results we observed that Random Forest performed better than the other classifiers. Support Vector Machine classification accuracies were more or less in par with the Random Forest as observed with some of the protein mutants but Random Forest mostly gave better accuracies. Therefore we stayed with Random Forest for all the further analysis.

Best Features Selection: The selection of the best of the features/attributes from the feature vector increased the prediction accuracies for some of the protein while it remained almost the same for the rest of the proteins. The Attribute selection functionality of the Weka software was utilized to pick the best features out of the feature vectors for each of the protein mutants. When compared across

all the protein mutants, the “CfsSubsetEval” attribute evaluator with a combination of the search methods “BestFirst” and “GeneticSearch” gave high prediction accuracies. The CfsSubsetEval evaluator evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred. The BestFirst search method searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point). The GeneticSearch search method is a Bayes Network learning algorithm which uses genetic search for finding a well scoring Bayes network structure. Genetic search works by having a population of Bayes network structures and allow them to mutate and apply cross over to get offspring. The best network structure found during the process is returned. It is seen that proteins that have shown higher classification accuracies after attribute selection, all used the “CfsSubsetEval” evalutator and the GeneticSearch method.

It was also observed that better the separation between the functional impact i.e. more different the mutational effect on the protein was, better were

the accuracies predicted by this approach. This was shown by the higher accuracies predicted for datasets with mutants mostly from the either ends of the spectrum of defined Functional Impact classes.

While there was not a clear indication from direct classification that the additional features added to the feature vector of the proteins did affect the prediction accuracies, higher accuracies of classification obtained of proteins after attribute selection showed the presence of the additional features in the selected attributes of the initial feature vectors of the mutants. The additional features mostly included were the actual Mutation and one amongst the wild-type aminoacid or the Mutant aminoacid.

Finally, with a two-class labeling of mutational effect, models for each protein are evaluated with a stratified tenfold cross-validation procedure. *Receiver operating characteristic* (ROC) curves were plotted and AUC calculated to test the robustness of the predictions, shown in figure 31. The AUC is equivalent to the non-parametric Wilcoxon-Mann-Whitney test of ranks and provides a measure of classifier performance that us insensitive to the distribution of the classes (impact classes here) [197].

As control set, the class labels were shuffled randomly and the classifiers run and ROC curves drawn. using these randomized impact The Random Forest results formed hyperbolic curves depicting the learning processes for each of the dataset/Protein.

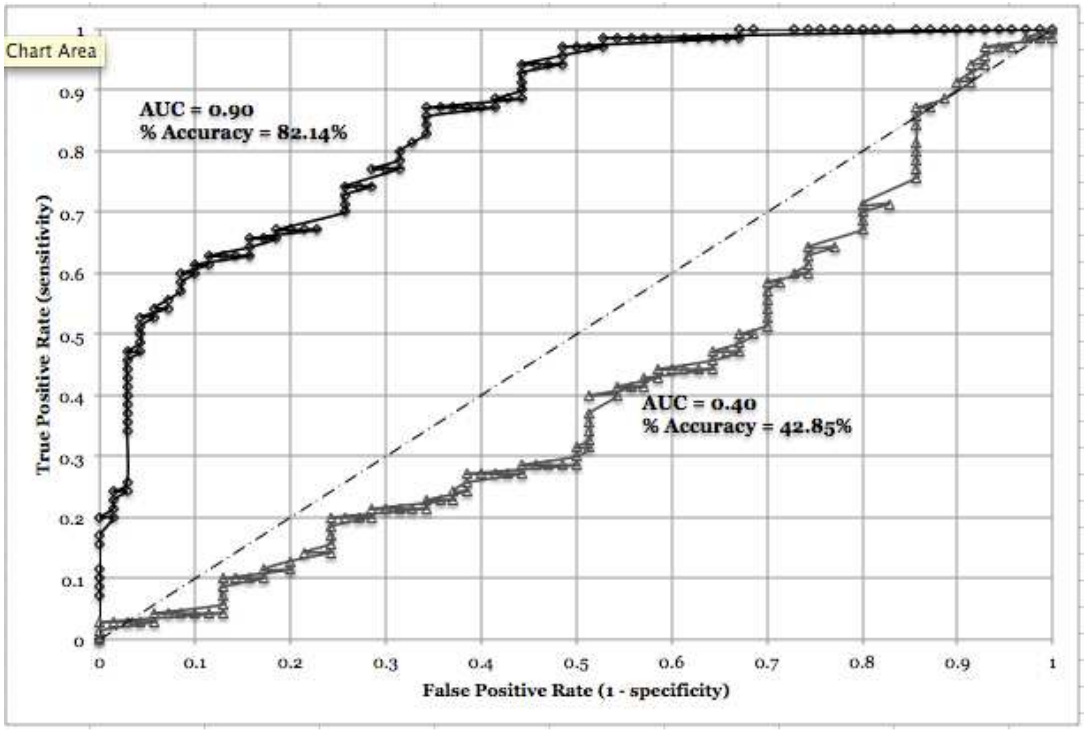


Figure 30: Two-class PTEN 1D5R ROC curves; ROC curves and the associated AUC, generating using Random Forest machine learning with the Residual profiles of PTEN 1D5R mutants. The mutants of PTEN are here labeled with high or low function impact for classification and ROC curves. Each point on the ROC curve is obtained via stratified tenfold cross-validation using a specific pair of misclassification costs. Control ROC curves were generated by random shuffling of the class labels among the training set mutants and reflect models that are near to random guessing.

The prediction accuracies, before attribute selection, ranged from 68% to 100%, and after attribute selection ranged from 76% to 95% amongst the proteins. This range of prediction accuracies definitely shows that the method used is quite robust.

Conclusions

The results of this study validate the use of the Delaunay tessellation approach for structural and functional analysis of different missense mutations originating from different proteins. Most mutant structures though are similar to

that of the wild-type such that local structural effects (with respect to the mutated residue) account for the majority of changes incurred by single point mutations. These local effects are captured by our potential profiles and the interesting trends observed in the results provide adequate proof of method using this method for various different cancer data sets.

With model performances reaching 0.9 and even more for some of the smaller datasets with mutants having better separation between functional impact classes, our residual profiles clearly catch the signal of mutational effect at both ends of the spectrum, with the help of machine learning classifiers. The uniformity of results for different proteins and a very huge number of their mutants certainly suggest that the potential and residual profiles generated by this approach can reliably used for predicting mutational impact for any kinds of proteins with available protein structure. These results support our assertion that the protein topological scores calculated based on Delaunay tessellation capture all the necessary structural information needed to study the structure to function relationships in proteins.

CHAPTER THREE: Machine Learning Models for Survival Prediction in Prostate Cancer using Gene Expression Data

Introduction

In the clinical picture of cancer, mutations can be considered as first steps towards cancer development. Mutations (including small insertions and deletions) instigate unfavorable changes in the cells such as rearrangements, copy number alterations, pathway alterations together with epigenetic and transcriptomic changes ultimately promoting cancer formation and progression [6]. Recent developments in comprehensive genomic characterization technologies have shown that this is however not a linear process and involves unanticipated complexities. It is therefore advantageous and sensible to able to complement mutational information with other genomic information such as gene expression and other genetic aberrations. The rapidly increasing size and availability of integrated genomic datasets is creating an unprecedented opportunity to study cancer biology and discover biomarkers and therapeutic targets in a novel way.

To complement our research work presented in the previous two chapters of this dissertation, we expand our focus towards building a basis for integrative

analysis. Part of our research demonstrated in this chapter extends the scope of our predictive model approach to gene expression data. Here, we performed a study to build models predicting survival in prostate cancer, based on gene expression data. Initially we had presented encouraging results in predicting mutational effects of missense mutation on protein structure and function with good accuracies. These results if complemented by similar results from gene expression analysis, would certainly provide more reliability and higher predictive power to the method. Our aim was the identification of genes whose expression levels are strongly associated with outcome. The final outcome can provide useful insights for developing targeted therapeutics and informative biomarkers. Apart from the clinical use as prognostic markers, such genes can shed light on the mechanisms causing the wide variations in survival outcomes. Quite a few approaches have been proposed and used in ranking genes according to predictive strength [198]. Genes implicated in cancer with independent prognostic value can be sought by estimating their relative importance in multivariate classification approaches {Citation}.

With this goal in mind we generated machine-learning models using gene expression profiles of genes implicated in prostate cancer to predict patient survival.

Prostate cancer and Survival modeling

Prostate cancer represents 14.4% of all new cancer cases in the US and is the second leading cause of cancer death in American men with estimated number of

233,000 new prostate cancer cases, and 29,480 estimated deaths in 2014 [199]. Massive cancer informatics efforts have been focused on discovery and validation of better diagnostic, prognostic and predictive biomarkers that aid in early diagnosis of cancer and assist in clinically relevant and reliable cancer therapies. Cancer prediction and prognosis now have equally important role, if not more, than cancer detection and diagnosis, owing to the fact that it can be done at a early stage of cancer progression. Cancer susceptibility and cancer survivability predictions are the main focus of cancer prognosis studies.

The current gold standard for diagnosis of prostate cancer includes serum prostate-specific antigen (PSA) measurement, digital rectal examination and histological inspection of prostate needle biopsies. This method has suboptimal sensitivity and specificity, and leads to many unnecessary initial and repeat biopsies. An important reason for that is, besides being a marker correlated with cancer, PSA is also a marker of increased prostate size. This points to an unrealized clinical potential for new biomarkers that can more accurately detect prostate cancer [200]. It has been seen that though patients with identical molecular, histological and clinical diagnostics are given the same treatments, survival amongst them varies a lot. This indicates that the present methods followed for deciding on cancer therapies are not clinically helpful enough. This shows the need to find other indicators to cancer prognosis and identify biomarkers that can help assess the risk at early stages, distinguish patients with aggressive disease and patients with indolent tumors and finally predict survival

rates. Research in biomarker discovery and validation has immensely helped in risk prognosis and improved treatment. However more promising novel biomarkers are needed for better detection and discovery of better therapeutic targets.

Some of the molecular biomarkers that have served as powerful prognostic or predictive indicators have been somatic mutations in certain genes (p53, BRCA1, BRCA2), the expression of tumor proteins (MUC1, HER2, PSA) or the chemical environment of the tumor (anoxic, hypoxic).[201, p. 2], [202], [203, p. 1]. Such biomarkers have been mostly identified from differentially expressed genes, transcripts, proteins or metabolites by comparing molecular profiles between benign tissue and cancer tissue and the comparison usually involves statistical hypothesis testing followed by some independent cross validation strategies to indicate significance of the results.

Advances in high-throughput molecular profiling such as microarray analysis have invigorated biomarker research and provided alternatives to the traditional methods. Enormous amounts of microarray data provide make it possible to identify gene expression profiles in cancer tissues and normal tissues. Molecular markers such as gene expression signatures have improved the predictive power of clinical nomograms eminently [204], [205]. Cancer-specific biomarker genes likely share gene expression profiles that are distinct in cancer samples as compared with normal samples. Various bioinformatics models have

demonstrated the potential of expression profiling for molecular diagnosis and survival analysis of human cancers.[206][207][208], [209]

High throughput technologies have been producing huge amounts of genomic and proteomic data leading to overwhelming number of molecular, cellular and clinical parameters. This has lead to an increasing reliance on protein markers, microarray data and non-traditional, computationally intensive machine learning approaches for conducing survival analysis [210]. Even comparative studies have observed this trend in cancer prognosis [211].

Micro array data and Gene Expression Profiles

Microarray technologies can map genome wide complex molecular divergence of cancer development and can be correlated to clinical data. Gene expression profiling measures the expression of thousands of genes at once thereby provides a complete picture of cellular function in a single experiment. These profiles can then be used to distinguish between cells where the genes show different levels of expression, for example, actively diving cells, cells reacting to a particular treatment etc. The microarray technology has been evolving from DNA microarrays to sequence based techniques like SAGE, SuperSAGE and more recently the sequence based expression analysis using RNA-Seq which is presumed to be the 'digital' alternate to microarrays.

Generally, nonhierarchical clustering, a kind of supervised classification, is used to analyze microarray data for identification of differentially expressed genes between predefined groups of samples. Microarray expression profiling is

based on the supposition that gene expression patterns can determine tumor samples from normal samples. Microarray analysis is an area of intense research and has evolved along with the technology. Identifying differentially expressed genes using a fold change cutoff has changed to using a variety of statistical tests such as ANOVA, which consider both fold change and variability to create a p-value, which is an estimate of how often we would observe the data by chance alone. However this approach looses robustness when there are a huge number of genes being analyzed. In such cases statistical methods are being developed and used such as SAM (Significance Analysis of Microarrays) [212] and a number of methods available from Bioconductor [213] etc. More recently microarray analysis techniques involve bootstrapping (statistics) and machine learning algorithms. These classifications and molecular characterization procedures are proving very valuable in providing insight into development of possible treatment stratergies for cancer.

Microarray expression profiling studies on prostate cancer have identified numerous protein coding genes with differential expression [214] [215][200], [216]. A study by Sørlie *et al.* [217] stratified the classifications described by Perou *et al.* [218] and explored the clinical value of the breast cancer subtypes. The authors separated the ER-positive tumors into two distinct groups and found that tumor classification based on gene expression was related to patient survival. In addition to identifying genes that correlate to survival, microarray analyses have also shown been utilized to establish gene expression profiles

associated with prognosis. van 't Veer L.J et al conducted studies that were able to identify "good-prognosis" and "bad-prognosis" signatures based on the expression of 70 genes that were better able to predict the likelihood of metastasis development within five years for breast cancer patients [206], [219].

Materials and Methods

Integrative Analysis

A number of studies have shown that integrating genomic data and gene expression profiles has not only provided better analysis and higher reliability on results but have also helped in complementing the other set of data. For example, Lindgren et al. combined genome profiling with global gene expression, gene mutation, and protein expression data and identified two major genomic circuits operating in urothelial carcinoma. This group of tumors showed no distinct pattern of genomic alterations, except for enrichment of CCND1 amplifications. Intriguingly, this group had the worst prognosis [220]. D'Antonio et al. integrated expression profiles, mutation effects, and systemic properties of mutated genes to identify novel cancer drivers. They were able to identify putative drivers in the majority of carcinomas without mutations in known cancer genes, thus suggesting that the method can be used as a complementary approach to find rare driver mutations that cannot be detected using frequency-

based approaches [221]. Curtis et al. [222] have shown that integrative clustering of copy number and gene expression in 2,000 breast tumors reveals novel subgroups beyond the classic expression subtypes that show distinct clinical outcomes. Mo et al. proposed a framework for joint modeling of discrete and continuous variables that arise from integrated genomic, epigenomic, and transcriptomic profiling. Using the cancer cell line encyclopedia dataset, they demonstrate that their method can accurately group cell lines by their cell-of-origin for several cancer types, and precisely pinpoint their known and potential cancer driver genes as well as demonstrate the power for revealing subgroups that are not lineage-dependent, but consist of different cancer types driven by a common genetic alteration [223]. Zhang et al. propose a network-based approach in which three kinds of data are integrated: somatic mutations, copy number variations (CNVs), and gene expressions. They applied their method, iMCMC to the Cancer Genome Atlas (TCGA) glioblastoma multiforme (GBM) and ovarian carcinoma data, and identified several mutated core modules, some of which are involved in known pathways. Most of the implicated genes they found were oncogenes or tumor suppressors previously reported to be related to carcinogenesis. Their results indicate that gene expressions or CNVs indeed provide extra useful information to the original data for the identification of core modules in cancer [224].

There are a numerous applications and software focused on the integration and analysis of oncogenomics and clinicopathological data such as

Oncomine [225, p.], PAPAyA [226], ITTACA [227], Cancer Genomics Browser [228], GenePattern [229].

Machine Learning

Machine learning methods are being used in a wide range of applications ranging from extracting patterns, detecting and classifying tumors to the classification of malignancies from proteomic and genomic assays. Different classification machine learning models applied to gene expression data have been shown to differentiate between different cancer subtypes as well as between normal and cancer samples [230].

Cruz et al. provide a good survey of machine learning applications in cancer prediction [231]. They found that almost all predictions are made using just four types of input data: genomic data (SNPs, mutations, microarrays), proteomic data (specific protein biomarkers, 2D gel data, mass spectral analyses), clinical data (histology, tumor staging, tumor size, age, weight, risk behavior, etc.) or combinations of these three. Models generated by combining both clinical and gene expression data show improved predictive accuracies of disease outcomes compared with predictions based on either data alone [232]. However, microarray data provides a huge number of features (e.g. genes) which require relatively a large training set to learn a classifier with a low error rate. Therefore feature selection becomes an important step before classification to avoid over-fitting of the model.

Feature Selection and Classification

Feature selection methods select the best features from the feature vector. These algorithms are based on the assumption that not all the features of the instances/vector are necessary for accurate classification; therefore try to identify a subset of the features, which still can accurately represent the characteristics of the instances. There are two commonly used methods for performing feature selection, filtering and wrapping. Filtering is mostly applied as a preprocessing procedure and assigns a score to each attribute and retains those, which have a score exceeding a threshold. On the other hand wrappers make the selection based on the prediction accuracy of a particular classification model [75].

In our research we use both kinds of feature selection evaluators and methods provided in the Weka software and pick the feature set that gives the highest accuracies of classification.

Classification is the process of assigning a category or class label to a sample, based on prior knowledge from a pre-defined set of classes. This prior knowledge is form the training set used in the approach, which learn a classification model and be applied to a test set predicting class to which it would belong.

In the study conducted by Cruz et al. they state from their survey that almost all machine learning algorithms used in cancer prediction and prognosis employ supervised learning [231].

We employed Decision trees (J48 and Random Forest) and Support vector machine for generating classification models, in this study.

Decision trees

The logic of decision trees is very easy to discern. Formally a decision tree is a structured graph or flow chart of decisions/nodes and their possible consequences, leaves or branches, used to create a plan to reach a goal [196]. While generating models, the leaves in the tree represent classifications and branches represent co-occurance of features that lead to those classifications. A decision tree can be learned by progressively splitting the labeled training data into subsets based on a numerical or logical test. This process is repeated on each derived subset in a recursive manner until further splitting is either not possible, or a singular classification is achieved. Decision trees have many advantages: they are simple to understand and interpret, they require little data preparation, they can handle many types of data including numeric, nominal (named) and categorical data, they generate robust classifiers, they are quick to "learn" and they can be validated using statistical tests. Random forests operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.

Support Vector Machines

Hijazi et al. [233] showed that the application of different classification methods (e.g., decision tree, k- nearest neighbor, support vector machine (SVM), bagging, and random forest) on 5 cancer datasets shows that no classification method universally outperforms all the others. However, k- nearest neighbor and linear SVM generally improve the classification performance over other classifiers.

Support Vector Machiness [234], [235] are well known in the world of machine learning and in the field of cancer prediction and prognosis. In this method two clusters are obviously evident. SVM machine learner finds the equation for a line that would separate the two clusters maximally. If there were more variables the line of separation would become a plane. If more variables were included the separation would be defined by a hyperplane. The hyperplane is determined by a subset of the points of the two classes, called support vectors. The SVM algorithm creates a hyperplane that separates the data into two classes with the maximum margin – meaning that the distance between the hyperplane and the closest examples (the margin) is maximized. SVMs can be used to perform non- linear classification using a non-linear kernel. A non-linear kernel is a mathematical function that transforms the data from a linear feature space to a non-linear feature space. Applying different kernels to different data sets can dramatically improve the performance of an SVM classifier.

The performance of the classification models is determined by the training and test errors. The training error is to the number of misclassified samples in the training set while the test error refers to the number of misclassified samples in the test set. The goal of all classification models is to achieve low training and test errors [233]. Any approach would have to deal with overfitting, which refers to the situation where there is a large set of features of the model compared to the size of the training samples, leading to a poor performance of the model. Frequently cross□validation strategy is used to avoid overfitting.

Here we extend our potentially useful machine learning methodology for accurately predicting, patient survival time, in Prostate cancer patients, based on biomarkers and their gene expression profiles. Using the Weka software, we perform feature selection and then with a two-class labeling of survival time, models are evaluated with a stratified tenfold cross-validation procedure. We generate predictive models that can anticipate short survival group or a long survival group from gene expression profiles incorporating known prostate cancer related genes implicated as biomarkers, collected from published literature.

Expression Data-set

Public, open–access gene-expression profiles of prostate cancer were downloaded from one of the primary repositories of functional genomic data, GEO (Gene Expression Omnibus) database [236]. The prostate cancer expression profiles were contributed and published by Taylor et al. [237]. This prostate oncogenome project worked on 181 primary, 37 metastatic prostate cancer samples, 12 prostate cancer cell lines and xenografts. The data statistics can be accessed at (http://www.cbioportal.org/public-portal/study.do?cancer_study_id=prad_mskcc).

Only Whole-transcript and exon-level expression data for human primary prostate cancer samples, obtained from hybridization done on Affymetrix Human Exon 1.0 ST Array platform, provided in the form of Series matrix file, was

downloaded . Series Matrix files are text files that include a tab-delimited value-matrix table generated from the 'Value' column of each Sample, headed by Sample and Series metadata. The GEO SuperSeries Id of the data set is GSE21032, SubSeries GSE 21032 and Platform id is GPL10264. The downloaded data file contained 150 primary tumor samples. Normal and control samples were excluded. Each gene expression profile had gene expression data for 43,419 genes. All genes had expression data available across all samples. Survival data was provided for 140 samples, by Taylor *et al.* in their publication. These were matched with the samples ids and the ten samples with no corresponding patient survival information were excluded.

With an aim to find out if biomarkers/genes implicated in prostate cancer can be separated or grouped based on survival class, i.e. investigate the set of genes with respect to the significance of their expression to survival we tested our classification approach, by extracting a subset of genes and their expression profiles from the expression data file. We first made a list of 53 genes and their transcript variants (total set-88) implicated in Prostate cancer, collected from published scientific literature about prostate cancer. This step backed up by the observation that using preexisting biological knowledge for survival prediction is not only reasonable, but also beneficial [238]. GenBank Accession numbers provided with expression profile data were mapped to Gene names using Ensembl biomart [239]. The subset of expression profile data were then extracted for these selected set of 53 genes and their transcript variants.

We reviewed the scientific literature to examine the reported associations of the selected genes to prostate cancer. A summary of the genes and their associations found with Prostate cancer and suggested cancer prognosis markers from the literature is given in Table 8. The list of genes and their transcript variants used are shown in Table 9.

Table 9: Summary of selected gene biomarkers for prostate cancer

Gene Name	Product	Biological Function/ Relation to PCa
AMACR	Racemase	Metabolize fatty acids in the body. Over-expressed in PCa tissue; detected with a high sensitivity and specificity in blood and urine.
ANXA3	Cell adhesion protein	A calcium and phospholipid binding protein, primarily found in urine. Implicated in cell differentiation, migration and immunomodulation. Increases the specificity and ability of PSA to discriminate between PCa stages.
ARHGDIB	Rho (or ARH) protein family	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
BRCA1/ BRCA2	Tumor suppressor	Both BRCA1 and BRCA2 are involved in maintaining genome stability as members of the ATM/ATR CHK2 DNA damage repair pathway. BRCA2 is associated with aggressive tumors and poor survival outcome. BRCA2 has prognostic ability however further experimental data is needed for BRCA1.
BSG/CD147	Membrane glycoprotein	Over-expressed in many human solid tumors. Involved in tumor invasion and angiogenesis. Increased expression of CD147 is associated with PCa progression and poor prognosis. May serve as an independent predictor of biochemical recurrence and development of PCa metastasis.
CAV1	Integral membrane protein	Mediates aspects of cholesterol and fatty acid metabolism. Circulating levels of serum Caveolin-1 correlate with extent of PCa.
CD44	Cell-surface glycoprotein	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
CHGA	Parathyroid secretory protein 1	CgA is a useful predictive marker in patients with prostatic cancer who have lower PSA. It is known that neuroendocrine cells in the prostate do not contain androgen receptors and are not regulated by androgens. PSA expression was stimulated by androgen through androgen receptors, so it is suggested that cases of prostate cancer associated with low serum PSA and high serum CgA, which would have more neuroendocrine cells with less androgen receptors, may show resistance to endocrine therapy and a poor prognosis. Therefore serum CgA tends to be elevated in high grade prostate cancer cases. Hence it can be used to fill the gap if any left by PSA when combined with serum PSA, the serum marker may effectively predict the prognosis after endocrine therapy. CgA expression in prostate cancer biopsies is an independent extrapolative factor of hormone refractory disease in patients with newly diagnosed prostate cancer on early androgen deprivation therapy.
CXCR3	Chemokine receptor 3	One of the genes from a panel of 7 genes derived from blood mRNA could distinguish between aggressive PCa and healthy patients with a high sensitivity (83%) and specificity (80%). Genes involved in regulating the immune response and gene transcription regulation in oncogenesis.
DAB2IP	Tumor suppressor	A single nucleotide polymorphism in the DAB2IP gene is associated with risk of aggressive prostate cancer (PCa), and loss of DAB2IP expression is frequently detected in metastatic PCa. The loss of DAB2IP expression initiates epithelial-to-mesenchymal transition (EMT), which is visualized by repression of E-cadherin and up-regulation of vimentin in both human normal prostate epithelial and prostate carcinoma cells as well as in clinical prostate-cancer specimens. Conversely, restoring DAB2IP in metastatic PCa cells reversed EMT.
E2F3	E2F transcription factor 3	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
EN2	Transcription factor	Involved in early embryonic development and re-expressed by PCa cells. EN-2 detection in urine as a test for diagnosing and detecting PCa. Although further validation is required, it appears it is more reliable than PSA and elevated expression is associated with increased tumor stage.
ENG	Trans membrane glycoprotein	Expressed by human vascular endothelial cells thought to play a pivotal role in endothelial cell proliferation. Elevated in prostatic fluid of men with large volume PCa.
ERG	Transforming protein	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
FCRL3	Fc receptor-like 3	One of the genes from a panel of 7 genes derived from blood mRNA could distinguish between aggressive PCa and healthy patients with a high sensitivity (83%) and specificity (80%). Genes involved

		in regulating the immune response and gene transcription regulation in oncogenesis.
FOLH1	Type II integral membrane glycoprotein	Overexpressed on prostate tumor cells and in the neovasculature of most solid prostate tumors, but not in the vasculature of normal tissues. May play an important role in the progression of PCa.
FZD7	Seven transmembrane spanning receptor	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
GOLM1	Golgi membrane protein 1	GOLM1 (Golgi membrane protein 1, Golm 1) is consistently up-regulated in clinically localized prostate cancer.Prostate epithelial cells were identified as the cellular source of GOLM1 expression.GOLM1 immunoreactivity was detected in the supernatants of prostate cell lines and in the urine of patients with prostate cancer.
GSTP1	Glutathione S-transferase pi 1	Pi-class glutathione-S-transferase (GSTP1) located on chromosome 11q13 encodes a phase II metabolic enzyme that detoxifies reactive electrophilic intermediates. GSTP1 plays an important role in protecting cells from cytotoxic and carcinogenic agents and is expressed in normal tissues at variable levels in different cell types. Altered GSTP1 activity and expression have been reported in many tumors and this is largely due to GSTP1 DNA hypermethylation at the CpG island in the promoter-5'. Hypermethylation of the GSTP1 promoter has been associated with gene silencing in prostate cancer and kidney cancer.
HDAC1	Histone deacetylase 1	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
IGF1	Insulin-like growth factor 1	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
IGFBP6	Insulin-like growth factor binding protein 6	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
IL6	Cytokine	Involved in hematopoiesis and mediates B cell differentiation. Clinical studies reveal increased serum IL-6 concentrations in patients are associated with advanced PCa tumor stage.
ING1	Tumor suppressor	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
KIAA1143	Uncharacterized protein	One of the genes from a panel of 7 genes derived from blood mRNA could distinguish between aggressive PCa and healthy patients with a high sensitivity (83%) and specificity (80%). Genes involved in regulating the immune response and gene transcription regulation in oncogenesis.
KLF12	Kruppel-like factor 12	One of the genes from a panel of 7 genes derived from blood mRNA could distinguish between aggressive PCa and healthy patients with a high sensitivity (83%) and specificity (80%). Genes involved in regulating the immune response and gene transcription regulation in oncogenesis.
KLK2	Serine protease	Serine protease that is highly expressed in prostate tissue and involved regulating semen liquefaction by activating pro-KLK3 to its active form (PSA), facilitating both tumorigenisis and disease progression to the advanced stages of PCa. Studies have shown a strong correlation with PCa- specific survival however further studies with larger cohorts are needed to confirm these observations.
MAF	Proto-oncogene	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
MKI67/Ki-67	Nuclear protein	Cell-cycle-proliferation marker. Possibly a prolific predictive marker for men with low grade, low volume PCa after radical prostatectomy. Associated with metastasis and survival outcome.
MME	Membrane metallo-endopeptidase /CD10	Inactivates several peptide hormones including glucagon, abundant in the kidney. Candidate cancer biomarker associated with PCa progression. A low level of CD10 is a possible prognostic indicator for biochemical relapse and early death as a result of lymph node metastases. Additionally may aid in personalized patient treatment/ management however this marker needs to be further validated.
MSH3	Divergent upstream protein	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
MSMB	Immunoglobulin binding factor	Secreted by epithelial cells of the prostate as well as other major organs. MSMB is a member of the immunoglobulin binding family. Exact function of MSMB is unknown but may have an autocrine (inhibin-like) role. The genetic variant rs10993994 is associated with PCa risk however further investigation is required to evaluate the predictive value of this marker.
MUC1	Mucin 1, cell surface associated	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
NME1	Tumor metastatic process-associated protein	Major role in the synthesis of nucleoside triphosphates other than ATP. Possesses nucleoside-diphosphate kinase, serine/threonine-specific protein kinase, geranyl and farnesyl pyrophosphate kinase, histidine protein kinase and 3'-5' exonuclease activities. Involved in cell proliferation, differentiation and development, signal transduction, G protein-coupled receptor endocytosis, and gene expression. Required for neural development including neural patterning and cell fate determination
PDLIM4	Reversion-induced lim protein	PDLIM4 mRNA and protein-expression levels were reduced in LNCaP, LAPC4, DU145, CWR22, and PC3 prostate cancer cells. The re-expression of PDLIM4 in prostate cancer cells has significantly reduced the cell growth and clonogenicity with G1 phase of cell-cycle arrest. We have shown the direct interaction of PDLIM4 with F-actin. Restoration of PDLIM4 expression resulted in reduction of tumor growth in xenografts. These results suggest that PDLIM4 may function as a tumor suppressor, involved in the control of cell proliferation by associating with actin in prostate cancer cells.
PIK3CA	Phosphoinositide-3- kinase; Protein kinase.	One of the most common genomic alterations in human PCa contributing to cellular transformation and cancer development. Possibly a key mechanism supporting progression toward androgen-independent PCa.
PSCA	Prostate Stem Cell Antigen, a	Involved in the regulation of cell proliferation. Up-regulated in the majority of PCas however, exact biological function is unknown. Increased expression is associated with Gleason score, seminal vesicle

	membrane glycoprotein	invasion, and capsular invasion in PCa.
PTEN	Phosphatase and Tensin homologue; protein phosphatase	Tumor suppressor involved in modulating the PI3- K/AKT signaling pathway. PTEN inactivating mutations/deletion occur in many tumors and result in rapid cell growth and division. It is associated with severe tumor stage; however, PTEN is not PCa specific It is among one of the most frequent genetic inactivation's present in PCa.
PTGS1	Prostaglandin-endoperoxide synthase 1	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.
RELB	V-rel avian reticuloendotheliosis viral oncogene homolog b	Inhibiting RelB in aggressive androgen-independent PC-3 cells by stable or conditional expression of a dominant-negative p100 mutant significantly reduced the incidence and growth rate of tumors. Consistently, down-regulation of RelB by small interfering RNA targeting also reduced tumor growth and decreased levels of IL-8. Conversely, stable expression of RelB in androgen-responsive LNCaP tumors increased the circulating IL-8 levels.
S100A11	Calcium-binding-protein family	Expressed in various solid tumors. Detection may be useful for diagnosis, monitoring and possible therapeutic targets. Involved in protein phosphorylation, enzyme activity, calcium homeostasis, and regulation of transcription factors, macrophage activators and modulators of cell proliferation. S100A2, S100A4, S100A8, S100A9 and S100A11 are associated with PCa recurrence and advanced pathological stage.
S100A2	Calcium-binding-protein family	Expressed in various solid tumors. Detection may be useful for diagnosis, monitoring and possible therapeutic targets. Involved in protein phosphorylation, enzyme activity, calcium homeostasis, and regulation of transcription factors, macrophage activators and modulators of cell proliferation. S100A2, S100A4, S100A8, S100A9 and S100A11 are associated with PCa recurrence and advanced pathological stage.
S100A4	Calcium-binding-protein family	Expressed in various solid tumors. Detection may be useful for diagnosis, monitoring and possible therapeutic targets. Involved in protein phosphorylation, enzyme activity, calcium homeostasis, and regulation of transcription factors, macrophage activators and modulators of cell proliferation. S100A2, S100A4, S100A8, S100A9 and S100A11 are associated with PCa recurrence and advanced pathological stage.
S100A8	Calcium-binding-protein family	Expressed in various solid tumors. Detection may be useful for diagnosis, monitoring and possible therapeutic targets. Involved in protein phosphorylation, enzyme activity, calcium homeostasis, and regulation of transcription factors, macrophage activators and modulators of cell proliferation. S100A2, S100A4, S100A8, S100A9 and S100A11 are associated with PCa recurrence and advanced pathological stage.
S100A9	Calcium-binding-protein family	Expressed in various solid tumors. Detection may be useful for diagnosis, monitoring and possible therapeutic targets. Involved in protein phosphorylation, enzyme activity, calcium homeostasis, and regulation of transcription factors, macrophage activators and modulators of cell proliferation. S100A2, S100A4, S100A8, S100A9 and S100A11 are associated with PCa recurrence and advanced pathological stage.
SAMSN1	Sam domain-containing protein samsn-1	One of the genes from a panel of 7 genes derived from blood mRNA could distinguish between aggressive PCa and healthy patients with a high sensitivity (83%) and specificity (80%). Genes involved in regulating the immune response and gene transcription regulation in oncogenesis.
SNRPA1	Small nuclear ribonucleoprotein polypeptide a	This protein is associated with sn-RNP U2. It helps the A' protein to bind stem loop IV of U2 snRNA.
TGFB1	Cytokine	Growth factor involved in the regulation of cellular proliferation, immune response and differentiation. Increased expression correlates with severe tumor grade, tumor invasion, PCa metastasis and biochemical recurrence. TGF-Beta needs to be validated before becoming a PCa biomarker.
TMEM204	Transmembrane protein 204	One of the genes from a panel of 7 genes derived from blood mRNA could distinguish between aggressive PCa and healthy patients with a high sensitivity (83%) and specificity (80%). Genes involved in regulating the immune response and gene transcription regulation in oncogenesis.
TRAF4	Tnf receptor-associated factor 4	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.Adapter protein and signal transducer that links members of the tumor necrosis factor receptor (TNFR) family to different signaling pathways. Plays a role in the activation of NF-kappa-B and JNK, and in the regulation of cell survival and apoptosis.
YES1	V-yes-1 yamaguchi sarcoma viral oncogene homolog 1	Associated with TMPRSS2–ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts of Prostate cancer.

Table 10: Selected List of Genes

ID	GENBANK_ACC	NM_GeneName	Gene Name
45	NM_000059	NM_000059_BRCA2	BRCA2
103	NM_000118	NM_000118_ENG_transcript variant 2	ENG
251	NM_000269	NM_000269_NME1_transcript variant 2	NME1
295	NM_000314	NM_000314_PTEN	PTEN

564	NM_000600	NM_000600_IL6	IL6
573	NM_000610	NM_000610_CD44_transcript variant 1	CD44
581	NM_000618	NM_000618_IGF1_transcript variant 4	IGF1
614	NM_000660	NM_000660_TGFB1	TGFB1
794	NM_000852	NM_000852_GSTP1	GSTP1
841	NM_000902	NM_000902_MME_transcript variant 1	MME
899	NM_000962	NM_000962_PTGS1_transcript variant 1	PTGS1
951	NM_001001389	NM_001001389_CD44_transcript variant 2	CD44
952	NM_001001390	NM_001001390_CD44_transcript variant 3	CD44
953	NM_001001391	NM_001001391_CD44_transcript variant 4	CD44
954	NM_001001392	NM_001001392_CD44_transcript variant 5	CD44
1157	NM_001002231	NM_001002231_KLK2_transcript variant 2	KLK2
2521	NM_001014986	NM_001014986_FOLH1_transcript variant 2	FOLH1
2662	NM_001018016	NM_001018016_MUC1_transcript variant 2	MUC1
2663	NM_001018017	NM_001018017_MUC1_transcript variant 3	MUC1
3060	NM_001031804	NM_001031804_MAF_transcript variant 2	MAF
3941	NM_001044390	NM_001044390_MUC1_transcript variant 5	MUC1
3943	NM_001044392	NM_001044392_MUC1_transcript variant 7	MUC1
5440	NM_001111283	NM_001111283_IGF1_transcript variant 1	IGF1
5441	NM_001111284	NM_001111284_IGF1_transcript variant 2	IGF1
5442	NM_001111285	NM_001111285_IGF1_transcript variant 3	IGF1
5593	NM_001114753	NM_001114753_ENG_transcript variant 1	ENG
5806	NM_001175	NM_001175_ARHGDIB	ARHGDIB
5898	NM_001275	NM_001275_CHGA	CHGA
6033	NM_001427	NM_001427_EN2	EN2
6104	NM_001504	NM_001504_CXCR3_transcript variant 1	CXCR3
6295	NM_001728	NM_001728_BSG_transcript variant 1	BSG
6320	NM_001753	NM_001753_CAV1_transcript variant 1	CAV1
6507	NM_001949	NM_001949_E2F3_transcript variant 1	E2F3
6728	NM_002178	NM_002178_IGFBP6	IGFBP6
6946	NM_002417	NM_002417_MKI67_transcript variant 1	MKI67
6968	NM_002439	NM_002439_MSH3	MSH3
6972	NM_002443	NM_002443_MSMB_transcript variant PSP94	MSMB
6984	NM_002456	NM_002456_MUC1_transcript variant 1	MUC1
7472	NM_002961	NM_002961_S100A4_transcript variant 1	S100A4
7475	NM_002964	NM_002964_S100A8	S100A8
7476	NM_002965	NM_002965_S100A9	S100A9
7594	NM_003090	NM_003090_SNRPA1	SNRPA1
7971	NM_003507	NM_003507_FZD7	FZD7
8142	NM_003687	NM_003687_PDLIM4_transcript variant 1	PDLIM4
8713	NM_004295	NM_004295_TRAF4	TRAF4
8858	NM_004449	NM_004449_ERG_transcript variant 2	ERG
8884	NM_004476	NM_004476_FOLH1_transcript variant 1	FOLH1
9345	NM_004964	NM_004964_HDAC1	HDAC1
9511	NM_005139	NM_005139_ANXA3	ANXA3
9716	NM_005360	NM_005360_MAF_transcript variant 1	MAF
9784	NM_005433	NM_005433_YES1	YES1
9884	NM_005537	NM_005537_ING1_transcript variant 4	ING1
9898	NM_005551	NM_005551_KLK2_transcript variant 1	KLK2
9963	NM_005620	NM_005620_S100A11	S100A11
10014	NM_005672	NM_005672_PSCA	PSCA
10285	NM_005978	NM_005978_S100A2	S100A2
10505	NM_006218	NM_006218_PIK3CA	PIK3CA
10780	NM_006509	NM_006509_RELB	RELB
11441	NM_007249	NM_007249_KLF12	KLF12
11477	NM_007287	NM_007287_MME_transcript variant 1bis	MME
11478	NM_007288	NM_007288_MME_transcript variant 2a	MME
11479	NM_007289	NM_007289_MME_transcript variant 2b	MME
11482	NM_007294	NM_007294_BRCA1_transcript variant 1	BRCA1
11485	NM_007297	NM_007297_BRCA1_transcript variant 3	BRCA1
11486	NM_007298	NM_007298_BRCA1_transcript variant 4	BRCA1
11487	NM_007299	NM_007299_BRCA1_transcript variant 5	BRCA1
11488	NM_007300	NM_007300_BRCA1_transcript variant 2	BRCA1
12373	NM_014324	NM_014324_AMACR_transcript variant 1	AMACR
14077	NM_016548	NM_016548_GOLM1_transcript variant 1	GOLM1
15424	NM_019554	NM_019554_S100A4_transcript variant 2	S100A4
15843	NM_020696	NM_020696_KIAA1143	KIAA1143

16575	NM_022136	NM_022136_SAMSN1_transcript variant 1	SAMSN1
17220	NM_024600	NM_024600_TMEM204_transcript variant 1	TMEM204
18545	NM_032552	NM_032552_DAB2IP_transcript variant 1	DAB2IP
19306	NM_052939	NM_052939_FCRL3	FCRL3
19570	NM_080591	NM_080591_PTGS1_transcript variant 2	PTGS1
20223	NM_138634	NM_138634_MSMB	MSMB
20243	NM_138709	NM_138709_DAB2IP_transcript variant 2	DAB2IP
22898	NM_177937	NM_177937_GOLM1_transcript variant 2	GOLM1
23709	NM_182918	NM_182918_ERG_transcript variant 1	ERG
23996	NM_198175	NM_198175_NME1_transcript variant 1	NME1
24021	NM_198217	NM_198217_ING1_transcript variant 3	ING1
24022	NM_198218	NM_198218_ING1_transcript variant 2	ING1
24023	NM_198219	NM_198219_ING1_transcript variant 1	ING1
24229	NM_198589	NM_198589_BSG_transcript variant 2	BSG
24230	NM_198591	NM_198591_BSG_transcript variant 4	BSG
24746	NM_203382	NM_203382_AMACR_transcript variant 3	AMACR

The data set generated from gene expression profiles of the 88 genes and their transcript variants, for all the 140 samples were fed into the classifiers as feature vectors. Java programs were written to process the data files, format conversions and running the Weka attribute selection and Classifiers.

Survival models

We carried out a systematic analysis of association between gene expression profiles of 88 genes and their transcript variants and patient survival using data from published gene expression data for prostate cancer and different sets of genes already implicated in prostate cancer and being validated as biomarkers. We performed a two-class classification, assigning 'Short' and 'Long' labels for recurrence-free (RF) survival time, to each of the samples. Three different machine-learning algorithms, support vector machines (weka implementation is called SMO), decision trees (J48) and Random Forest (RF) were run. These algorithms implemented are available as part of an extensive suite of machine

learning tools referred to as Weka (Waikato Environment for Knowledge Analysis; (http://www.cs.waikato.ac.nz/ml/weka/index.html) [195].

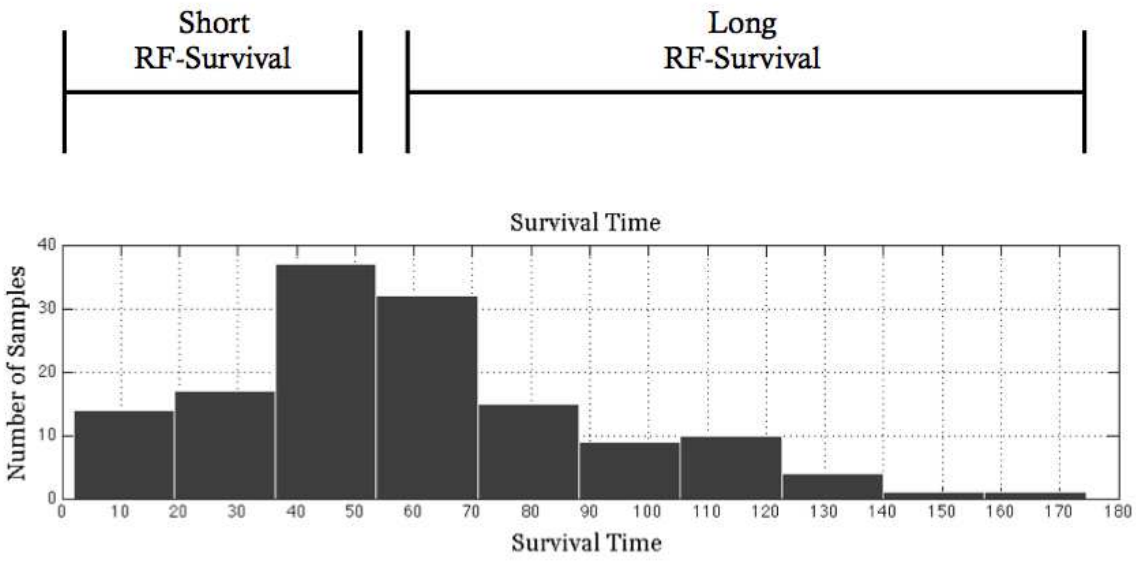


Figure 31: Recurrence Free Survival Classes

In order to examine changes in prediction accuracies with better separation in RF survival classes, the datasets were broken down into smaller sets with each subsequent dataset having 5 more samples from the 'Long' RF survival class and 5 from the 'Short' RF survival class. Refined classification accuracy was obtained by performing feature selection from with in the feature vector using Weka evaluator and search algorithms to perform feature selection. We utilized the attribute/feature selection feature of Weka and performed all available feature selection evaluators and methods.

We also performed a 3-class classification on recurrence-free survival prediction, but did not pursue it after seeing low prediction accuracies.

Results and Discussion

Our approach demonstrated the valid use of microarray profile data with respect to measuring gene expression in a predictive prognosis of prostate cancer and indicated our approach used here can be integrated with our functional impact predictive models to perform more reliable screening for clinical biomarkers. Identification of class-separable biomarkers was accomplished via classification with feature selection. Here we performed and compared a number of feature selection methods and then applied 3 different classifiers.

Features selected by almost all methods under CfsSubsetEval evaluator gave the best classification accuracies for all the three classifiers. Amongst the classifiers Support vector machine and J48 decision tree had better accuracies than RandomForest. Henceforth we will be discussing results from J48 and SVM classifiers only, further in this manuscript. Random Forest showed lower accuracies but definitely higher than random predictions. However, the accuracy distribution across incremental datasets (datasets with different number of samples) and also across different feature selection methods is very consistent with that of the distribution shown by the other two classifiers, J48 and SVM. This certainly attaches a factor of reliability to the result set.

Accuracies were high when there was optimal separation i.e. not too much or too less separation between recurrence-free survival time classes. Around this class

separation the prediction accuracies ranged from 71.6% to 63% for J48 classifier and from 67.7% to 58.8% for support vector machine (Figure 1).

Figure 32 shows accuracies plotted against datasets containing different number of samples taken from each of the recurrence-free survival classes. J48 gave the highest accuracy of 71.6% and Support Vector Machine gave a highest accuracy of 67.7%. Figure 32 also shows accuracies achieved across different feature selection methods used in this approach. Though these accuracies are not very high they certainly do provide a positive signal of classification by our approach. At the same time these observations can be considered non-random, clearly indicative signs of possible recurrence-free survival prediction based on use of gene expression data. The observations are further strongly supported by very consistent results across all the feature selection methods used and also across all the classifiers. This is shown in Figures 33-37.

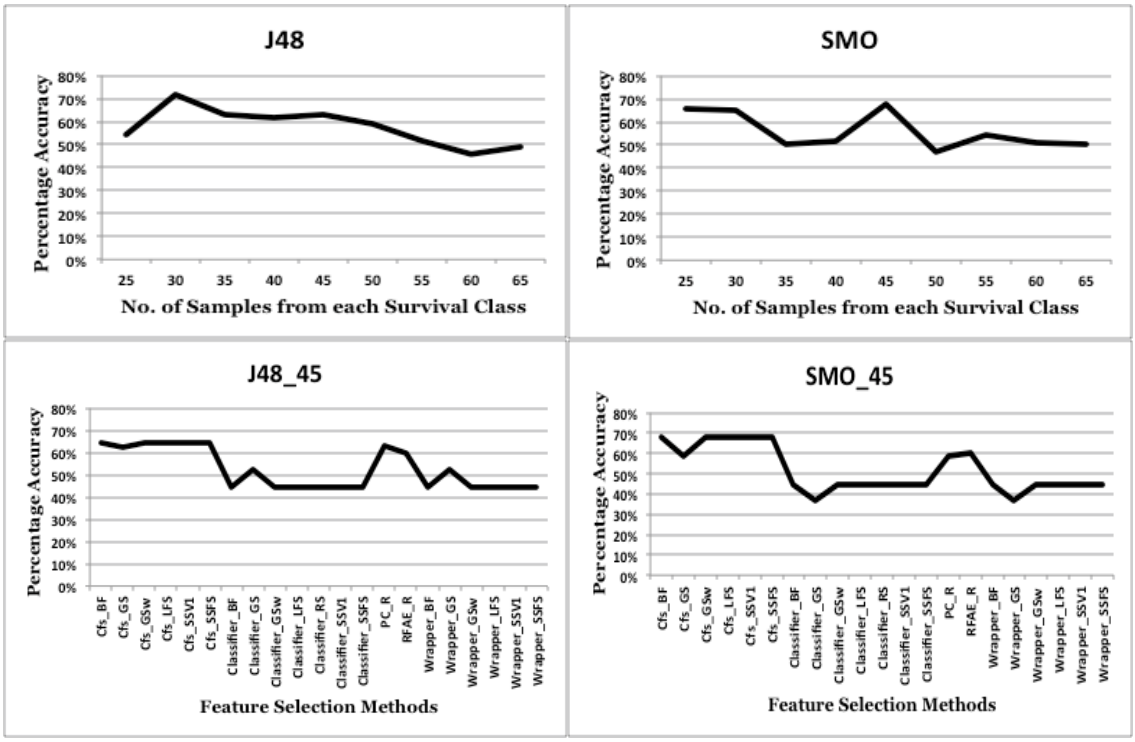


Figure 31: Percentage Accuracies from classifiers: J48 and Support Vector Machine (SMO)

These graphs clearly indicate that, each feature selection method though gives slightly different accuracy graphs they look consistent across classifiers.

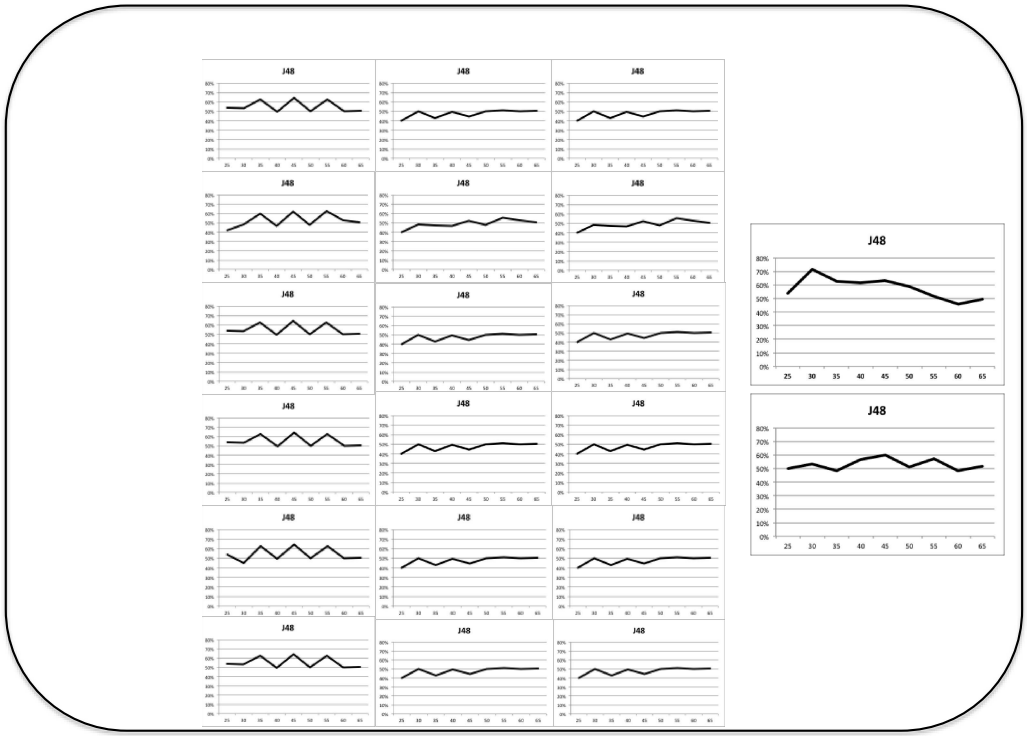


Figure 32: J48 classification accuracies across feature selection methods

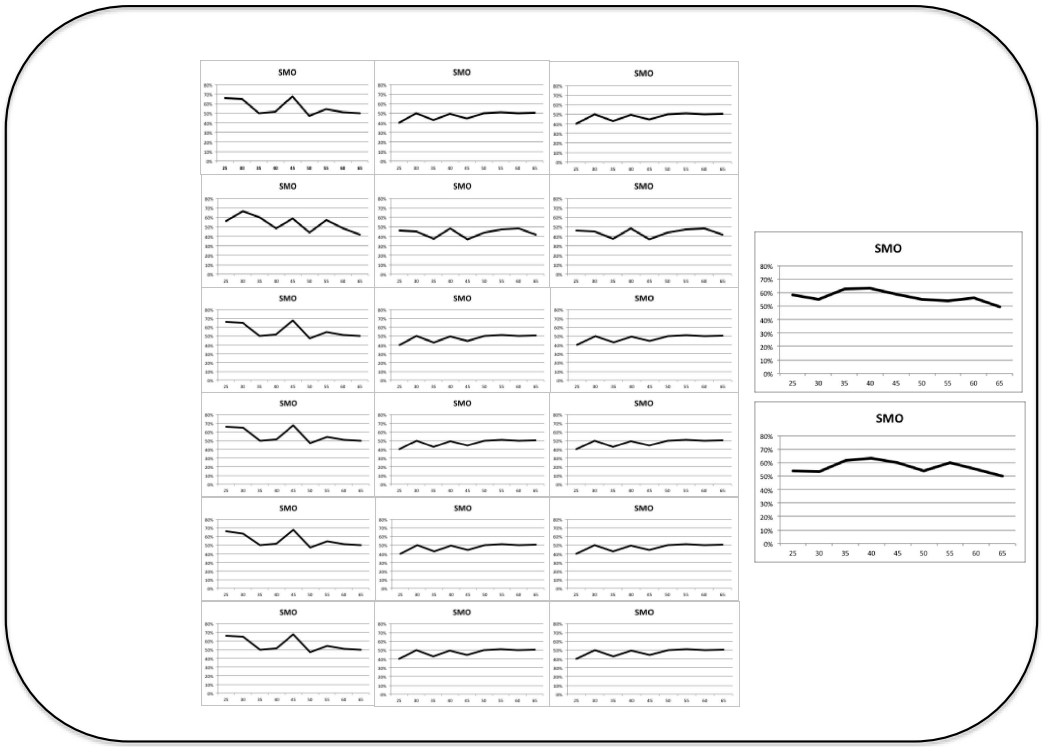


Figure 2: SVM classification accuracies across feature selection methods

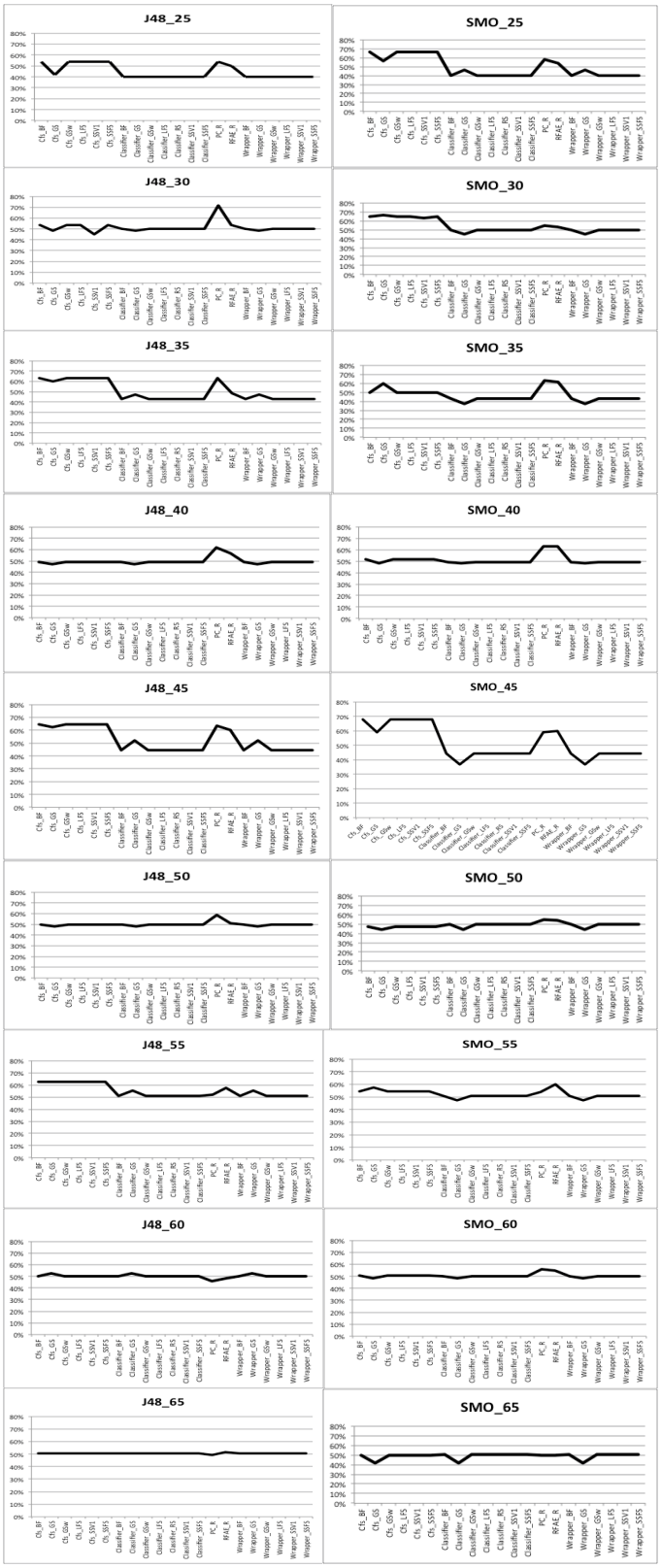


Figure 34: Show here are prediction accuracies obtained from the two classifiers J48 and SVM in two columns of graphs. The left column shows results using J48 and the right column shows those from SVM.

Finally, with the two-classes of recurrence-free survival data, performance of prediction models were evaluated with a stratified tenfold cross-validation procedure. Receiver operating characteristic (ROC) curves were plotted and AUC calculated to test the robustness of the predictions (Figure 5). As control set, the class labels were shuffled randomly and the classifiers run and ROC curves drawn.

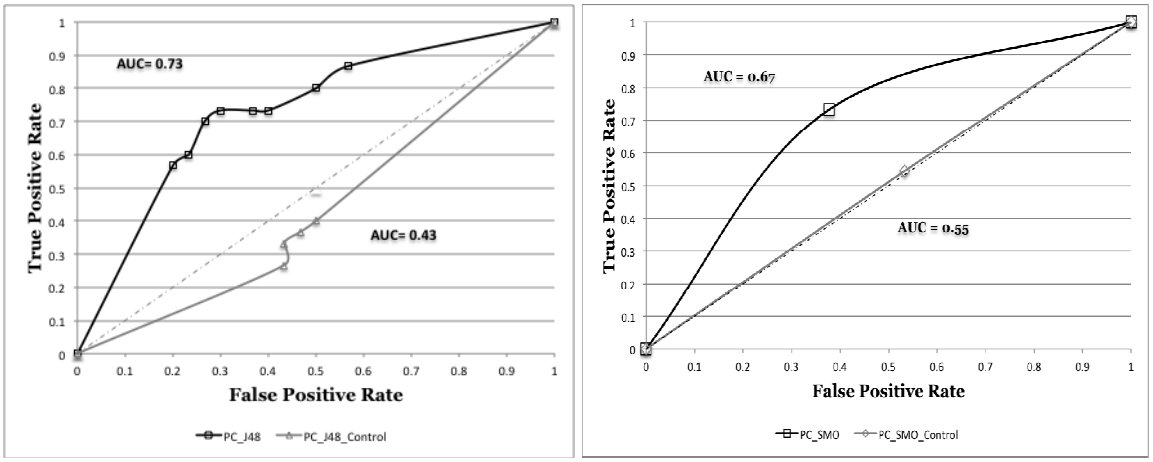


Figure 35: ROC Curves obtained by applying J48 classifier to the 70 TB dataset and SVM to the 45TB dataset.Tenfold cross-validation is used to test the learned models and generate the ROC curves, and the area under each ROC curve (AUC) provides a measure of model performance. Each training set differs by the number of samples incremented by 5 from each survival class in the next subsequent dataset. In addition to the survival class label, PC dataset used for training, is represented by the gene expression profiles of genes implicated in PC.

Genes with highest predictive accuracies of recurrence-free survival

Some of the highest predictive accuracies achieved by both J48 and SVM classifiers were using dataset with 45 samples from each recurrence-free survival

class. Five of the feature selection methods of the 'CfsSubsetEval' evaluator produced the same highest accuracy of 67.78% with both the classifiers.

The highest accuracy, 71.67%, was obtained by applying the J48 classifier, using Principal Components feature selection evaluator with the Ranker search method. Very interestingly all these high accuracies across five feature selection methods and all three classifiers had a single common gene, Endoglin, represented as ENG or CD105 as the single feature selected by feature selection methods to classify the recurrence-free survival groups. Endoglin is a Trans membrane glycoprotein, expressed by human vascular endothelial cells and is thought to play a pivotal role in endothelial cell proliferation. It is elevated in prostatic fluid of patients with large volume PCa. ENG encodes a homodimeric transmembrane protein, which is a major glycoprotein of the vascular endothelium. It is involved in the regulation of angiogenesis and may play a critical role in the binding of endothelial cells to integrins and/or other RGD receptors. It acts as TGF-beta coreceptor and is involved in the TGF-beta/BMP signaling cascade, required for GDF2/BMP9 signaling through SMAD1 in endothelial cells and modulates TGF-beta1 signaling through SMAD3.

The gene that produces the second highest accuracy of 62.8% applying J48 classifier, is ARHGDIB, Rho GDP Dissociation Inhibitor (GDI) Beta. ARHGDIB consistently produces the same accuracy and is selected by five different feature selection evaluator (CfsSubsetEval') methods, as the only feature that can

separate the recurrence-free survival classes. ARHGDIB is a member of the Rho (or ARH) protein family and other Ras-related small GTP-binding proteins that are involved in diverse cellular events, including cell signaling, proliferation, cytoskeletal organization, and secretion. Significance testing of genes differentially regulated in TMPRSS2–ERG fusion-positive prostate tumours in the Toronto cohort of 139 patients characterized for 502 genes was validated in a Swedish cohort (Setlur et al, 2008) of 455 patients characterized for 6144 genes. ARHGDIB was found to be upregulated with the TMPRSS2 – ERG fusion in both cohorts, along with 8 other genes.

The third highest accuracies are obtained by the gene, KLK2 transcript variant 2. It is a serine protease, that is highly expressed in prostate tissue and involved in regulating semen liquefaction by activating pro-KLK3 to its active form (PSA), facilitating both tumorigenisis and disease progression to the advanced stages of PCa. Studies have shown a strong correlation with PCa- specific survival however further studies with larger cohorts are needed to confirm these observations.

All these genes provide good separation of patients with good outcome from those with poorer outcome in terms of recurrence-free survival duration (Long vs. Short). All these genes with expression profiles showing significant association to survival, all represent highly relevant candidates to be examined as prognostic markers. In this work here, we have not examined the underlying quality of the expression data used, and the number of samples is also relatively

small, preventing an extensive survey of all relevant genes as candidate markers. After decreasing the size of the gene set with retaining only those with strong prognostic value, the independency of these parameters, can be assessed using our approach.

In conclusion, we believe that our approach can decipher gene expression data for survival marker predictions and can be used in combination with functional impact data to build accurate predictive model for cancer prognosis. Moreover, our can be applied to other clinical studies to increase accuracy of the methods and give it a better clinical relevance. It can be used in conjunction with other clinical predictive tools to develop prognostic genes or gene signatures of cancer development and progression, with high reliability.

Future Directions

A very interesting extension of research work presented here and an immediate future task that can be accomplished using the predictive approach proposed and validated in this study is to integrate the mutational data and the gene expression data and build predictive models for cancer prediction and prognosis. Exemplary studies, detailed in the 'Integrative Analysis' section of chapter three in this manuscript, have been carried out by other researchers and have shown promising results. Results from our approach using topological scores of wild-type and mutant proteins show significant structure-function correlations and machine learning classifiers trained using the residual profiles of these mutants have provided impressive model performances based on accuracy measures and ROC curves. Similarly, models build for survival prediction based on our approach using gene expression profiles also indicate positive pointers towards having efficient predictive power for survival prediction. Thus a logical and next step would be to integrate these analyses and build stronger, more reliable, clinically significant models predicting structure to function to survival correlations.

Another promising and very useful extension to this work would be automating and maintaining the integrated database IDHCMM. IDHCMM was built as a

basic one-stop-shop service for obtaining huge experimentally validated missense mutational data sets enhanced with availability of x-ray crystallographic 3D structures cross-linked to all of the mutations. However, this repository can be further improved in many ways, mentioned in chapter two of this dissertation. One of the most required and useful improvements would be to updating the data in IDHCMM by downloading and restoring latest data from the source databases, especially the protein 3D structures from PDB. This task does have its own challenges when working with the comprehensive data from TCGA and ICGC, but can be managed by designing a data frame for IDHCMM and sticking to it and not making data changes along the sources. This useful resource can be made more widely used by expanding mutational effect prediction data for the mutations by adding predictions obtained by Delaunay tessellation approach used here and also from other renowned predictive tools such as MutationAssessor, SIFT and PolyPhen. As of now only very few records have these predictions associated to them. At the next level, IDHCMM has links to neXtProt and PharmGKB Ids. This can be utilized to complement mutational models and associate them to higher level protein data from neXtProt and further to potentially clinically actionable gene-drug associations and genotype-phenotype relationships thereby deciphering knowledge about the impact of human genetic variation on drug responses. This entire knowledge mine can then be used to predict cancer outcomes and design targeted clinical therapies for cancer.

Finally the long term objectives of this research work is to be able to contribute to the technological revolution driving our pursuit of transforming approaches able of molecular characterization, enabling proper and timely cancer management with improved patient outcomes. This improved link between the ability to characterize the cancer genome and changes in tumors for biomarker identification, predicting patient response to targeted therapies, is a step further towards implementation of personalized cancer therapy.

Supplementary Material

Table 11: IDHCMM Data elements and mapping to source database-data elements

IDHCMM Data Elements	**MSKCC_Sarcoma**	**MSKCC_Prostate**	**TCGA**	**COSMIC**	**BIC db**	**ICGC**	**IARC_TP53**
MutationID							
SourceTableID							
SourceDB	MSKCC_Sarcoma	MSKCC_Prostate	TCGA	COSMIC	BIC	ICGC	IARC_TP53
Source_Version			v2.3	COSMIC61		Rel 9	R16
SourceDB_Mutation_ID				COSMIC Mutation ID (520)	accession number (4347)	Mutation_ID	MUT_ID
SourceDB_Sample_ID	Tumor_Sample_Barcode (PT10DD)	Sample	Tumor_Sample_Barcode	COSMIC Sample ID		Analyzed_sample_ID	Sample_ID
SourceDB_Patient_ID					ID Number (F2743)	Donor_ID	
CDS_Mutation	cDNA_Change_Broad		ChromChange	CDS Mutation Syntax (c.1500C>G)	HGVS cDNA (c.181T>G)	CDS mutation	c_description
CDS_Mutation_Type				CDS Mutation Type (Substitution)		Mutation Type	
CDS_Mutation_Start			ChromChange	CDS Mutation Start (1500)		CDS mutation	c_description
CDS_Mutation_Stop				CDS Mutation Stop (1500)		CDS mutation	c_description
AA_Mutation	FAM_variant (A212T)	Protein (W742C)	AAChange	AA Mutation Syntax (p.I500M)	BIC Designation (M1V)	AA mutation	ProtDescription
AA_Mutation_Type	Variant_Clas	Mutation	Variant_Clas	AA Mutation	mutation type	Consequence	Effect

	sification (Missense)	type (Missense)	sification (Missense_ Mutation)	Type (Substitution - Missense)	(M)	type (non-synonymousCoding)	
AA_Mutation_Start_ Position				AA Mutation Start (500)			ProtDescription
AA_Mutation_Stop_ Position				AA Mutation Stop (500)			ProtDescription
Genomic_Mutation		Reference> Mutant	Genome_Change		Base Change (A to G)	Mutation	Description
Genomic_Mutation_ Start_Position	Start_Position (56934311)	Position (66854097)	Start_Position (56934311)	Genomic Coordinates (NCBI36)		Chromosome start	Genomic_nt
Genomic_Mutation_ Stop_Position	End_Position (56934311)		End_Position (56934311)	Genomic Coordinates (NCBI36)		Chromosome end	
Gene_Name	Hugo_Symbol (PTPN14)	Gene (AR)	Hugo_Symbol (PTPN14)	Gene Name (KRAS)	BRCA1 / BRCA2	Gene _Name	TP53
Chromosome	Chromosome (1)	Chr (chr17)	Chromosome (1)	Genomic Coordinates (NCBI36)- (12:25289551-25289551)		Chromosome (12)	
Mutation_Validation _Status	Validation_Status (Valid)		Validation_Status	Validation_Status	Mutation_Validation_Status	Validation_Status	
Mutation_Detection_ Platform	Sequencer	Method (Sanger)	Sequencer		Detection Method	Platform	
Mutation_Validation _Platform	Validation_ Method		Validation_ Method			Validation platform	
Domain_Affected	domain_WU	Affected domain (PFAM) - (PF00104 // Ligand-binding domain of nuclear hormone receptor)	domain				
MutationAssessor_Prediction	FAM_Fimpact(M)	Predicted Funtional Impact					

		(High)					
Polyphen_Prediction			polyphen				
SIFT_Prediction			sift				SIFTClass
MSA_Link	FAM_link_MSA						
PDB_Link	FAM_link_PDB						
dbSNP_ID	dbSNP_RS		dbSNP_RS		dbSNP		
Mapping_ID							
Entrez_Gene_ID	Entrez_Gene_Id (2041)		Entrez_Gene_ID (1956)	Entrez Gene ID (3845)			
Swissprot_AC_ID			Swissprot_ACC_ID	Swissprot ID (P01116)			
Swissprot_Entry_ID			Swissprot_Entry_ID				
Ensembl_Gene_ID				Ensembl Gene ID (ENSG00000176601)		Ensembl_Gene_ID	
Transcript_ID	transcript_name_WU		TranscriptID	Gene_Name(ENST);Accession Number (ENST)		Transcript_affected	
Refseq_ID		Refseq ID (NM_000044)	Refseq_Prot_ID	Gene_Name (NM_, XM_)			
Pfam_Accession_ID	FAM_Pfam_domain	Affected domain (PFAM) - (PF00104 // Ligand-binding domain of nuclear hormone receptor)					
PubMed_ID				Pubmed ID			PubMed
Ethnicity					Ethnicity		Ethnicity
Cancer_Type					Breast Cancer	Cancer Type(Chronic	

						Lymphocytic Leukemia (ISC/MICINN, ES))	
Primary_Site				Primary Site			Topography
Sample_Name							Sample_Name
Sample_Source				Sample_Source			Sample_Source
Primary_Histology				Primary_Histology			
Site_Subtype_1				Site_Subtype_1			
Site_Subtype_2				Site_Subtype_2			
Site_Subtype_3				Site_Subtype_3			
Histology_Subtype_1				Histology_Subtype_1			
Histology_Subtype_2				Histology_Subtype_2			
Histology_Subtype_3				Histology_Subtype_3			
Sample_Type						Sample_Type	
Normal_Sample_ID						Matched_Sample_ID	
Specimen_ID						Specimen_ID	
Specimen_Type						Specimen_Type	
Gender						Sex	Sex
Age_at_Diagnosis						Age_at_Diagnosis	Age
Age_at_Enrollment						Age_at_Enrollment	
Population						Population	
Country					Nationality	Country	
Geo_Area						Geo_Area	
Family_History						Family_History	
Tobacco						Tobacco	
Alcohol						Alcohol	
Exposure						Exposure	
Infectious_Agent						Infectious_Agent	
Depositor					Depositor		
Number_Reported					Number_Repor		

					ted		
Contact_Person					Contact_Perso n		
Tumour_Source				Tumour_Source			
Tumor_Seq_Allele1			Tumor_Seq_ Allele1				
Tumor_Seq_Allele2			Tumor_Seq_ Allele2				
Tumor_Sample_Barc ode			Tumor_Sam ple_Barcode				
Matched_Norm_Sam ple_Barcode			Matched_No rm_Sample_ Barcode				
Match_Norm_Seq_A llele1			Match_Nor m_Seq_Allel e1				
Match_Norm_Seq_A llele2			Match_Nor m_Seq_Allel e2				
Tumor_Validation_A llele1			Tumor_Vali dation_Allel e1				
Tumor_Validation_A llele2			Tumor_Vali dation_Allel e2				
Match_Norm_Valida tion_Allele1			Match_Nor m_Validatio n_Allele1				
Match_Norm_Valida tion_Allele2			Match_Nor m_Validatio n_Allele2				
normal_depth			normal_dept h				
normal_vaf			normal_vaf				
tumor_depth			tumor_depth				
tumor_vaf			tumor_vaf				
rna_depth			rna_depth				
rna_vaf			rna_vaf				
Tumour_Confirmed						Tumour_Confirm ed	

Tumour_Grade						Tumour_Grade	Grade
Control_Genotype							
Tumour_Genotype							
Donor_Tumour_Stage_at_Diagnosis						Donor_Tumour_Stage_at_Diagnosis	
Tumour_Stage						Tumour_Stage	Stage
ICD_10						ICD_10	
Short_Topo							Short_Topo
Topo_Code							Topo_Code
Morphology							Morphology
Morpho_Code							Morpho_Code
TNM							TNM
p53_IHC							p53_IHC
Specimen_Donor_Treatment_Type						Specimen_Donor_Treatment_Type	
Specimen_Donor_Treatment_Type_Other						Specimen_Donor_Treatment_Type_Other	
Add_Info					Notes		Add_Info

REFERENCES

[1] H. Liang, L. W. T. Cheung, J. Li, Z. Ju, S. Yu, K. Stemke-Hale, T. Dogruluk, Y. Lu, X. Liu, C. Gu, W. Guo, S. E. Scherer, H. Carter, S. N. Westin, M. D. Dyer, R. G. W. Verhaak, F. Zhang, R. Karchin, C.-G. Liu, K. H. Lu, R. R. Broaddus, K. L. Scott, B. T. Hennessy, and G. B. Mills, "Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer," *Genome Res.*, vol. 22, no. 11, pp. 2120–2129, Oct. 2012.

[2] L. Ding, T. J. Ley, D. E. Larson, C. A. Miller, D. C. Koboldt, J. S. Welch, J. K. Ritchey, M. A. Young, T. Lamprecht, M. D. McLellan, J. F. McMichael, J. W. Wallis, C. Lu, D. Shen, C. C. Harris, D. J. Dooling, R. S. Fulton, L. L. Fulton, K. Chen, H. Schmidt, J. Kalicki-Veizer, V. J. Magrini, L. Cook, S. D. McGrath, T. L. Vickery, M. C. Wendl, S. Heath, M. A. Watson, D. C. Link, M. H. Tomasson, W. D. Shannon, J. E. Payton, S. Kulkarni, P. Westervelt, M. J. Walter, T. A. Graubert, E. R. Mardis, R. K. Wilson, and J. F. DiPersio, "Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing," *Nature*, vol. 481, no. 7382, pp. 506–510, Jan. 2012.

[3] C. E. Barbieri, S. C. Baca, M. S. Lawrence, F. Demichelis, M. Blattner, J.-P. Theurillat, T. A. White, P. Stojanov, E. Van Allen, N. Stransky, E. Nickerson, S.-S. Chae, G. Boysen, D. Auclair, R. C. Onofrio, K. Park, N. Kitabayashi, T. Y. MacDonald, K. Sheikh, T. Vuong, C. Guiducci, K. Cibulskis, A. Sivachenko, S. L. Carter, G. Saksena, D. Voet, W. M. Hussain, A. H. Ramos, W. Winckler, M. C. Redman, K. Ardlie, A. K. Tewari, J. M. Mosquera, N. Rupp, P. J. Wild, H. Moch, C. Morrissey, P. S. Nelson, P. W. Kantoff, S. B. Gabriel, T. R. Golub, M. Meyerson, E. S. Lander, G. Getz, M. A. Rubin, and L. A. Garraway, "Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer," *Nat. Genet.*, vol. 44, no. 6, pp. 685–689, May 2012.

[4] D. Tian, "Remarkable difference of somatic mutation patterns between oncogenes and tumor suppressor genes," *Oncol. Rep.*, Sep. 2011.

[5] C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson,

R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y.-E. Chiew, A. deFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M.-H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton, "Patterns of somatic mutation in human cancer genomes," *Nature*, vol. 446, no. 7132, pp. 153–158, Mar. 2007.

[6] E. D. Pleasance, R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M.-L. Lin, G. R. Ordóñez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton, "A comprehensive catalogue of somatic mutations from a human cancer genome," *Nature*, vol. 463, no. 7278, pp. 191–196, Dec. 2009.

[7] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjoblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, R. Croshaw, J. Willis, D. Dawson, M. Shipitsin, J. K. V. Willson, S. Sukumar, K. Polyak, B. H. Park, C. L. Pethiyagoda, P. V. K. Pant, D. G. Ballinger, A. B. Sparks, J. Hartigan, D. R. Smith, E. Suh, N. Papadopoulos, P. Buckhaults, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, V. E. Velculescu, and B. Vogelstein, "The Genomic Landscapes of Human Breast and Colorectal Cancers," *Science*, vol. 318, no. 5853, pp. 1108–1113, Nov. 2007.

[8] E. D. Green, M. S. Guyer, E. D. Green, M. S. Guyer, T. A. Manolio, and J. L. Peterson, "Charting a course for genomic medicine from base pairs to bedside," *Nature*, vol. 470, no. 7333, pp. 204–213, Feb. 2011.

[9] E. R. Mardis, "A decade's perspective on DNA sequencing technology," *Nature*, vol. 470, no. 7333, pp. 198–203, Feb. 2011.

[10] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, G. M. Mastrogianakis, J. J. Olson, T. Mikkelsen, N. Lehman, K. Aldape, W. K. Alfred Yung, O. Bogler, S. VandenBerg, M. Berger, M. Prados, D. Muzny, M. Morgan, S. Scherer, A. Sabo, L. Nazareth, L. Lewis, O. Hall, Y. Zhu, Y. Ren, O. Alvi, J. Yao, A. Hawes, S. Jhangiani, G. Fowler, A. San Lucas, C. Kovar, A. Cree, H. Dinh, J. Santibanez, V. Joshi, M. L. Gonzalez-Garay, C. A. Miller, A. Milosavljevic, L. Donehower, D. A. Wheeler, R. A. Gibbs, K. Cibulskis, C. Sougnez, T. Fennell, S. Mahan, J. Wilkinson, L. Ziaugra, R. Onofrio, T. Bloom, R. Nicol, K. Ardlie, J. Baldwin, S. Gabriel, E. S. Lander, L. Ding, R. S. Fulton, M. D. McLellan, J. Wallis,

D. E. Larson, X. Shi, R. Abbott, L. Fulton, K. Chen, D. C. Koboldt, M. C. Wendl, R. Meyer, Y. Tang, L. Lin, J. R. Osborne, B. H. Dunford-Shore, T. L. Miner, K. Delehaunty, C. Markovic, G. Swift, W. Courtney, C. Pohl, S. Abbott, A. Hawkins, S. Leong, C. Haipek, H. Schmidt, M. Wiechert, T. Vickery, S. Scott, D. J. Dooling, A. Chinwalla, G. M. Weinstock, E. R. Mardis, R. K. Wilson, G. Getz, W. Winckler, R. G. W. Verhaak, M. S. Lawrence, M. O'Kelly, J. Robinson, G. Alexe, R. Beroukhim, S. Carter, D. Chiang, J. Gould, S. Gupta, J. Korn, C. Mermel, J. Mesirov, S. Monti, H. Nguyen, M. Parkin, M. Reich, N. Stransky, B. A. Weir, L. Garraway, T. Golub, M. Meyerson, L. Chin, A. Protopopov, J. Zhang, I. Perna, S. Aronson, N. Sathiamoorthy, G. Ren, J. Yao, W. R. Wiedemeyer, H. Kim, S. Won Kong, Y. Xiao, I. S. Kohane, J. Seidman, P. J. Park, R. Kucherlapati, P. W. Laird, L. Cope, J. G. Herman, D. J. Weisenberger, F. Pan, D. Van Den Berg, L. Van Neste, J. Mi Yi, K. E. Schuebel, S. B. Baylin, D. M. Absher, J. Z. Li, A. Southwick, S. Brady, A. Aggarwal, T. Chung, G. Sherlock, J. D. Brooks, R. M. Myers, P. T. Spellman, E. Purdom, L. R. Jakkula, A. V. Lapuk, H. Marr, S. Dorton, Y. Gi Choi, J. Han, A. Ray, V. Wang, S. Durinck, M. Robinson, N. J. Wang, K. Vranizan, V. Peng, E. Van Name, G. V. Fontenay, J. Ngai, J. G. Conboy, B. Parvin, H. S. Feiler, T. P. Speed, J. W. Gray, C. Brennan, N. D. Socci, A. Olshen, B. S. Taylor, A. Lash, N. Schultz, B. Reva, Y. Antipin, A. Stukalov, B. Gross, E. Cerami, W. Qing Wang, L.-X. Qin, V. E. Seshan, L. Villafania, M. Cavatore, L. Borsu, A. Viale, W. Gerald, C. Sander, M. Ladanyi, C. M. Perou, D. Neil Hayes, M. D. Topal, K. A. Hoadley, Y. Qi, S. Balu, Y. Shi, J. Wu, R. Penny, M. Bittner, T. Shelton, E. Lenkiewicz, S. Morris, D. Beasley, S. Sanders, A. Kahn, R. Sfeir, J. Chen, D. Nassau, L. Feng, E. Hickey, J. Zhang, J. N. Weinstein, A. Barker, D. S. Gerhard, J. Vockley, C. Compton, J. Vaught, P. Fielding, M. L. Ferguson, C. Schaefer, S. Madhavan, K. H. Buetow, F. Collins, P. Good, M. Guyer, B. Ozenberger, J. Peterson, and E. Thomson, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, pp. 1061–1068, Sep. 2008.

[11] S. A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. W. Teague, P. A. Futreal, and M. R. Stratton, "The Catalogue of Somatic Mutations in Cancer (COSMIC)," *Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al*, vol. Chapter 10, p. Unit 10.11, Apr. 2008.

[12] T. J. Hudson (Chairperson), W. Anderson, A. Aretz, A. D. Barker, C. Bell, R. R. Bernabé, M. K. Bhan, F. Calvo, I. Eerola, D. S. Gerhard, A. Guttmacher, M. Guyer, F. M. Hemsley, J. L. Jennings, D. Kerr, P. Klatt, P. Kolar, J. Kusuda, D. P. Lane, F. Laplace, Y. Lu, G. Nettekoven, B. Ozenberger, J. Peterson, T. S. Rao, J. Remacle, A. J. Schafer, T. Shibata, M. R. Stratton, J. G. Vockley, K. Watanabe, H. Yang, M. M. F. Yuen, B. M. Knoppers (Leader), M. Bobrow, A. Cambon-Thomsen, L. G. Dressler, S. O. M. Dyke, Y. Joly, K. Kato, K. L. Kennedy, P. Nicolás, M. J. Parker, E. Rial-Sebbag, C. M. Romeo-Casabona, K. M. Shaw, S. Wallace, G. L. Wiesner, N. Zeps, P. Lichter (Leader), A. V. Biankin, C. Chabannon, L. Chin, B. Clément, E. de Alava, F. Degos, M. L. Ferguson, P. Geary, D. N. Hayes, T. J. Hudson, A. L. Johns,

A. Kasprzyk, H. Nakagawa, R. Penny, M. A. Piris, R. Sarin, A. Scarpa, T. Shibata, M. van de Vijver, P. A. Futreal (Leader), H. Aburatani, M. Bayés, D. D. L. Bowtell, P. J. Campbell, X. Estivill, D. S. Gerhard, S. M. Grimmond, I. Gut, M. Hirst, C. López-Otín, P. Majumder, M. Marra, J. D. McPherson, H. Nakagawa, Z. Ning, X. S. Puente, Y. Ruan, T. Shibata, M. R. Stratton, H. G. Stunnenberg, H. Swerdlow, V. E. Velculescu, R. K. Wilson, H. H. Xue, L. Yang, P. T. Spellman (Leader), G. D. Bader, P. C. Boutros, P. J. Campbell, P. Flicek, G. Getz, R. Guigó, G. Guo, D. Haussler, S. Heath, T. J. Hubbard, T. Jiang, S. M. Jones, Q. Li, N. López-Bigas, R. Luo, L. Muthuswamy, B. F. Francis Ouellette, J. V. Pearson, X. S. Puente, V. Quesada, B. J. Raphael, C. Sander, T. Shibata, T. P. Speed, L. D. Stein, J. M. Stuart, J. W. Teague, Y. Totoki, T. Tsunoda, A. Valencia, D. A. Wheeler, H. Wu, S. Zhao, G. Zhou, L. D. Stein (Leader), R. Guigó, T. J. Hubbard, Y. Joly, S. M. Jones, A. Kasprzyk, M. Lathrop, N. López-Bigas, B. F. Francis Ouellette, P. T. Spellman, J. W. Teague, G. Thomas, A. Valencia, T. Yoshida, K. L. Kennedy (Leader), M. Axton, S. O. M. Dyke, P. A. Futreal, D. S. Gerhard, C. Gunter, M. Guyer, T. J. Hudson, J. D. McPherson, L. J. Miller, B. Ozenberger, K. M. Shaw, A. Kasprzyk (Leader), L. D. Stein (Leader), J. Zhang, S. A. Haider, J. Wang, C. K. Yung, A. Cross, Y. Liang, S. Gnaneshan, J. Guberman, J. Hsu, M. Bobrow (Leader), D. R. C. Chalmers, K. W. Hasel, Y. Joly, T. S. H. Kaan, K. L. Kennedy, B. M. Knoppers, W. W. Lowrance, T. Masui, P. Nicolás, E. Rial-Sebbag, L. Lyman Rodriguez, C. Vergely, T. Yoshida, S. M. Grimmond (Leader), A. V. Biankin, D. D. L. Bowtell, N. Cloonan, A. deFazio, J. R. Eshleman, D. Etemadmoghadam, B. A. Gardiner, J. G. Kench, A. Scarpa, R. L. Sutherland, M. A. Tempero, N. J. Waddell, P. J. Wilson, J. D. McPherson (Leader), S. Gallinger, M.-S. Tsao, P. A. Shaw, G. M. Petersen, D. Mukhopadhyay, L. Chin, R. A. DePinho, S. Thayer, L. Muthuswamy, K. Shazand, T. Beck, M. Sam, L. Timms, V. Ballin, Y. Lu (Leader), J. Ji, X. Zhang, F. Chen, X. Hu, G. Zhou, Q. Yang, G. Tian, L. Zhang, X. Xing, X. Li, Z. Zhu, Y. Yu, J. Yu, H. Yang, M. Lathrop (Leader), J. Tost, P. Brennan, I. Holcatova, D. Zaridze, A. Brazma, L. Egevad, E. Prokhortchouk, R. Elizabeth Banks, M. Uhlén, A. Cambon-Thomsen, J. Viksna, F. Ponten, K. Skryabin, M. R. Stratton (Leader), P. A. Futreal, E. Birney, A. Borg, A.-L. Børresen-Dale, C. Caldas, J. A. Foekens, S. Martin, J. S. Reis-Filho, A. L. Richardson, C. Sotiriou, H. G. Stunnenberg, G. Thomas, M. van de Vijver, L. van't Veer, F. Calvo (Leader), D. Birnbaum, H. Blanche, P. Boucher, S. Boyault, C. Chabannon, I. Gut, J. D. Masson-Jacquemier, M. Lathrop, I. Pauporté, X. Pivot, A. Vincent-Salomon, E. Tabone, C. Theillet, G. Thomas, J. Tost, I. Treilleux, F. Calvo (Leader), P. Bioulac-Sage, B. Clément, T. Decaens, F. Degos, D. Franco, I. Gut, M. Gut, S. Heath, M. Lathrop, D. Samuel, G. Thomas, J. Zucman-Rossi, P. Lichter (Leader), R. Eils (Leader), B. Brors, J. O. Korbel, A. Korshunov, P. Landgraf, H. Lehrach, S. Pfister, B. Radlwimmer, G. Reifenberger, M. D. Taylor, C. von Kalle, P. P. Majumder (Leader), R. Sarin, T. S. Rao, M. K. Bhan, A. Scarpa (Leader), P. Pederzoli, R. T. Lawlor, M. Delledonne, A. Bardelli, A. V. Biankin, S. M. Grimmond, T. Gress, D. Klimstra, G. Zamboni, T. Shibata (Leader), Y. Nakamura, H. Nakagawa, J. Kusuda, T. Tsunoda, S. Miyano, H. Aburatani, K. Kato, A. Fujimoto, T. Yoshida, E. Campo (Leader), C. López-Otín, X. Estivill, R. Guigó, S. de

Sanjosé, M. A. Piris, E. Montserrat, M. González-Díaz, X. S. Puente, P. Jares, A. Valencia, H. Himmelbaue, V. Quesada, S. Bea, M. R. Stratton (Leader), P. A. Futreal, P. J. Campbell, A. Vincent-Salomon, A. L. Richardson, J. S. Reis-Filho, M. van de Vijver, G. Thomas, J. D. Masson-Jacquemier, S. Aparicio, A. Borg, A.-L. Børresen-Dale, C. Caldas, J. A. Foekens, H. G. Stunnenberg, L. van't Veer, D. F. Easton, P. T. Spellman, S. Martin, A. D. Barker, L. Chin, F. S. Collins, C. C. Compton, M. L. Ferguson, D. S. Gerhard, G. Getz, C. Gunter, A. Guttmacher, M. Guyer, D. N. Hayes, E. S. Lander, B. Ozenberger, R. Penny, J. Peterson, C. Sander, K. M. Shaw, T. P. Speed, P. T. Spellman, J. G. Vockley, D. A. Wheeler, R. K. Wilson, T. J. Hudson (Chairperson), L. Chin, B. M. Knoppers, E. S. Lander, P. Lichter, L. D. Stein, M. R. Stratton, W. Anderson, A. D. Barker, C. Bell, M. Bobrow, W. Burke, F. S. Collins, C. C. Compton, R. A. DePinho, D. F. Easton, P. A. Futreal, D. S. Gerhard, A. R. Green, M. Guyer, S. R. Hamilton, T. J. Hubbard, O. P. Kallioniemi, K. L. Kennedy, T. J. Ley, E. T. Liu, Y. Lu, P. Majumder, M. Marra, B. Ozenberger, J. Peterson, A. J. Schafer, P. T. Spellman, H. G. Stunnenberg, B. J. Wainwright, R. K. Wilson, and H. Yang, "International network of cancer genome projects," *Nature*, vol. 464, no. 7291, pp. 993–998, Apr. 2010.

[13] J. G. Lohr, P. Stojanov, S. L. Carter, P. Cruz-Gordillo, M. S. Lawrence, D. Auclair, C. Sougnez, B. Knoechel, J. Gould, G. Saksena, K. Cibulskis, A. McKenna, M. A. Chapman, R. Straussman, J. Levy, L. M. Perkins, J. J. Keats, S. E. Schumacher, M. Rosenberg, G. Getz, and T. R. Golub, "Widespread Genetic Heterogeneity in Multiple Myeloma: Implications for Targeted Therapy," *Cancer Cell*, vol. 25, no. 1, pp. 91–101, Jan. 2014.

[14] C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, M. D. M. Leiserson, C. A. Miller, J. S. Welch, M. J. Walter, M. C. Wendl, T. J. Ley, R. K. Wilson, B. J. Raphael, and L. Ding, "Mutational landscape and significance across 12 major cancer types," *Nature*, vol. 502, no. 7471, pp. 333–339, Oct. 2013.

[15] M. -d.-M. Inda, R. Bonavia, A. Mukasa, Y. Narita, D. W. Y. Sah, S. Vandenberg, C. Brennan, T. G. Johns, R. Bachoo, P. Hadwiger, P. Tan, R. A. DePinho, W. Cavenee, and F. Furnari, "Tumor heterogeneity is an active process maintained by a mutant EGFR-induced cytokine circuit in glioblastoma," *Genes Dev.*, vol. 24, no. 16, pp. 1731–1745, Aug. 2010.

[16] D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, "The Next-Generation Sequencing Revolution and Its Impact on Genomics," *Cell*, vol. 155, no. 1, pp. 27–38, Sep. 2013.

[17] L. A. Garraway and E. S. Lander, "Lessons from the Cancer Genome," *Cell*, vol. 153, no. 1, pp. 17–37, Mar. 2013.

[18] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, “Cancer Genome Landscapes,” *Science*, vol. 339, no. 6127, pp. 1546–1558, Mar. 2013.

[19] T. Sjoblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. V. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, and V. E. Velculescu, “The Consensus Coding Sequences of Human Breast and Colorectal Cancers,” *Science*, vol. 314, no. 5797, pp. 268–274, Oct. 2006.

[20] Z. Shi and J. Moult, “Structural and Functional Impact of Cancer-Related Missense Somatic Mutations,” *J. Mol. Biol.*, vol. 413, no. 2, pp. 495–512, Oct. 2011.

[21] S. Stefl, H. Nishi, M. Petukh, A. R. Panchenko, and E. Alexov, “Molecular Mechanisms of Disease-Causing Missense Mutations,” *J. Mol. Biol.*, vol. 425, no. 21, pp. 3919–3936, Nov. 2013.

[22] H. Nishi, M. Tyagi, S. Teng, B. A. Shoemaker, K. Hashimoto, E. Alexov, S. Wuchty, and A. R. Panchenko, “Cancer Missense Mutations Alter Binding Properties of Proteins and Their Interaction Networks,” *PLoS ONE*, vol. 8, no. 6, p. e66273, Jun. 2013.

[23] H. Carter, S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, and R. Karchin, “Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations,” *Cancer Res.*, vol. 69, no. 16, pp. 6660–6667, Aug. 2009.

[24] M. Krawczak, E. V. Ball, I. Fenton, P. D. Stenson, S. Abeysinghe, N. Thomas, and D. N. Cooper, “Human gene mutation database—a biomedical information and research resource,” *Hum. Mutat.*, vol. 15, no. 1, pp. 45–51, 2000.

[25] A. Azia, V. N. Uversky, A. Horovitz, and R. Unger, “The Effects of Mutations on Protein Function: A Comparative Study of Three Databases of Mutations in Humans,” *Isr. J. Chem.*, vol. 53, no. 3–4, pp. 217–226, Apr. 2013.

[26] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, “Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2,” in *Current Protocols in Human Genetics*, J. L. Haines, B. R. Korf, C. C. Morton, C. E. Seidman, J. G. Seidman, and D. R. Smith, Eds. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013.

[27] Y. Mao, H. Chen, H. Liang, F. Meric-Bernstam, G. B. Mills, and K. Chen, "CanDrA: Cancer-Specific Driver Missense Mutation Annotation with Optimized Features," *PLoS ONE*, vol. 8, no. 10, p. e77945, Oct. 2013.

[28] H.-H. Won, J.-W. Kim, and D. Lee, "A Bayesian ensemble approach with a disease gene network predicts damaging effects of missense variants of human cancers," *Hum. Genet.*, vol. 132, no. 1, pp. 15–27, Aug. 2012.

[29] T. M. K. Cheng, L. Goehring, L. Jeffery, Y.-E. Lu, J. Hayles, B. Novák, and P. A. Bates, "A Structural Systems Biology Approach for Quantifying the Systemic Consequences of Missense Mutations in Proteins," *PLoS Comput. Biol.*, vol. 8, no. 10, p. e1002738, Oct. 2012.

[30] E. Capriotti and R. B. Altman, "A new disease-specific machine learning approach for the prediction of cancer-causing missense variants," *Genomics*, vol. 98, no. 4, pp. 310–317, Oct. 2011.

[31] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: application to cancer genomics," *Nucleic Acids Res.*, p. gkr407, Jul. 2011.

[32] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nat. Methods*, vol. 7, no. 4, pp. 248–249, Apr. 2010.

[33] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nat. Protoc.*, vol. 4, no. 8, pp. 1073–1081, Jun. 2009.

[34] S. Teng, E. Michonova-Alexova, and E. Alexov, "Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions," *Curr. Pharm. Biotechnol.*, vol. 9, no. 2, pp. 123–133, 2008.

[35] M. Barenboim, M. Masso, I. I. Vaisman, and D. C. Jamison, "Statistical geometry based prediction of nonsynonymous SNP functional effects using random forest and neuro-fuzzy classifiers," *Proteins Struct. Funct. Bioinforma.*, vol. 71, no. 4, pp. 1930–1939, Jan. 2008.

[36] C. J. Needham, J. R. Bradford, A. J. Bulpitt, and D. R. Westhead, "Predicting the effect of missense mutations on protein function: analysis with Bayesian networks," *BMC Bioinformatics*, vol. 7, no. 1, p. 405, 2006.

[37] V. G. Krishnan and D. R. Westhead, “A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function,” *Bioinformatics*, vol. 19, no. 17, pp. 2199–2209, Nov. 2003.

[38] D. Chasman and R. M. Adams, “Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation,” *J. Mol. Biol.*, vol. 307, no. 2, pp. 683–706, Mar. 2001.

[39] I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein, and M. A. Nowak, “Accumulation of driver and passenger mutations during tumor progression,” *Proc. Natl. Acad. Sci.*, vol. 107, no. 43, pp. 18545–18550, 2010.

[40] F. Gnad, A. Baucom, K. Mukhyala, G. Manning, and Z. Zhang, “Assessment of computational methods for predicting the effects of missense mutations in human cancers,” *BMC Genomics*, vol. 14, no. Suppl 3, p. S7, 2013.

[41] S. Gong, C. L. Worth, T. M. K. Cheng, and T. L. Blundell, “Meet Me Halfway: When Genomics Meets Structural Bioinformatics,” *J Cardiovasc. Transl. Res.*, vol. 4, no. 3, pp. 281–303, Feb. 2011.

[42] J. S. Kaminker, Y. Zhang, A. Waugh, P. M. Haverty, B. Peters, D. Sebisanovic, J. Stinson, W. F. Forrest, J. F. Bazan, S. Seshagiri, and Z. Zhang, “Distinguishing Cancer-Associated Missense Mutations from Common Polymorphisms,” *Cancer Res.*, vol. 67, no. 2, pp. 465–473, Jan. 2007.

[43] P. C. Ng, “SIFT: predicting amino acid changes that affect protein function,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, Jul. 2003.

[44] S. V. Tavtigian, M. S. Greenblatt, F. Lesueur, G. B. Byrnes, and for the IARC Unclassified Genetic Variants Working Group, “In silico analysis of missense substitutions using sequence-alignment based methods,” *Hum. Mutat.*, vol. 29, no. 11, pp. 1327–1336, Nov. 2008.

[45] D. M. Jordan, V. E. Ramensky, and S. R. Sunyaev, “Human allelic variation: perspective from protein function, structure, and evolution,” *Curr. Opin. Struct. Biol.*, vol. 20, no. 3, pp. 342–350, Jun. 2010.

[46] D. R. Walker, J. P. Bond, R. E. Tarone, C. C. Harris, W. Makalowski, M. S. Boguski, and M. S. Greenblatt, “Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features.,” *Oncogene*, vol. 18, no. 1, 1999.

[47] M. P. Miller and S. Kumar, “Understanding human disease mutations through the use of interspecific genetic variation,” *Hum. Mol. Genet.*, vol. 10, no. 21, pp. 2319–2328, Oct. 2001.

[48] M. S. Greenblatt, L. C. Brody, W. D. Foulkes, M. Genuardi, R. M. W. Hofstra, M. Olivier, S. E. Plon, R. H. Sijmons, O. Sinilnikova, A. B. Spurdle, and for the IARC Unclassified Genetic Variants Working Group, “Locus-specific databases and recommendations to strengthen their contribution to the classification of variants in cancer susceptibility genes,” *Hum. Mutat.*, vol. 29, no. 11, pp. 1273–1281, Nov. 2008.

[49] D. E. Goldgar, D. F. Easton, A. M. Deffenbaugh, A. N. Monteiro, S. V. Tavtigian, and F. J. Couch, “Integrated Evaluation of DNA Sequence Variants of Unknown Clinical Significance: Application to< i> BRCA1</i> and< i> BRCA2</i>,” *Am. J. Hum. Genet.*, vol. 75, no. 4, pp. 535–544, 2004.

[50] S. V. Tavtigian, “Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral,” *J. Med. Genet.*, vol. 43, no. 4, pp. 295–305, Sep. 2005.

[51] E. A. Stone and A. Sidow, “Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity,” *Genome Res.*, vol. 15, no. 7, pp. 978–986, 2005.

[52] C. Ferrer-Costa, J. L. Gelpi, L. Zamakola, I. Parraga, X. de la Cruz, and M. Orozco, “PMUT: a web-based tool for the annotation of pathological mutations on proteins,” *Bioinformatics*, vol. 21, no. 14, pp. 3176–3178, May 2005.

[53] V. Ramensky, P. Bork, and S. Sunyaev, “Human non□synonymous SNPs: server and survey,” *Nucleic Acids Res.*, vol. 30, no. 17, pp. 3894–3900, Sep. 2002.

[54] Y. Bromberg, G. Yachdav, and B. Rost, “SNAP predicts effect of mutations on protein function,” *Bioinformatics*, vol. 24, no. 20, pp. 2397–2398, Aug. 2008.

[55] A. Bare\vsic and A. C. Martin, “Compensated pathogenic deviations.”

[56] R. J. Kulathinal, “Compensated Deleterious Mutations in Insect Genomes,” *Science*, vol. 306, no. 5701, pp. 1553–1554, Nov. 2004.

[57] I. F. Tsigelny, K. Vladimir, and W. Linda, “SNP analysis combined with protein structure prediction defines structure-functional relationships in cancer related cytochrome P450 estrogen metabolism,” *Curr. Med. Chem.*, vol. 11, no. 5, pp. 525–538, 2004.

[58] P. Yue, Z. Li, and J. Moult, “Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease,” *J. Mol. Biol.*, vol. 353, no. 2, pp. 459–473, Oct. 2005.

[59] J. Cheng, A. Randall, and P. Baldi, “Prediction of protein stability changes for single-site mutations using support vector machines,” *Proteins Struct. Funct. Bioinforma.*, vol. 62, no. 4, pp. 1125–1132, Dec. 2005.

[60] E. Capriotti, P. Fariselli, and R. Casadio, “I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure,” *Nucleic Acids Res.*, vol. 33, no. Web Server, pp. W306–W310, Jul. 2005.

[61] P. Yue, E. Melamud, and J. Moult, “SNPs3D: candidate gene and SNP selection for association studies,” *BMC Bioinformatics*, vol. 7, no. 1, p. 166, 2006.

[62] R. Karchin, M. Diekhans, L. Kelly, D. J. Thomas, U. Pieper, N. Eswar, D. Haussler, and A. Sali, “LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources,” *Bioinformatics*, vol. 21, no. 12, pp. 2814–2820, Apr. 2005.

[63] E. Capriotti, R. Calabrese, and R. Casadio, “Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information,” *Bioinformatics*, vol. 22, no. 22, pp. 2729–2734, Aug. 2006.

[64] Y. Bromberg and B. Rost, “SNAP: predict effect of non-synonymous polymorphisms on function,” *Nucleic Acids Res.*, vol. 35, no. 11, pp. 3823–3835, May 2007.

[65] B. Li, V. G. Krishnan, M. E. Mort, F. Xin, K. K. Kamati, D. N. Cooper, S. D. Mooney, and P. Radivojac, “Automated inference of molecular mechanisms of disease from amino acid substitutions,” *Bioinformatics*, vol. 25, no. 21, pp. 2744–2750, Sep. 2009.

[66] L. Bao, M. Zhou, and Y. Cui, “nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms,” *Nucleic Acids Res.*, vol. 33, no. Web Server, pp. W480–W482, Jul. 2005.

[67] Z. Wang and J. Moult, “SNPs, protein structure, and disease,” *Hum. Mutat.*, vol. 17, no. 4, pp. 263–270, 2001.

[68] M. Masso and I. I. Vaisman, “AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements,” *Protein Eng. Des. Sel.*, vol. 23, no. 8, pp. 683–687, Jun. 2010.

[69] Y. Dehouck, A. Grosfils, B. Folch, D. Gilis, P. Bogaerts, and M. Rooman, “Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0,” *Bioinformatics*, vol. 25, no. 19, pp. 2537–2543, Aug. 2009.

[70] R. Guerois, J. E. Nielsen, and L. Serrano, “Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations,” *J. Mol. Biol.*, vol. 320, no. 2, pp. 369–387, Jul. 2002.

[71] V. Parthiban, M. M. Gromiha, and D. Schomburg, “CUPSAT: prediction of protein stability upon point mutations,” *Nucleic Acids Res.*, vol. 34, no. Web Server, pp. W239–W242, Jul. 2006.

[72] C. Deutsch and B. Krishnamoorthy, “Four-Body Scoring Function for Mutagenesis,” *Bioinformatics*, vol. 23, no. 22, pp. 3009–3015, Oct. 2007.

[73] H. Zhou and Y. Zhou, “Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction,” *Protein Sci.*, vol. 11, no. 11, pp. 2714–2726, Apr. 2009.

[74] C. T. Saunders and D. Baker, “Evaluation of Structural and Evolutionary Contributions to Deleterious Mutation Prediction,” *J. Mol. Biol.*, vol. 322, no. 4, pp. 891–901, Sep. 2002.

[75] P. Larranaga, “Machine learning in bioinformatics,” *Brief. Bioinform.*, vol. 7, no. 1, pp. 86–112, Feb. 2006.

[76] S. Prompramote, Y. Chen, and Y.-P. P. Chen, “Machine learning in bioinformatics,” in *Bioinformatics technologies*, Springer, 2005, pp. 117–153.

[77] S. From, “How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis,” 2012.

[78] R. J. Dobson, P. B. Munroe, M. J. Caulfield, and M. A. Saqi, “Predicting deleterious nsSNPs: an analysis of sequence and structural attributes,” *BMC Bioinformatics*, vol. 7, no. 1, p. 217, 2006.

[79] M. Michael Gromiha and L.-T. Huang, “Machine learning algorithms for predicting protein folding rates and stability of mutant proteins: Comparison with statistical methods,” *Curr. Protein Pept. Sci.*, vol. 12, no. 6, pp. 490–502, 2011.

[80] P. S. Nair and M. Vihinen, “VariBench: A Benchmark Database for Variations,” *Hum. Mutat.*, vol. 34, no. 1, pp. 42–49, Jan. 2013.

[81] J. Thusberg, A. Olatubosun, and M. Vihinen, “Performance of mutation pathogenicity prediction methods on missense variants,” *Hum. Mutat.*, vol. 32, no. 4, pp. 358–368, Apr. 2011.

[82] W. Lee, Y. Zhang, K. Mukhyala, R. A. Lazarus, and Z. Zhang, “Bi-Directional SIFT Predicts a Subset of Activating Mutations,” *PLoS ONE*, vol. 4, no. 12, p. e8311, Dec. 2009.

[83] G. C. Johnson and J. A. Todd, “Strategies in complex disease mapping,” *Curr. Opin. Genet. Dev.*, vol. 10, no. 3, pp. 330–334, 2000.

[84] A. Krogh, B. Larsson, G. von Heijne, and E. L. . Sonnhammer, “Predicting transmembrane protein topology with a hidden markov model: application to complete genomes,” *J. Mol. Biol.*, vol. 305, no. 3, pp. 567–580, Jan. 2001.

[85] P. C. Ng, J. G. Henikoff, and S. Henikoff, “PHAT: a transmembrane-specific substitution matrix,” *Bioinformatics*, vol. 16, no. 9, pp. 760–766, Sep. 2000.

[86] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne, “Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.,” *Protein Eng.*, vol. 10, no. 1, pp. 1–6, Jan. 1997.

[87] S. R. Sunyaev, F. Eisenhaber, I. V. Rodchenkov, B. Eisenhaber, V. G. Tumanyan, and E. N. Kuznetsov, “PSIC: profile extraction from sequence alignments with position-specific counts of independent observations,” *Protein Eng.*, vol. 12, no. 5, pp. 387–394, May 1999.

[88] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[89] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn, “The Pfam protein families database,” *Nucleic Acids Res.*, vol. 40, no. D1, pp. D290–D301, Nov. 2011.

[90] J. Kyte and R. Doolittle, “A simple method for displaying the hydropathic character of a protein.,” *J. Mol. Biol.*, vol. 157, pp. 105–132, Jan. 1982.

[91] L. Stryer, *Biochemistry*, 4th ed. W.H. Freeman and Co., 1995.

[92] V. Munoz and L. Serrano, “Intrinsic secondary structure propensities of the amino acids, using statistical matrices: Comparison with experimental scales.,” *PROTEINS Struct. Funct. Genet.*, vol. 20, pp. 301–311, 1994.

[93] J.-M. Chandonia, “The ASTRAL Compendium in 2004,” *Nucleic Acids Res.*, vol. 32, no. 90001, p. 189D–192, Jan. 2004.

[94] J. Bowie, R. Luthy, and D. Eisenberg, “A METHOD TO IDENTIFY PROTEIN SEQUENCES THAT FOLD INTO A KNOWN THREE-DIMENSIONAL STRUCTURE,” *Science*, pp. 164–170, Jul. 1991.

[95] D. Frishman and P. Argos, “Knowledge-based protein secondary structure assignment.,” *PROTEINS Struct. Funct. Genet.*, vol. 23, pp. 566–579, 1995.

[96] R. Calabrese, E. Capriotti, P. Fariselli, P. L. Martelli, and R. Casadio, “Functional annotations improve the predictive score of human disease-related mutations in proteins,” *Hum. Mutat.*, vol. 30, no. 8, pp. 1237–1244, Aug. 2009.

[97] P. D. Thomas and A. Kejariwal, “Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 43, pp. 15398–15403, 2004.

[98] E. Capriotti, P. Fariselli, and R. Casadio, “A neural-network-based method for predicting protein stability changes upon single point mutations,” *Bioinformatics*, vol. 20, no. Suppl 1, pp. i63–i68, Jul. 2004.

[99] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[100] B. Rost, “PHD: predicting one-dimensional protein structure by profile-based neural networks.,” *Methos Enzymol.*, vol. 266, pp. 525–39, 1996.

[101] B. Rost and C. Sander, “Prediction of Protein Secondary Structure at Better than 70% Accuracy,” *J. Mol. Biol.*, vol. 232, pp. 584–599, 1993.

[102] A. Schlessinger, G. Yachdav, and B. Rost, “PROFbval: predict flexible and rigid residues in proteins,” *Bioinformatics*, vol. 22, no. 7, pp. 891–893, Apr. 2006.

[103] C. Ferrer-Costa, M. Orozco, and X. de la Cruz, “Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties,” *J. Mol. Biol.*, vol. 315, no. 4, pp. 771–786, Jan. 2002.

[104] C. Ferrer-Costa, M. Orozco, and X. de la Cruz, “Sequence-based prediction of pathological mutations,” *Proteins Struct. Funct. Bioinforma.*, vol. 57, no. 4, pp. 811–819, Dec. 2004.

[105] P. Yue and J. Moult, “Identification and Analysis of Deleterious Human SNPs,” *J. Mol. Biol.*, vol. 356, no. 5, pp. 1263–1274, Mar. 2006.

[106] A. Uzun, C. M. Leslin, A. Abyzov, and V. Ilyin, “Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways,” *Nucleic Acids Res.*, vol. 35, no. Web Server, pp. W384–W392, May 2007.

[107] D. Gilis and M. Rooman, “PoPMuSiC, an algorithm for predicting protein mutant stability changes. Application to prion proteins,” *Protein Eng.*, vol. 13, no. 12, pp. 849–856, Dec. 2000.

[108] K. A. Bava, “ProTherm, version 4.0: thermodynamic database for proteins and mutants,” *Nucleic Acids Res.*, vol. 32, no. 90001, p. 120D–121, Jan. 2004.

[109] H.-Y. Yuan, J.-J. Chiou, W.-H. Tseng, C.-H. Liu, C.-K. Liu, Y.-J. Lin, H.-H. Wang, A. Yao, Y.-T. Chen, and C.-N. Hsu, “FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization,” *Nucleic Acids Res.*, vol. 34, no. Web Server, pp. W635–W641, Jul. 2006.

[110] S. T. Sherry, M. Ward, and K. Sirotkin, “dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation,” *Genome Res.*, vol. 9, no. 8, pp. 677–679, Aug. 1999.

[111] T. Heinemeyer, E. Wingender, I. Reuter, H. Hermjakob, A. E. Kel, O. V. Kel, E. V. Ignatieva, E. A. Ananko, O. A. Podkolodnaya, and F. A. Kolpakov, “Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL,” *Nucleic Acids Res.*, vol. 26, no. 1, pp. 362–367, 1998.

[112] L. Cartegni, “ESEfinder: a web resource to identify exonic splicing enhancers,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3568–3571, Jul. 2003.

[113] W. G. Fairbrother, G. W. Yeo, R. Yeh, P. Goldstein, M. Mawson, P. A. Sharp, and C. B. Burge, “RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons,” *Nucleic Acids Res.*, vol. 32, no. Web Server, pp. W187–W190, Jul. 2004.

[114] Z. Wang, M. E. Rolish, G. Yeo, V. Tung, M. Mawson, and C. B. Burge, “Systematic identification and analysis of exonic splicing silencers,” *Cell*, vol. 119, no. 6, pp. 831–845, 2004.

[115] D. Karolchik, “The UCSC Genome Browser Database,” *Nucleic Acids Res.*, vol. 31, no. 1, pp. 51–54, Jan. 2003.

[116] S. McGinnis and T. L. Madden, “BLAST: at the core of a powerful and diverse set of sequence analysis tools,” *Nucleic Acids Res.*, vol. 32, no. Web Server, pp. W20–W25, Jul. 2004.

[117] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch’ang, W. Huang, B. Liu, and Y. Shen, “The international HapMap project,” *Nature*, vol. 426, no. 6968, pp. 789–796, 2003.

[118] T. M. K. Cheng, Y.-E. Lu, M. Vendruscolo, P. Lio’, and T. L. Blundell, “Prediction by Graph Theoretic Measures of Structural Effects in Proteins Arising from Non-Synonymous Single Nucleotide Polymorphisms,” *PLoS Comput. Biol.*, vol. 4, no. 7, p. e1000135, Jul. 2008.

[119] P. D. Thomas, “PANTHER: A Library of Protein Families and Subfamilies Indexed by Function,” *Genome Res.*, vol. 13, no. 9, pp. 2129–2141, Sep. 2003.

[120] H. Mi, Q. Dong, A. Muruganujan, P. Gaudet, S. Lewis, and P. D. Thomas, “PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium,” *Nucleic Acids Res.*, vol. 38, no. Database, pp. D204–D210, Dec. 2009.

[121] N. O. Stitziel, “topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association,” *Nucleic Acids Res.*, vol. 32, no. 90001, p. 520D–522, Jan. 2004.

[122] N. O. Stitziel, Y. Y. Tseng, D. Pervouchine, D. Goddeau, S. Kasif, and J. Liang, “Structural Location of Disease-associated Single-nucleotide Polymorphisms,” *J. Mol. Biol.*, vol. 327, no. 5, pp. 1021–1030, Apr. 2003.

[123] A. González-Pérez and N. López-Bigas, “Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel,” *Am. J. Hum. Genet.*, vol. 88, no. 4, pp. 440–449, Apr. 2011.

[124] M. C. Lopes, C. Joyce, G. R. S. Ritchie, S. L. John, F. Cunningham, J. Asimit, and E. Zeggini, “A Combined Functional Annotation Score for Non-Synonymous Variants,” *Hum. Hered.*, vol. 73, no. 1, pp. 47–51, 2012.

[125] X. Liu, X. Jian, and E. Boerwinkle, “dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations,” *Hum. Mutat.*, vol. 34, no. 9, pp. E2393–E2402, Sep. 2013.

[126] X. Liu, X. Jian, and E. Boerwinkle, “dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions,” *Hum. Mutat.*, vol. 32, no. 8, pp. 894–899, Aug. 2011.

[127] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, “Detection of nonneutral substitution rates on mammalian phylogenies,” *Genome Res.*, vol. 20, no. 1, pp. 110–121, Oct. 2009.

[128] K. Wang, M. Li, and H. Hakonarson, “ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data,” *Nucleic Acids Res.*, vol. 38, no. 16, pp. e164–e164, Jul. 2010.

[129] P. H. Lee and H. Shatkay, “An integrative scoring system for ranking SNPs by their potential deleterious effects,” *Bioinformatics*, vol. 25, no. 8, pp. 1048–1055, Feb. 2009.

[130] P. H. Lee and H. Shatkay, “F-SNP: computationally predicted functional SNPs for disease association studies,” *Nucleic Acids Res.*, vol. 36, no. Database, pp. D820–D824, Dec. 2007.

[131] B. Reva, Y. Antipin, and C. Sander, “Determinants of protein function revealed by combinatorial entropy optimization,” *Genome Biol*, vol. 8, no. 11, p. R232, 2007.

[132] J. S. Kaminker, Y. Zhang, C. Watanabe, and Z. Zhang, “CanPredict: a computational tool for predicting cancer-associated missense mutations,” *Nucleic Acids Res.*, vol. 35, no. Web Server, pp. W595–W598, May 2007.

[133] R. J. Clifford, M. N. Edmonson, C. Nguyen, and K. H. Buetow, “Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms,” *Bioinformatics*, vol. 20, no. 7, pp. 1006–1014, Jan. 2004.

[134] D. W. Parsons, S. Jones, X. Zhang, J. C.-H. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, I.-M. Siu, G. L. Gallia, A. Olivi, R. McLendon, B. A. Rasheed, S. Keir, T. Nikolskaya, Y. Nikolsky, D. A. Busam, H. Tekleab, L. A. Diaz, J. Hartigan, D. R. Smith, R. L. Strausberg, S. K. N. Marie, S. M. O. Shinjo, H. Yan, G. J. Riggins, D. D. Bigner, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu, and K. W. Kinzler, “An Integrated Genomic Analysis of Human Glioblastoma Multiforme,” *Science*, vol. 321, no. 5897, pp. 1807–1812, Sep. 2008.

[135] W. C. Wong, D. Kim, H. Carter, M. Diekhans, M. C. Ryan, and R. Karchin, “CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer,” *Bioinformatics*, vol. 27, no. 15, pp. 2147–2148, Jun. 2011.

[136] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham, "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor," *Bioinformatics*, vol. 26, no. 16, pp. 2069–2070, Jun. 2010.

[137] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, "Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++," *PLoS Comput. Biol.*, vol. 6, no. 12, p. e1001025, Dec. 2010.

[138] S. Chun and J. C. Fay, "Identification of deleterious mutations within three human genomes," *Genome Res.*, vol. 19, no. 9, pp. 1553–1561, Jul. 2009.

[139] P. Yue, W. F. Forrest, J. S. Kaminker, S. Lohr, Z. Zhang, and G. Cavet, "Inferring the functional effects of mutation through clusters of mutations in homologous proteins," *Hum. Mutat.*, vol. 31, no. 3, pp. 264–271, Mar. 2010.

[140] S. R. Eddy, "Profile hidden Markov models.," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, Jan. 1998.

[141] A. Gonzalez-Perez, J. Deu-Pons, and N. Lopez-Bigas, "Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation," *Genome Med*, vol. 4, no. 11, pp. 89–89, 2012.

[142] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson, and L. Ding, "MuSiC: Identifying mutational significance in cancer genomes," *Genome Res.*, vol. 22, no. 8, pp. 1589–1598, Jul. 2012.

[143] M. C. Wendl, J. W. Wallis, L. Lin, C. Kandoth, E. R. Mardis, R. K. Wilson, and L. Ding, "PathScan: a tool for discerning mutational significance in groups of putative cancer genes," *Bioinformatics*, vol. 27, no. 12, pp. 1595–1602, Apr. 2011.

[144] M. Ohanian, R. Otway, and D. Fatkin, "Heuristic Methods for Finding Pathogenic Variants in Gene Coding Sequences," *J. Am. Heart Assoc.*, vol. 1, no. 5, pp. e002642–e002642, Sep. 2012.

[145] J. Thusberg and M. Vihinen, "Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods," *Hum. Mutat.*, vol. 30, no. 5, pp. 703–714, May 2009.

[146] S. Castellana and T. Mazza, "Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools," *Brief. Bioinform.*, vol. 14, no. 4, pp. 448–459, Mar. 2013.

[147] R. Karchin, “Next generation tools for the annotation of human SNPs,” *Brief. Bioinform.*, vol. 10, no. 1, pp. 35–52, Oct. 2008.

[148] V. G. Krishnan and P. C. Ng, “Predicting cancer drivers: are we there yet?,” *Genome Med.*, vol. 4, no. 11, pp. 1–3, 2012.

[149] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, and M. J. Bamshad, “Exome sequencing identifies the cause of a mendelian disorder,” *Nat. Genet.*, vol. 42, no. 1, pp. 30–35, Nov. 2009.

[150] D. E. Goldgar, D. F. Easton, G. B. Byrnes, A. B. Spurdle, E. S. Iversen, M. S. Greenblatt, and for the IARC Unclassified Genetic Variants Working Group, “Genetic evidence and integration of various data sources for classifying uncertain variants into a single model,” *Hum. Mutat.*, vol. 29, no. 11, pp. 1265–1272, Nov. 2008.

[151] S. Hicks, D. A. Wheeler, S. E. Plon, and M. Kimmel, “Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed,” *Hum. Mutat.*, vol. 32, no. 6, pp. 661–668, Jun. 2011.

[152] P. A. Chan, S. Duraisamy, P. J. Miller, J. A. Newell, C. McBride, J. P. Bond, T. Raevaara, S. Ollila, M. Nyström, A. J. Grimm, J. Christodoulou, W. S. Oetting, and M. S. Greenblatt, “Interpreting missense variants: comparing computational methods in human disease genes *CDKN2A* , *MLH1* , *MSH2* , *MECP2* , and tyrosinase (*TYR*),” *Hum. Mutat.*, vol. 28, no. 7, pp. 683–693, Jul. 2007.

[153] A. C. R. Martin, A. M. Facchiano, A. L. Cuff, T. Hernandez-Boussard, M. Olivier, P. Hainaut, and J. M. Thornton, “Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein,” *Hum. Mutat.*, vol. 19, no. 2, pp. 149–164, Feb. 2002.

[154] Z. Dosztanyi, C. Magyar, G. Tusnady, and I. Simon, “SCide: identification of stabilization centers in proteins,” *Bioinformatics*, vol. 19, no. 7, pp. 899–900, May 2003.

[155] Z. Dosztányi, A. Fiser, and I. Simon, “Stabilization centers in proteins: identification, characterization and predictions,” *J. Mol. Biol.*, vol. 272, no. 4, pp. 597–612, 1997.

[156] C. Magyar, M. M. Gromiha, G. Pujadas, G. E. Tusnady, and I. Simon, “SRide: a server for identifying stabilizing residues in proteins,” *Nucleic Acids Res.*, vol. 33, no. Web Server, pp. W303–W305, Jul. 2005.

[157] C. M. Topham, N. Srinivasan, and T. L. Blundell, “Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables.,” *Protein Eng.*, vol. 10, no. 1, pp. 7–21, 1997.

[158] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, “The FoldX web server: an online force field,” *Nucleic Acids Res.*, vol. 33, no. Web Server, pp. W382–W388, Jul. 2005.

[159] B. Krishnamoorthy and A. Tropsha, “Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations,” *Bioinformatics*, vol. 19, no. 12, pp. 1540–1548, Aug. 2003.

[160] R. I. Dima, G. Settanni, C. Micheletti, J. R. Banavar, and A. Maritan, “Extraction of interaction potentials between amino acids from native protein structures,” *J. Chem. Phys.*, vol. 112, no. 20, p. 9151, 2000.

[161] C. Micheletti, F. Seno, J. R. Banavar, and A. Maritan, “Learning effective amino acid interactions through iterative stochastic techniques,” *Proteins Struct. Funct. Bioinforma.*, vol. 42, no. 3, pp. 422–431, 2001.

[162] G. A. Lazar and T. M. Handel, “Hydrophobic Core Packing and Protein Design,” *Curr. Opin. Chem. Biol.*, vol. 2, no. 6, pp. 675–679, 1998.

[163] J. Chen and W. E. Stites, “Higher-Order Packing Interactions in Triple and Quadruple Mutants of Staphylococcal Nuclease †,” *Biochemistry (Mosc.)*, vol. 40, no. 46, pp. 14012–14019, Nov. 2001.

[164] J. Khatun, S. D. Khare, and N. V. Dokholyan, “Can contact potentials reliably predict stability of proteins?,” *J. Mol. Biol.*, vol. 336, no. 5, pp. 1223–1238, 2004.

[165] P. Sundaramurthy, K. Shameer, R. Sreenivasan, S. Gakkhar, and R. Sowdhamini, “HORI: a web server to compute Higher Order Residue Interactions in protein structures,” *BMC Bioinformatics*, vol. 11, no. Suppl 1, p. S24, 2010.

[166] A. A. Jalan and J. D. Hartgerink, “Pairwise interactions in collagen and the design of heterotrimeric helices,” *Curr. Opin. Chem. Biol.*, vol. 17, no. 6, pp. 960–967, Dec. 2013.

[167] A. Tropsha, C. Carter, S. Cammer, and I. I. Vaisman, “Simplicial neighborhood analysis of protein packing (SNAPP): a computational geometry approach to studying proteins.,” *Methods Enzymol.*, vol. 374, pp. 509–544.

[168] C. W. Carter, B. C. LeFebvre, S. A. Cammer, A. Tropsha, and M. H. Edgell, “Four-body potentials reveal protein-specific correlations to stability changes caused

by hydrophobic core mutations," *J. Mol. Biol.*, vol. 311, no. 4, pp. 625–638, Aug. 2001.

[169] F. M. Richards, "The interpretation of protein structures: Total volume, group volume distributions and packing density," *J. Mol. Biol.*, vol. 82, no. 1, pp. 1–14, Jan. 1974.

[170] A. Poupon, "Voronoi and Voronoi-related tessellations in studies of protein structure and interaction," *Curr. Opin. Struct. Biol.*, vol. 14, no. 2, pp. 233–241, Apr. 2004.

[171] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Trans. Math. Softw. TOMS*, vol. 22, no. 4, pp. 469–483, 1996.

[172] E. Jeremy, O. Christophe, and G. de B. Alexandre, "A novel evaluation of residue and protein volumes by means of Laguerre tessellation.," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 947–960, Apr. 2010.

[173] A. Soyer, J. Chomilier, J.-P. Momon, R. Jullien, and J.-F. Sadoc, "Voronoi tessellation reveals the condensed matter character of folded proteins.," *Phys. Rev. Lett.*, vol. 85, no. 3532, Oct. 2000.

[174] S. Chakravarty, "A Procedure for Detection and Quantitation of Cavity Volumes in Proteins. APPLICATION TO MEASURE THE STRENGTH OF THE HYDROPHOBIC DRIVING FORCE IN PROTEIN FOLDING," *J. Biol. Chem.*, vol. 277, no. 35, pp. 31345–31353, Jun. 2002.

[175] T. J. Taylor and I. I. Vaisman, "Discrimination of thermophilic and mesophilic proteins," *BMC Struct. Biol.*, vol. 10, no. Suppl 1, p. S5, 2010.

[176] N. Lindow, D. Baum, A.-N. Bondar, and H.-C. Hege, "Exploring cavity dynamics in biomolecular systems," *BMC Bioinformatics*, vol. 14, no. Suppl 19, p. S5, 2013.

[177] Z. S. Apte and W. F. Marshall, "Statistical method for comparing the level of intracellular organization between cells," *Proc. Natl. Acad. Sci.*, vol. 110, no. 11, pp. E1006–E1015, Nov. 2012.

[178] R. K. Singh, A. Tropsha, and I. I. Vaisman, "Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues," *J. Comput. Biol.*, vol. 3, no. 2, pp. 213–221, 1996.

[179] A. Tropsha, R. . Singh, I. I. Vaisman, and W. Zheng, "STATISTICAL GEOMETRY ANALYSIS OF PROTEINS: IMPLICATIONS FOR INVERTED STRUCTURE PREDICTION," *Pac. Symp. Biocomput.*, pp. 614–623, 1996.

[180] W. Zheng, S. cho, I. I. Vaisman, and A. Tropsha, “A new approach to protein fold recognition based on Delaunay tessellation of protein structure.,” *Pac. Symp. Biocomput.*, pp. 486–497, 1997.

[181] P. J. Munson and R. K. Singh, “Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment,” *Protein Sci.*, vol. 6, no. 7, pp. 1467–1481, 1997.

[182] G. Weberndorfer, I. . Hofacker, and P. . Stadler, “An Efficient Potential For Protein Sequence Design,” *Proc Ger. Conf. Bioinforma.*, pp. 107–112, 1999.

[183] H. H. Gan, A. Tropsha, and T. Schlick, “Lattice protein folding with two and four-body statistical potentials,” *Proteins Struct. Funct. Bioinforma.*, vol. 43, no. 2, pp. 161–174, 2001.

[184] M. Masso and I. I. Vaisman, “Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach,” *Biochem. Biophys. Res. Commun.*, vol. 305, no. 2, pp. 322–326, May 2003.

[185] D. Bostick and I. I. Vaisman, “A new topological method to measure protein structure similarity,” *Biochem. Biophys. Res. Commun.*, vol. 304, no. 2, pp. 320–325, May 2003.

[186] D. L. Bostick, M. Shen, and I. I. Vaisman, “A simple topological representation of protein structure: Implications for new, fast, and robust structural classification,” *Proteins Struct. Funct. Bioinforma.*, vol. 56, no. 3, pp. 487–501, May 2004.

[187] M. Barenboim, D. C. Jamison, and I. I. Vaisman, “Statistical geometry approach to the study of functional effects of human nonsynonymous SNPs,” *Hum. Mutat.*, vol. 26, no. 5, pp. 471–476, Nov. 2005.

[188] I. I. Vaisman, A. Tropsha, and W. Zheng, “Compositional preferences in quadruplets of nearest neighbor residues in protein structures: statistical geometry analysis,” in *Intelligence and Systems, 1998. Proceedings., IEEE International Joint Symposia on*, 1998, pp. 163–168.

[189] I. Hedenfalk, “Gene-expression profiles in hereditary breast cancer.,” *N. Engl. J. Meidcine*, vol. 344, pp. 539–548, Feb. 2001.

[190] E. Mathe, M. Olivier, S. Kato, C. Ishioka, I. Vaisman, and P. Hainaut, “Predicting the transactivation activity of p53 missense mutants using a four-body potential score derived from Delaunay tessellations,” *Hum. Mutat.*, vol. 27, no. 2, pp. 163–172, Feb. 2006.

[191] T. Taylor, M. Rivera, G. Wilson, and I. I. Vaisman, “New method for protein secondary structure assignment based on a simple topological descriptor,” *Proteins Struct. Funct. Bioinforma.*, vol. 60, no. 3, pp. 513–524, May 2005.

[192] M. A. Care, C. J. Needham, A. J. Bulpitt, and D. R. Westhead, “Deleterious SNP prediction: be mindful of your training data!,” *Bioinformatics*, vol. 23, no. 6, pp. 664–672, Jan. 2007.

[193] M. Masso and I. I. Vaisman, “Accurate prediction of enzyme mutant activity based on a multibody statistical potential,” *Bioinformatics*, vol. 23, no. 23, pp. 3155–3161, Oct. 2007.

[194] M. Masso and I. I. Vaisman, “Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis,” *Bioinformatics*, vol. 24, no. 18, pp. 2002–2009, Jul. 2008.

[195] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[196] J. R. Quinlan, “Induction of decision trees,” *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[197] T. Fawcett, “ROC graphs: Notes and practical considerations for researchers,” *Mach. Learn.*, vol. 31, pp. 1–38, 2004.

[198] S. M. Dhanasekaran, T. R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan, “Delineation of prognostic biomarkers in prostate cancer.,” *Nature*, vol. 412, pp. 822–826, Aug. 2001.

[199] American Cancer Society, “Cancer Facts & Figures2014,” *Am. Cancer Soc.*, 2014.

[200] K. D. Sørensen and T. F. Ørntoft, “Discovery of Prostate Cancer Biomarkers by Microarray Gene Expression Profiling,” *Expert Rev. Mol. Diagn.*, p. 2010.

[201] M. Piccart, C. Lohrisch, A. Di Leo, and D. Larsimont, “The predictive value of HER2 in breast cancer,” *Oncology*, vol. 61, no. Suppl. 2, pp. 73–82, 2001.

[202] M. J. Duffy, “Predictive Markers in Breast and Other Cancers: A Review,” *Clin. Chem.*, vol. 51, no. 3, pp. 494–503, Mar. 2005.

[203] S. E. Baldus, K. Engelmann, and F.-G. Hanisch, “MUC1 and the MUCs: A Family of Human Mucins with Impact in Cancer Biology,” *Crit. Rev. Clin. Lab. Sci.*, vol. 41, no. 2, pp. 189–231, Jan. 2004.

[204] G. V. Glinsky, A. B. Glinskii, A. J. Stephenson, R. M. Hoffman, and W. L. Gerald, “Gene expression profiling predicts clinical outcome of prostate cancer,” *J. Clin. Invest.*, vol. 113, no. 6, pp. 913–923, Mar. 2004.

[205] A. J. Stephenson, A. Smith, M. W. Kattan, J. Satagopan, V. E. Reuter, P. T. Scardino, and W. L. Gerald, “Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy,” *Cancer*, vol. 104, no. 2, pp. 290–298, 2005.

[206] L. J. van ’t Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, “Gene expression profiling predicts clinical outcome of breast cancer,” *Nature*, vol. 415, pp. 530–536, Jan. 2002.

[207] A. Y. Salmon, M. Salmon-Divon, T. Zahavi, Y. Barash, R. S. Levy-Drummer, J. Jacob-Hirsch, and T. Peretz, “Determination of Molecular Markers for BRCA1 and BRCA2 Heterozygosity Using Gene Expression Profiling,” *Cancer Prev. Res. (Phila. Pa.)*, vol. 6, no. 2, pp. 82–90, Feb. 2013.

[208] R. Xu, X. Cai, and D. C. Wunsch, “Gene expression data for DLBCL cancer survival prediction with a combination of machine learning technologies,” in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, 2006, pp. 894–897.

[209] M. Burton, M. Thomassen, Q. Tan, and T. A. Kruse, “Gene Expression Profiles for Predicting Metastasis in Breast Cancer: A Cross-Study Comparison of Classification Methods,” *Sci. World J.*, vol. 2012, pp. 1–11, 2012.

[210] A. D. Weston and L. Hood, “Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine,” *J. Proteome Res.*, vol. 3, no. 2, pp. 179–196, Mar. 2004.

[211] S. Y. Kim, J. W. Lee, and I. S. Sohn, “Comparison of various statistical methods for identifying differential gene expression in replicated microarray data,” *Stat. Methods Med. Res.*, vol. 15, no. 1, pp. 3–20, Feb. 2006.

[212] V. G. Tusher, R. Tibshirani, and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 9, pp. 5116–5121, 2001.

[213] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, and J. Gentry, “Bioconductor: open software development for computational biology and bioinformatics,” *Genome Biol.*, vol. 5, no. 10, p. R80, 2004.

[214] O. J. Halvorsen, A. M. Øyan, T. H. Bø, S. Olsen, K. Rostad, S. A. Haukaas, A. M. Bakke, B. Marzolf, K. Dimitrov, L. Stordrange, B. Lin, I. Jonassen, L. Hood, L. A. Akslen, and K.-H. Kalland, “Gene expression profiles in prostate cancer: Association with patient subgroups and tumour differentiation,” *Int. J. Oncol.*, vol. 26, no. 2, 2005.

[215] L. True, I. Coleman, S. Hawley, C.-Y. Huang, D. Gifford, R. Coleman, T. M. Beer, E. Gelmann, M. Datta, and E. Mostaghel, “A molecular correlate to the Gleason grading system for prostate adenocarcinoma,” *Proc. Natl. Acad. Sci.*, vol. 103, no. 29, pp. 10991–10996, 2006.

[216] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D’Amico, and J. P. Richie, “Gene expression correlates of clinical prostate cancer behavior,” *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[217] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, and S. S. Jeffrey, “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 19, pp. 10869–10874, 2001.

[218] P. J. Coopman, M. T. Do, M. Barth, E. T. Bowden, A. J. Hayes, E. Basyuk, J. K. Blancato, P. R. Vezza, S. W. McLeskey, and P. H. Mangeat, “Molecular portraits of human breast tumours,” *Nature*, vol. 406, no. 6797, pp. 742–747, 2000.

[219] M. J. Van De Vijver, Y. D. He, L. J. van’t Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, and M. J. Marton, “A gene-expression signature as a predictor of survival in breast cancer,” *N. Engl. J. Med.*, vol. 347, no. 25, pp. 1999–2009, 2002.

[220] D. Lindgren, G. Sjödahl, M. Lauss, J. Staaf, G. Chebil, K. Lövgren, S. Gudjonsson, F. Liedberg, O. Patschan, W. Månsson, M. Fernö, and M. Höglund, “Integrated Genomic and Gene Expression Profiling Identifies Two Major Genomic Circuits in Urothelial Carcinoma,” *PLoS ONE*, vol. 7, no. 6, p. e38863, Jun. 2012.

[221] M. D’Antonio and F. D. Ciccarelli, “Integrated analysis of recurrent properties of cancer genes to identify novel drivers,” *Genome Biol.*, vol. 14, no. 5, p. R52, 2013.

[222] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, C. Caldas, S. Aparicio, C. Curtis†, S. P. Shah, C. Caldas, S. Aparicio, J. D. Brenton, I. Ellis, D. Huntsman, S. Pinder, A. Purushotham, L. Murphy, C. Caldas, S. Aparicio, C. Caldas, H. Bardwell, S.-F. Chin, C. Curtis, Z. Ding, S. Gräf, L. Jones, B. Liu, A. G. Lynch, I. Papatheodorou, S. J. Sammut, G. Wishart, S. Aparicio, S. Chia, K. Gelmon, D. Huntsman, S. McKinney, C. Speers, G. Turashvili, P. Watson, I. Ellis, R. Blamey, A. Green, D. Macmillan, E. Rakha, A. Purushotham, C. Gillett, A. Grigoriadis, S. Pinder, E. di Rinaldis, A. Tutt, L. Murphy, M. Parisien, S. Troup, C. Caldas, S.-F. Chin, D. Chan, C. Fielding, A.-T. Maia, S. McGuire, M. Osborne, S. M. Sayalero, I. Spiteri, J. Hadfield, S. Aparicio, G. Turashvili, L. Bell, K. Chow, N. Gale, D. Huntsman, M. Kovalik, Y. Ng, L. Prentice, C. Caldas, S. Tavaré, C. Curtis, M. J. Dunning, S. Gräf, A. G. Lynch, O. M. Rueda, R. Russell, S. Samarajiwa, D. Speed, F. Markowetz, Y. Yuan, J. D. Brenton, S. Aparicio, S. P. Shah, A. Bashashati, G. Ha, G. Haffari, S. McKinney, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio, “The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups,” *Nature*, Apr. 2012.

[223] Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, and R. Shen, “Pattern discovery and cancer gene identification in integrated cancer genomic data,” *Proc. Natl. Acad. Sci.*, vol. 110, no. 11, pp. 4245–4250, Mar. 2013.

[224] J. Zhang, S. Zhang, Y. Wang, and X.-S. Zhang, “Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data,” *BMC Syst. Biol.*, vol. 7, no. Suppl 2, p. S4, 2013.

[225] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan, “ONCOMINE: a cancer microarray database and integrated data-mining platform,” *Neoplasia N. Y. NY*, vol. 6, no. 1, p. 1, 2004.

[226] A. Janevski, S. Kamalakaran, N. Banerjee, V. Varadan, and N. Dimitrova, “PAPAyA: a platform for breast cancer biomarker signature discovery, evaluation and assessment,” *BMC Bioinformatics*, vol. 10, no. Suppl 9, p. S7, 2009.

[227] A. Elfilali, “ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis,” *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D613–D616, Jan. 2006.

[228] M. Goldman, B. Craft, T. Swatloski, K. Ellrott, M. Cline, M. Diekhans, S. Ma, C. Wilks, J. Stuart, D. Haussler, and J. Zhu, "The UCSC Cancer Genomics Browser: update 2013," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D949–D954, Jan. 2013.

[229] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and J. P. Mesirov, "GenePattern 2.0," *Nat. Genet.*, vol. 38, no. 5, pp. 500–501, 2006.

[230] A. Chen, Y. Tsau, and Y. Wang, "A novel multi-task support vector sample learning technique to predict classification of cancer.," *New Trends Inf. Sci. Serv. Sci. NISS*, pp. 196–200, 2010.

[231] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Inform.*, vol. 2, p. 59, 2006.

[232] J. Pittman, E. Huang, H. Dressman, C.-F. Horng, S. H. Cheng, M.-H. Tsou, C.-M. Chen, A. Bild, E. S. Iversen, and A. T. Huang, "Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 22, pp. 8431–8436, 2004.

[233] H. Hijazi and C. Chan, "A Classification Framework Applied to Cancer Gene Expression Profiles," *J. Healthc. Eng.*, vol. 4, no. 2, pp. 255–284, Jun. 2013.

[234] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[235] V. Vapnik, *Estimation of Dependences Based on Empirical Data*, 1st ed. 1982.

[236] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets--update," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D991–D995, Jan. 2013.

[237] B. S. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, B. S. Carver, V. K. Arora, P. Kaushik, E. Cerami, B. Reva, Y. Antipin, N. Mitsiades, T. Landers, I. Dolgalev, J. E. Major, M. Wilson, N. D. Socci, A. E. Lash, A. Heguy, J. A. Eastham, H. I. Scher, V. E. Reuter, P. T. Scardino, C. Sander, C. L. Sawyers, and W. L. Gerald, "Integrative Genomic Profiling of Human Prostate Cancer," *Cancer Cell*, vol. 18, no. 1, pp. 11–22, Jul. 2010.

[238] H. Kim and M. Bredel, "Feature selection and survival modeling in The Cancer Genome Atlas," *Int. J. Nanomedicine*, vol. 8, no. Suppl 1, pp. 57–62, 2013.

[239] A. Kasprzyk, “BioMart: driving a paradigm change in biological data management,” *Database*, vol. 2011, no. 0, pp. bar049–bar049, Nov. 2011.

Biography

KanakaDurga Addepalli was born in India but is a US citizen now. She received her Bachelor of Science from University of Mumbai, in 1997. She received her Masters in Biochemistry and Biotechnology also from University of Mumbai in 1999. She did a summer project at the prestigious institute BARC, from Oct 1998 to Jan 1999. She was awarded the Advanced Diploma in Bioinformatics from University of Pune in the year 2000. She was employed as a Bioinformatics Officer at ForScience Inc and then as Bioinformatics Executive at Dr.Reddy's Labs. She moved to USA in December 2001. She joined GMU in 2003 and worked as Graduate Research Assistant and a Summer Intern in 2003. She was employed with SAIC in 2005, contracting for NCI/CBIIT and CTEP (sub-contract via CTIS) as Analyst, for 5+ years. She was later employed with SAIC-Frederick, contracting for NCI/CCR at Bethesda for 1 year as Bioinformatics Analyst in the NGS initiative.