Carlos Henrique Miranda Rodrigues

Identification and Understanding of Kinase Activating Missense Mutations

Belo Horizonte

July 2017

Carlos Henrique Miranda Rodrigues

Identification and Understanding of Kinase Activating Missense Mutations

Dissertation presented to Programa de Pós-Graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais as requirement for obtaining the title of Master in Bioinformatics.

Universidade Federal de Minas Gerais - UFMG

Instituto de Ciências Biológicas

Programa de Pós-Graduação em Bioinformática

Supervisor: Douglas Eduardo Pires Valente

Co-supervisor: David Benjamin Ascher

Belo Horizonte

July 2017

Carlos Henrique Miranda Rodrigues
Identification and Understanding of Kinase Activating Missense Mutations / Carlos Henrique Miranda Rodrigues. – Belo Horizonte, July 2017-
102 p. : il. (algumas color.) ; 30 cm.

Supervisor: Douglas Eduardo Pires Valente

Dissertation (Master) – Universidade Federal de Minas Gerais - UFMG
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática, July 2017.

1. Bioinformatics. 2. Activating Mutations. 3. Protein Kinase. 4. Machine Learning I. Pires, Douglas Eduardo Valente. II. UFMG. III. ICB. IV. Title

This work is dedicated to Carlinhos, Jane, Vinicius, Ana and Lála,
I love you.

Acknowledgements

I would first like to thank my supervisor Dr. Douglas Pires from Fiocruz Minas - René Rachou Institute. Prof. Pires desk was always open whenever I ran into a trouble spot or had a question about my research or writing. I would also like to thank my co-supervisor Dr David Ascher of School of Biomedical Sciences at the University of Melbourne. Even though he was mostly on the other side of the globe he always provide invaluable insights to this work. Both of them, consistently allowed this thesis to be my own work, but steered me in the right direction whenever they thought I need it. Thank you Drs!

Secondly, I would like to thank all the professors and staff of the MSc degree Bioinformatics Programme at the Universidade Federal de Minas Gerais, especially Prof. Dr. Vasco Azevedo for always looking for what it is best for his students and the Programme, and also Sheila Santana for her tireless hard work.

I would also like to acknowledge my colleagues of the MSc Programme, Dhiego, Felipe and Edson, as well as my co-workers at the room 108, Leilaine, Grace, Fred, João, Henrique, Amanda and Paul. Also my friends in CEBIO at Fiocruz Minas, Fausto, Laura, Larissa, Juliana, Wagner, Fabiano, Gabriel. Thank you for contributions and insights on this work during the last years.

I must also express my profound gratitude to my parents, Jane and Carlos, and to my girlfriend, Yara, for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Finally, to Fundação de Amparo a Pesquisa de Minas Gerais (FAPEMIG) and Conselho Nacional de Desenvolvimento Científico e Tecnológico for providing the financial support to the development of this work.

“The real voyage of discovery consists not in seeking new landscapes, but in having new eyes.”

(Marcel Proust)

Abstract

Protein phosphorylation and dephosphorylation play vital roles in a variety of cellular processes, and the balance between them must be closely regulated. Disturbances in the harmonic relationship between protein phosphorylation and dephosphorylation, through the introduction of dominant activating missense mutations in protein kinases, are known to be driver events of many cancer. Despite this, the identification of potential activating mutations has proven to be a difficult task, and has been limited to evolutionary and sequence-based comparisons with previously characterised mutations. This study aims to fill this gap by proposing a novel machine learning method for predicting missense activating mutations on protein kinases, named Kinact. Experimental data on 384 point mutations in 42 different protein kinases was collected from Kin-Driver, Clinvar and Ensembl databases. The resulting data sample was then manually curated and 258 mutations were mapped into solved 3D structures of the Protein Data Bank. Each protein was classified into one group of the Kinase Classification and a set of *in-silico* analysis were performed with sequence and structure data. The most descriptive features were then used as input for training and testing supervised learning algorithms and predictive classification models that rely on attributes solely from sequence level, structural level and in combination were generated. The best performing model was observed when a combination of structural and sequence-based features were used as evidence during the learning task, achieving a precision of up to 90% and Area Under ROC Curve of 0.96 under 10-fold cross-validation and precision of 81% and Area Under ROC Curve of 0.89 on blind tests. We show the best performing model of Kinact significantly outperforms the gold-standard methods used by clinical geneticists (p-value < 0.01), SIFT and PolyPhen-2, which achieved Area Under ROC Curve of 0.49 and 0.63 on the training data set, respectively and 0.67 and 0.53, respectively, on the blind test. Kinact conveniently combines high-performance open source web visualization tools to assist further research on how mutations affect protein kinases activity. The method is freely available as a user friendly, easy to use web server at <http://biosig.unimelb.edu.au/kinact/>

Keywords: bioinformatics. kinase. activating mutations, machine learning.

List of Figures

List of Tables

Contents

1 Introduction

1.1 Cell Signalling and Protein Phosphorylation

1.1.1 Role of Phosphorylation in Signalling

The ability of cells to recognize and correctly respond to their microenvironment is crucial for survival. Cell signaling is a complex communication process that controls most basic cellular activities and actions, allowing cells to quickly respond to internal and external stimuli (ALBERTS et al., 2014). In order to dynamically respond to cellular signals, however, fast dynamic switches are required. Protein phosphorylation is the most widespread type of post-translational modification, and acts as a rapid way of regulating protein behaviour and activity across most signalling pathways. It is estimated that one-third of the proteins in the human proteome are substrates for phosphorylation (COHEN, 2002). Phosphorylation exists in a delicate balance with dephosphorylation to orchestrate the activity of almost all cellular processes, including signal transduction, growth, division, differentiation, motility, organelle trafficking and membrane transport (COHEN, 1982; HUMPHREY; JAMES; MANN, 2015).

This process is coordinated by two large protein families: kinases and phosphatases. Protein kinases catalyze the transfer of the terminal phosphate group of Adenosine Triphosphate (ATP) to the hydroxyl group of a Serine (S), Threonine (T), or Tyrosine (Y) side chain of the target protein. This is counteracted by protein phosphatases, which catalyze the reverse reaction of phosphate removal (dephosphorylation). Therefore, the activity of any protein regulated by phosphorylation depends on the balance between the activities of the kinases that phosphorylate it and of the phosphatases that dephosphorylate it (ALBERTS et al., 2014).

Given the importance and range of cellular processes affected, the equilibrium between phosphorylation and dephosphorylation is stringently regulated. Loss of control over this regulation process, through the introduction of dominant activating mutations in kinases, are frequent driver events in many tumor types, contributing for the development and metastasis of many cancers (BOSE et al., 2013; GRABINER et al., 2014; TIACCI et al., 2017), along with the development of other metabolic disorders (LAHIRY et al., 2010).

1.1.2 Protein Kinases

Phosphorylation is catalysed by a type of enzymes called Kinases, and primarily occurs at tyrosines, serines and threonines. There are over 500 different kinases in the human proteome (MANNING et al., 2002), and while these can in part be differentiated by their substrate specificities, molecular characterisations have shown that this is not necessarily straightforward. Many kinases (for example the Serine-Threonine Kinases), exhibit substrate preferences for both Serines and Threonines; and there are also triple specificity kinase that recognize all three possible residues (MARSHALL, 1994). Furthermore, there are also approximately 50 pseudokinases that have lost the phosphorylation activity but are used as the building blocks for the assembly of multi-protein complexes, regulating other cellular processes such as proliferation and apoptosis (BOUDEAU et al., 2006).

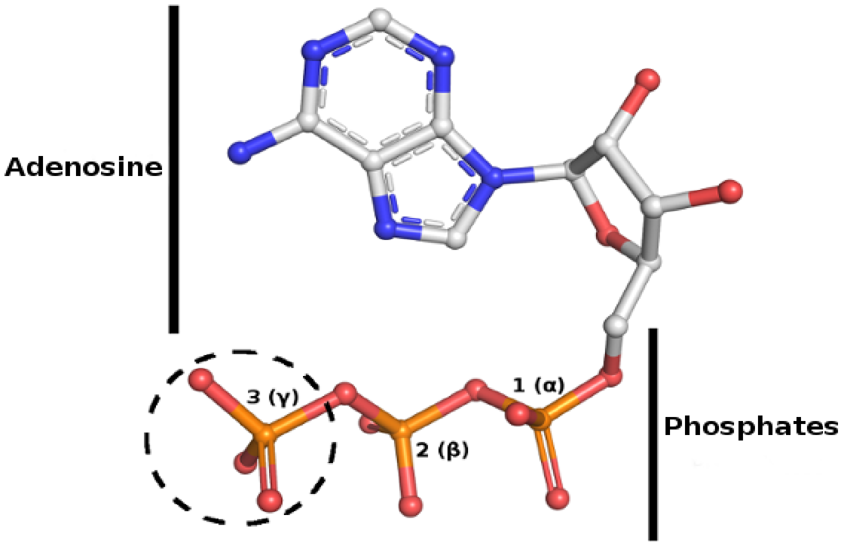
To assist with the analysis and comprehension of kinase function and evolution, as well as to compare related protein kinases in model organisms, a standard classification was proposed based on a set of features such as sequence similarity in the kinase domain and domain structure outside the catalytic domain, evolutionary conservation and known biological functions (MANNING et al., 2002). This classification clusters the kinases into 11 major groups that can be subdivided further into families and subfamilies. Table 1 presents a short description for each of the groups in the classification scheme.

1.1.3 Phosphorylation Reaction

Phosphorylation involves the transfer of a phosphate from ATP onto a particular amino acid within a protein. ATP is a source of chemical energy for the cell (BOYER, 1998) and is formed by two main chemical units, adenosine and phosphates (Figure 1). While the covalent bonds in the adenosine part are stable, the triphosphate is quite labile and reactive, allowing some proteins to transfer the gamma (γ) phosphate to the end of a specific amino acid. The three phosphates in ATP are connected by oxygens to each other and there are also negatively charged oxygens on the side of each phosphate. This highly negative charge density destabilizes the molecule and the electrostatic repulsion makes it a favorable high-energy reaction. A set of conserved residues on the binding pocket region of the kinase play key roles in the transfer of the phosphate group to a target protein.

Figure 2 summarizes the steps of the phosphorylation mechanism of B-Raf kinase. In step 1, binding of ATP to the kinase is strengthened by the presence of Mg^{2+} ions in the active site and an Asparagine in position 581 (Asn581). This induces the formation of the transition state of the phosphoryl reaction due electrostatic interactions that stabilise the ATP molecule in optimal orientation in the active site of the kinase (YU et al., 2011). An Aspartic Acid in position 576 (Asp576) deprotonates the Ser/Thr residue of the target protein. The negatively charged oxygen in the side chain of Ser/Thr of the target protein nucleophilic attacks the gamma phosphate of ATP in step 2. A Lysine residue in position 578 (Lys578) helps to preserve the conformation of the kinase when the attack to the gamma phosphate occurs. The magnesium complex formed in step 1 is broken down in step 3 and the Asp576 deprotonates releasing a proton (H^+). The products of the reaction

Figure 1 – ATP molecule structure and chemical reaction.



The molecule is divided into two parts: adenosine and phosphates. The phosphates are labeled as alpha (α), beta (β) and gamma (γ). Some enzymes are able to remove the γ phosphate from an ATP molecule and add it to the end of specific amino acids in a process known as phosphorylation.

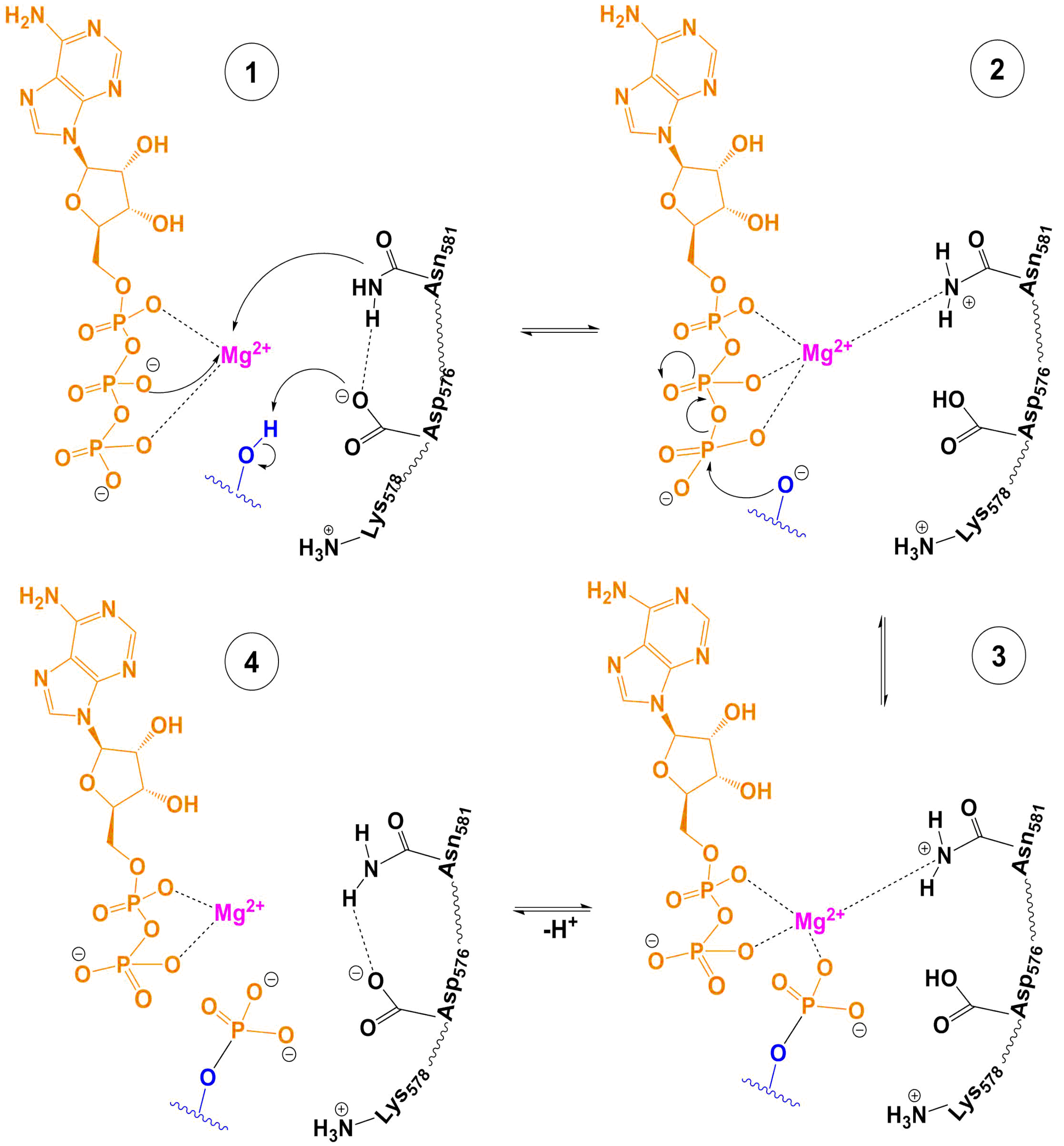
(Adenosine Diphosphate, also known as ADP, phosphorylated target protein and the kinase) are released in the last step (HANKS; HUNTER, 1995).

1.1.4 Conserved Structural Features of Kinases

Even though at the sequence level kinases are highly diverse, the core of protein kinases adopt a common fold with two well conserved subdomains, also known as N-terminal lobe and C-terminal lobe (Figure 3B). The N lobe is comprised of five strands of antiparallel β sheets (β1, β2, β3, β4 and β5) and one α-helix, called C helix. The larger C lobe is comprised of six α helices (αD, αE, αF, αG, αH and αI) and also two short antiparallel β sheets (β7 with β8 and β6 with β9). These two lobes are connected by a hinge region that allows the two lobes to articulate (Figure 3C). Studies suggest that role of this hinge region is vital for the enzyme to toggle between open and closed conformations as it goes through the catalytic cycle (JOHNSON; NOBLE; OWEN, 1996; KORNEV et al., 2006).

The ATP binding site is situated at the interface of these lobes and conserved residues in these two regions (Figure 3D) are essential for the phosphorylation process to occur (HANKS; QUINN; HUNTER, 1988; GIBBS; ZOLLER, 1991). A glycine-rich loop between β1 and β2 extends on top of ATP and then a Lysine residue in β3 binds α and β phosphates to keep the ATP molecule in place (HUSE; KURIYAN, 2002). If a mutations occurs in any of these regions the kinase will not be able to bind ATP and consequently it

Figure 2 – Phosphorylation mechanism of B-Raf kinase in 4 steps.



Mg^{2+} íon (colored in pink) and Asn581 residue in the ATP binding site of the kinase help stabilise the ATP molecule (colored in orange). The protein target (colored in blue) binds to the substrate pocket of the kinase (colored in black) near Asp576 of the kinase such that it can deprotonate the Ser/Thr side chain of the target protein. Step 2 shows the nucleophilic attack of the negatively charged Ser/Thr of the target protein to the gamma phosphate of ATP. The magnesium complex with the three phosphates is disassembled on the phosphoanhydride bond between beta and gamma phosphates, and Asp576 deprotonates in step 3. Lastly, the products of the reaction (ADP molecule, phosphorylated target protein and kinase) are released in step 4. Modified from <https://commons.wikimedia.org/wiki/File:B-Raf_Phosphorylation_Mechanism.png>.

will lose its phosphorylation activity. An interaction between a conserved Glutamic Acid in C helix interacts with the Lysine from $\beta 3$ is also important for keeping the conformational stability of the kinase.

On the other side of the binding pocket there is a highly conserved DFG motif (Aspartic Acid (D), Phenylalanine (F) and Glycine (G), respectively) in the loop that connects strands $\beta 1$ and $\beta 2$. This motif is the start of a large loop called activation loop. The amino acid sequence in this loop determines whether the kinase will recognise Tyrosines/Serines or Threonines.

Finally, a YRD motif, or HRD in many kinases (Tyrosine/Histidine (Y/H), Arginine (R) and Aspartic Acid (D), respectively), in the catalytic loop on C lobe is another crucial region in protein kinases due a conserved Aspartic Acid responsible for the actual transfer of the gamma phosphate from ATP onto a substrate Serine/Threonine or Tyrosine. If a mutation occurs in this Aspartic Acid the kinase will still bind to ATP, but it will not actually transfer a phosphate (JOHNSON et al., 1998).

1.2 Activating Mutations in Kinases

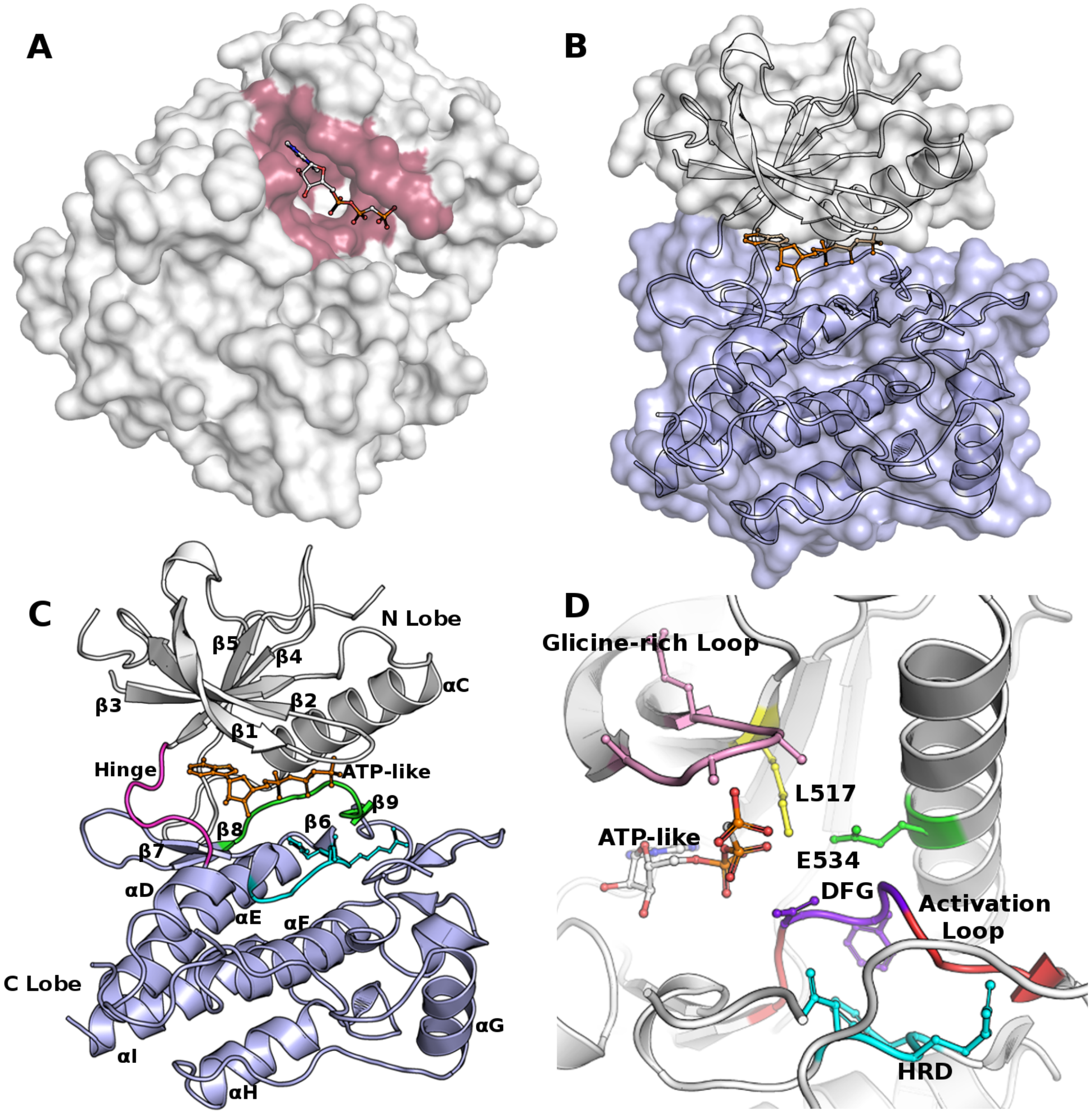
Single-nucleotide variants (SNVs) are mutations in a single nucleotide that occur at a specific positions in the genome. Such variations are crucial for evolution by the introduction of diversity into genomes. This work is particularly interested in SNVs that occur within coding sequences of genes, in other words, single point mutations that fall in regions of the genetic code that encode proteins. These coding SNVs can be further grouped into Synonymous and Non-synonymous mutations.

Synonymous substitutions, often called silent mutations, are those that the variation of one base for another within a protein-coding portion of a gene does not alter the produced amino acid. This is possible because the proteins are encoded by "triplets" of nucleotides, called codons, that are responsible for adding a particular amino acid to the protein chain. However, due the redundancy of the genetic code, different codons code for the same amino acid, as summarized in Figure 4. For example, all codons starting with GG (GGA, GGC, GGG and GGU) are translated to the amino acid Glycine, making those codons synonyms. In this sense, a mutation that alters a nucleotide but produces a synonymous codon is a silent mutation (CHAMARY; HURST, 2009).

Unlike Synonymous SNVs, Non-synonymous single-nucleotide variants (nsSNVs) within the protein coding regions of the genome replaces the amino acid at specific positions and can have two effects: a codon that codes for a different amino acid (missense mutations) or the introduction of a stop-codon that results in a truncated protein (non-sense mutation).

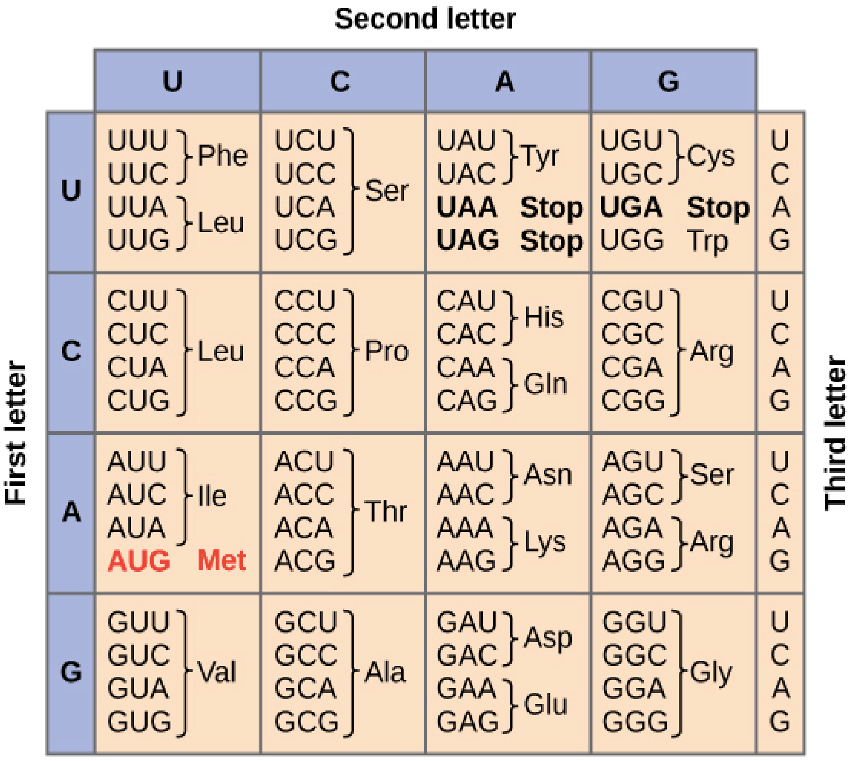
As mentioned in section 1.1.1, protein kinases are involved in many complex

Figure 3 – FGF Receptor 2 (FGFR2) kinase in complex with ATP analog molecule (PDB: 2PVF).



A) depicts the surface of the kinase and the binding site highlighted in red with the ATP analog bound. B) highlights the two lobes that comprise the kinase core structure, N-terminal Lobe (light gray) and C-terminal Lobe (light blue). C) shows the cartoon representation of the same molecule with the conserved secondary structures identified for the N lobe ($\beta1$, $\beta2$, $\beta3$, $\beta4$, $\beta5$ and C helix) and C lobe (αD, αE, αF, αG, αH, αI, $\beta6$, $\beta7$, $\beta8$, $\beta9$). The hinge loop which connects both subdomains is colored in pink, the loop that contains HRD motif is highlighted in cyan and the activation loop is colored in green. D) displays a closer look at the binding site with conserved regions and important residues highlighted. The Glycine-rich loop on top of the ATP-like molecule is colored in light pink. The Lysine 517 responsible for binding the beta phosphate is colored in yellow. The Glutamic Acid 534 responsible for stabilization of the binding site is colored in green. The DFG motif necessary for transferring the phosphate to the protein is colored in purple. The activation loop, in which the DFG motif is contained is colored in red. The HRD motif containing an Aspartic Acid (626) residue that also plays a role in transferring the phosphate from the ATP to the protein is colored in light blue.

Figure 4 – Genetic code table.



	Second letter								
First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		**UAA**	**Stop**	**UGA**	**Stop**	A
	UUG		UCG		**UAG**	**Stop**	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

The outer columns and row present the letter options for a codon. The table summarizes all combinations and what they implicate. For example, a UAU codon will be coded as a Tyrosine (Y), while UAA will be identified as a stop signal to translation. The table also shows that most amino acids can be generated by more than one 3 letter sequence. Source: "The genetic code," by OpenStax (2015).

cellular processes and because of that they need to be tightly regulated. The introduction of dominant activating (gain of function) mutations in these proteins are associated with disturbance of that regulation due the hyperphosphorylation of their targets, contributing to the transformation of proto-oncogenes into oncogenes. For example, Hairy Cell Leukemia (HCL) is a rare and slow-growing blood cancer in which the bone marrow produces too many B cells (lymphocytes). A recent study reported that an activating mutation in the BRAF kinase, in which a Valine in position 600 is replaced by a Glutamic Acid, plays key role in the development of HCL and is also related to the cascade activation of three other kinases: RAF, MEK and ERK (TIACCI et al., 2017).

The development of small-molecule kinase inhibitors has therefore been seen as an attractive alternative to conventional (cytotoxic) chemotherapy. The goal of these inhibitors is to reduce the activity of kinases that promote cancer development, survival or metastasis (GHARWAN; GRONINGER, 2015). The mechanisms by which the inhibition is performed may vary from simple ATP-competitive small-molecules, such as gefitinib that have been used for treatment of non-small-cell lung cancer through the inhibition of the Epidermal Growth Factor Receptor (EGFR), to more complex and flexible allosteric inhibitors that can bind either to the kinase domain or to sites outside the kinase domain (ALBANAZ et al., 2017).

However, studies have shown that activating mutations have also been identified to contribute for the development of mechanisms of resistance to treatments with these kinase inhibitors. For instance substitutions of an Aspartic Acid in position 1067 by either a Tyrosine, Alanine or a Valine were identified to confer resistance to PI3K inhibitors used in the treatment of breast cancer to suppress the activated pathway of Phosphoinositide 3-kinase (PI3K) (NAKANISHI et al., 2016).

1.3 Computational Studies for Predicting Effects of Mutations

The development of databases with experimental biological data have been crucial in the field of bioinformatics. This has allowed scientists to access a large variety of relevant biological information from a vast range of organisms, which supported the development of computational approaches that help to compare, understand and elucidate major challenges in the field based on curated data (BAXEVANIS; BATEMAN, 2015).

Several computational approaches built upon different biological assumptions and applicability for unraveling genotype-phenotype correlations have been proposed for predicting the effects of mutations. These methods can be broadly classified into those that explore effects of mutations based on the amino acid sequence of a protein, and those that analyze the structural information of proteins in an attempt to elucidate mutation mechanisms and molecular effects.

1.3.1 Sequence-based Methods

The two main methods used by clinical geneticists to study the effects of coding mutations in the human genomes are SIFT (NG; HENIKOFF, 2001) and Polyphen-2 (ADZHUBEI et al., 2010). The predictions from these two methods strongly rely on information from the sequence of a protein, such as the analysis of residue conservation at the mutated position. Both methods present a scoring system to denote how likely a mutation is to affect protein function.

SIFT (Sorting Intolerant From Tolerant) is an algorithm that uses multiple sequence alignment (MSA) information to predict tolerant and deleterious substitutions for every position of a query sequence. It compares conserved residues within the protein family and assesses specific amino acid positions for their ability to tolerate substitution by different classes of amino acids. It assumes that structurally and functionally important positions should be conserved in an alignment of the protein family, whereas non-essential residues will be under less selective pressure and should thus appear more diverse across the alignment. For example, if a position in a MSA of a protein family has only a conserved

Aspartic Acid, as seen in protein kinases catalytic sites described in subsection 1.1.4, it is assumed that substitution to any other amino acid must not be tolerated and that the Aspartic Acid is a key residue for protein function at that specific position. As a result, a mutation to any other amino acid residue will be predicted to be deleterious to protein function. Likewise, if a position in an alignment contains amino acids with hydrophobic side chains, such as Tyrosine, Isoleucine and Valine, SIFT infers that this position can only contain amino acid residues with hydrophobic side chains, and substitutions to charged or polar residues, will be predicted to affect protein function (NG; HENIKOFF, 2003). A web server is freely available[1] and allows users to run single and batch predictions as well as downloading the source code of SIFT for local running.

PolyPhen-2 (Polymorphism Phenotyping) is a tool for predicting the impact of an amino acid replacement on the structure and function of a human protein. In contrast to SIFT, which relies solely on sequence conservation, PolyPhen-2 performs functional annotation of single nucleotide variants using annotated UniProt entries, maps mutations in coding regions to gene transcripts, extracts protein sequence annotations and structural attributes, and builds conservation profiles. It then estimates the probability of the missense mutations being damaging based on a combination of all these properties using a Naïve Bayes classifier. PolyPhen-2 is available as stand-alone software and via a web server[2] (ADZHUBEI et al., 2010).

Studies assessing the predictive value of SIFT and PolyPhen-2, for the analysis of mutations identified in genes associated with human disease, have demonstrated that the predictive outcome of both methods have to be analyzed with caution due relatively low specificity especially when dealing with substitutions identified in genes that harbor gain-of-function mutations (DOSS; SETHUMADHAVAN, 2009; VALDMANIS; VERLAAN; ROULEAU, 2009; FLANAGAN; PATCH; ELLARD, 2010). Combinations of predictive values were also evaluated for SIFT, PolyPhen-2 and a set of other predictive models, but no significant improvement was observed (GNAD et al., 2013).

1.3.2 Structure-based Methods

Structure-based approaches, on the other hand, use protein structural data from the 3D space of a natively folded protein and try to predicts the impact of a mutation on this space. The Protein Data Bank (PDB) (BERMAN et al., 2000) is one of the main sources from which such data can be easily extracted. Even though these methods are essentially based on the same structural data, they based their assumptions through broadly different, but sophisticated, approaches, such as statistical potential function energy calculations

[1] <http://sift.bii.a-star.edu.sg/>

[2] <http://genetics.bwh.harvard.edu/pph2/>

and mining of structural patterns.

SDM (Site Directed Mutator) is a method that relies on amino acid propensities derived from environment-specific substitution tables for homologous protein families that serve as input for a statistical potential energy function and encompass an evolutionary view of the constraints from the immediate residue environment. This approach examines amino acid susceptibility for the wild-type in contrast with mutant proteins in the folded and unfolded states in order to estimate the free energy differences between them (TOPHAM; SRINIVASAN; BLUNDELL, 1997; WORTH; PREISSNER; BLUNDELL, 2011; PANDURANGAN et al., 2017). A freely available web server for running SDM is available[3].

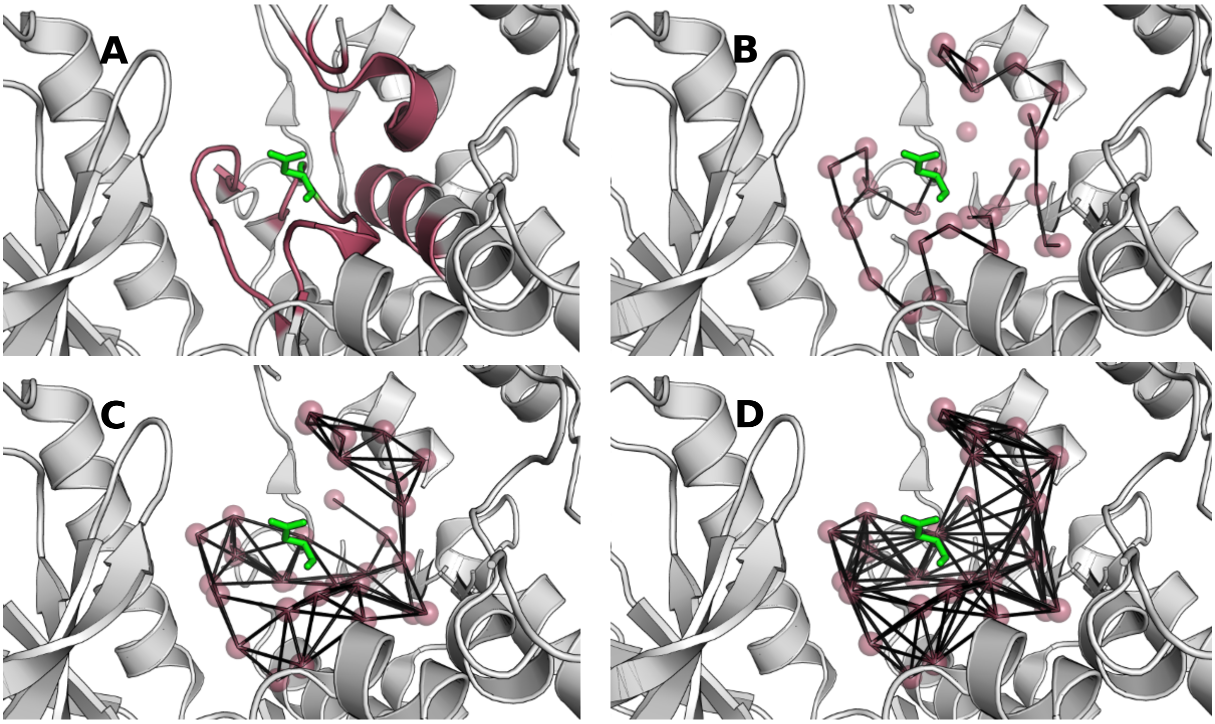
mCSM is a machine learning method to predict the effects of missense mutations that relies on structural signatures. mCSM was built upon a graph-based concept (PIRES et al., 2011) used for representing network topology by distance patterns in the study of biological systems. For this approach, residue environments are represented as graphs where nodes are the atoms and the edges are the physicochemical interactions established among them. Figure 5 shows the network topology of the contact graph for the structural environment of Aspartic Acid in position 626 (Asp626) in the active site of the kinase FGFR2 at different cutoff values. mCSM uses such graph representations to extract geometric and physicochemical patterns that define the chemical environment used in the supervised learning task (PIRES; ASCHER; BLUNDELL, 2014b). Like the other methods presented, mCSM also has a freely available web server[4]. Such graph-based signatures approach has been successfully employed on variety of tasks including large-scale receptor-based protein ligand prediction (PIRES et al., 2013), quantification of effects of mutations on protein-small molecule affinity in genetic disease (PIRES; BLUNDELL; ASCHER, 2016), antibody-antigen affinity changes (PIRES; ASCHER, 2016) and more recently for predicting the effects of mutations on protein-nucleic acids interactions (PIRES; ASCHER, 2017). As it is the only available method that can accommodate within a single framework all the different measures of interaction and stability. Specific attention will be made to its use.

DUET is an integrated approach for predicting the effects of mutations on protein stability that takes advantage of the distinct techniques and property evaluation between SDM and mCSM by trying to combine them in a consensus prediction (PIRES; ASCHER; BLUNDELL, 2014a). DUET unifies the results of the separate methods in an optimized predictor using Support Vector Machines (SVMs) trained with Sequential Minimal Optimization (SHEVADE et al., 2000). It is shown that DUET improves overall accuracy of the predictions of both methods on their own. Like the other methods presented in this

[3] <http://structure.bioc.cam.ac.uk/sdm2>

[4] <http://structure.bioc.cam.ac.uk/mcsm>

Figure 5 – Topology of the contact graph of FGF Receptor 2 (FGFR2) kinase protein active site structure (PDB: 2PVF).



A conserved Aspartic Acid in position 626 (Asp626), plays a key role in phosphorylation as described in previous section. Different cutoff values results in different distance patterns. The residues surrounding the Asp626 with a maximum distance of 10Å are represented by its alpha carbon as red spheres. A) presents the structure of the kinase catalytic site with the Asp626 highlighted in green and the residues surrounding as red spheres in a cartoon representation. B) shows the signatures for residues with a 5Å distance. C) shows the signatures for the residues with a 7Å. Finally, D) shows the signatures for the residues with 9Å distance from the Asp626.

section, DUET also is freely available as a website[5].

1.4 Motivation

Advances in next generation sequencing techniques are leading to the identification of many novel mutations, including in kinases. In the absence of experimental information, it is currently quite difficult to identify mutations that are likely to lead to the loss of control over kinase regulation, through the introduction of dominant activating mutations. However, it is important to recognise these variants, as they can drive the development and metastasis of many cancers. In this sense, the understanding of the impact of mutations upon kinase activity has significant influence on patient outcomes and treatment.

Over the last two decades, several computational approaches have been proposed

[5] <http://biosig.unimelb.edu.au/duet>

for predicting the effects of mutations from sequence and structure, built upon different biological premises and used for unraveling genotype-phenotype correlations. While the effects of mutations that disrupt activity have been well studied, no robust computation methods for identifying activating mutations have been proposed.

This thesis aims to fill this gap by proposing a novel machine learning method for predicting missense activating mutations in kinases from structure, sequence and structure-sequence perspectives.

Table 1 – Standard Kinase Classification Scheme.

Group name	Description
AGC	Named after the Protein Kinase A,G and C families (PKA, PKC, PKG), this group contains many core intracellular signaling kinases which are modulated by cyclic nucleotides, phospholipids and calcium.
CMGC	Named after another set of families (CDK, MAPK, GSK3 and CLK), this group has a diversity of functions in cell cycle control, MAPK signaling, splicing and other unknown functions.
CAMK	Best known for the Calmodulin/Calcium regulated kinases (CAMK) in CAMK1 and CAMK2 families, this also has several families of non-calcium regulated kinases.
CK1	A small but ancient family. Originally known as Casein Kinase 1 (from a biochemically assay with a non-physiological substrate), and now renamed to Cell Kinase 1.
STE	Homologs of the yeast STE7, STE11 and STE20 genes, which form the MAPK cascade, transducing signals from the surface of the cell to the nucleus.
TK (Tyrosine Kinase)	This group phosphorylates almost exclusively on tyrosine residues, as opposed to most other kinases that are selective for serine or threonine
TKL (Tyrosine Kinase-Like)	The group most similar to tyrosine kinases, but whose activities are generally on serine/threonine substrates.
RGC	Receptor Guanylate Cyclases. This small group contains an active guanylate cyclase domain, which generates the cGMP second messenger, and a catalytically inactive kinase domain, which appears to have a regulatory function.
PKL	Contains a number of diverse families that share a PKL fold and catalytic mechanism with the ePKs but do not have substantial sequence similarity. This group also contains a number of lipid, sugar, and other small-molecule kinases.
Atypical	Diverse group of kinases and candidate kinases with no structural similarity to ePKs.
Other	This group consists of several families, and some unique kinases that are clearly ePKs but do not fit into the other ePK groups.

Kinases were clustered by sequence similarity in the kinase domain and additional information from domains outside the catalytic domain, evolutionary conservation and known functions were also added. All the groups are named and a short description provided. Source: Manning et al. (2002)

2 Aims

2.1 General Aim

To propose, implement and validate a novel machine learning model for predicting missense activating mutations in kinases based on sequence and structural data, aiming for a better understanding of the role of these mutations in diseases and guide the development of improved and more personalized treatment strategies.

2.2 Specific Aims

- Data acquisition and curation of experimentally characterized missense mutations in kinases;
- Feature engineering in order to identify the most descriptive attributes from sequence and structure levels;
- Training and testing supervised learning algorithms with the selected features;
- Development of a user-friendly, freely available web server with the best performing predictive model;

3 Materials and Methods

The methodology workflow of this work is composed of four main steps, as displayed in Figure 6, in consonance with the four specific aims described in section 2.2. First of all, experimental well-characterised mutations were collected from publicly available databases, as discussed in section 3.1. This data had to be carefully curated in order to filter the mutations that satisfied strict inclusion criteria for the subsequent *in-silico* analyses and model development. All the data generated from the *in-silico* analyses of the mutations was then supplied as input for training machine learning algorithms. The best performing models were selected and a web server (Kinact) was developed. Each step is described in detail in the following sections.

3.1 Data Sets

3.1.1 Data collecting

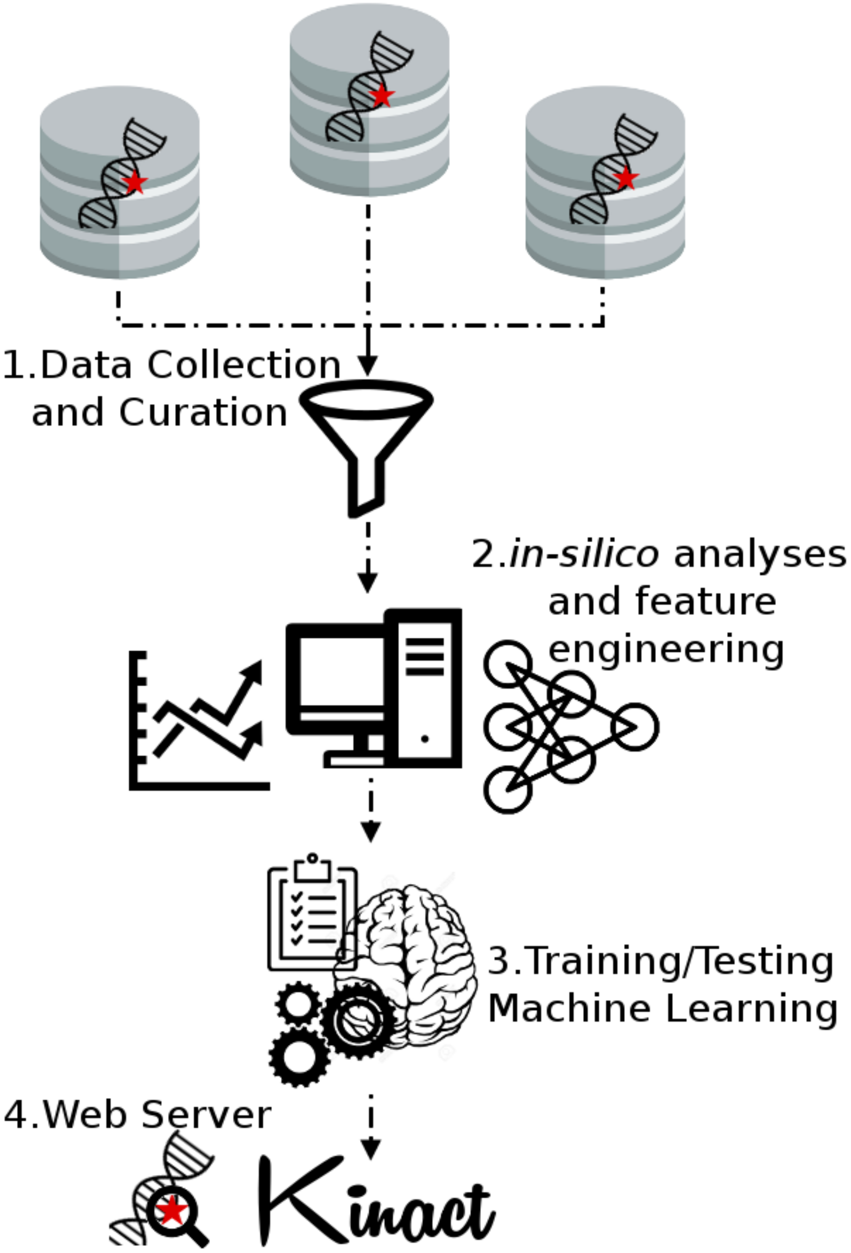
Driver mutations are somatic mutations that result in a selective advantage for tumor cells, and for the scope of this work they will be divided into two groups: activating and non-activating mutations. The non-activating mutations contain mutations that disrupt activity (inactivating) and those that have no significant biological effect (neutral mutations). On the other hand, activating or gain-of-function mutations result in the increased activity of the kinase, and in proto-oncogenes can lead to their activation into oncogenes.

The data set used in this work is derived from three different databases of mutations. These are described below as well as how the data that comprises our mutation data set was extracted from them.

Kin-Driver database (SIMONETTI et al., 2014) is a manually curated database of driver mutations in protein kinases with experimental evidence demonstrating their functional role. Kin-Driver is a MySQL relational database offering structural and sequence data cross-referenced with the database of Catalogue of Somatic Mutations in Cancer (COSMIC) (FORBES et al., 2009) and with a set of manually curated mutations. Each mutation in Kin-Driver is displayed with its validation status (activating, inactivating or unknown), the mutation type (missense, insertion, deletion, nonsense, frameshift or indel), its absolute and relative frequencies in human tumors and the PubMed reference describing a particular mutation as activating/inactivating. The Kin-Driver database is available as a website[1].

[1] <http://kin-driver.leloir.org.ar/>

Figure 6 – Methodology workflow.


1.Data Collection and Curation
2.*in-silico* analyses and feature engineering
3.Training/Testing Machine Learning
4.Web Server
Kinact

The Kinact methodology can be divided into four steps to help fulfill the specific aims from section 2.2. In step 1, data was collected from databases of mutations on kinases with experimental evidence. The resulting data was then curated to identify which mutations had their regions mapped into solved PDB structure. Each mutation was then classified into one of the groups of kinases, described in section 1.1.2, with Kinannote (GOLDBERG et al., 2013). In step 2 a set of *in-silico* analyses were performed with sequence and structure information providing features that were used as input for training supervised learning algorithms in step 3. After evaluating all the algorithms from previous step, in the last step, a web server was implemented relying on the best models identified.

The data curated from this database results in 318 mutations across 42 different proteins, from which 191 were mapped to known protein structures in the Protein Data Bank.

The second database used on data collection was ClinVar (LANDRUM et al., 2014). This is a medical genetics resource that collects assertions of the relationships between human sequence variations and phenotypes. Submissions to ClinVar may specify the variation, the phenotype, the interpretation of the medical importance of the variation, the date in which the interpretation was last evaluated and the evidence supporting that interpretation, along with information about the author of the submission. For this database, 40 missense mutations were collected from 14 different protein kinases with clinical testing evidence. From these, 23 mutations could be mapped to experimentally determined structures in the PDB. Of the remaining 17 mutations, 14 had no experimentally determined structures, and 3 were present in disordered regions of known structures, limiting structural analyses.

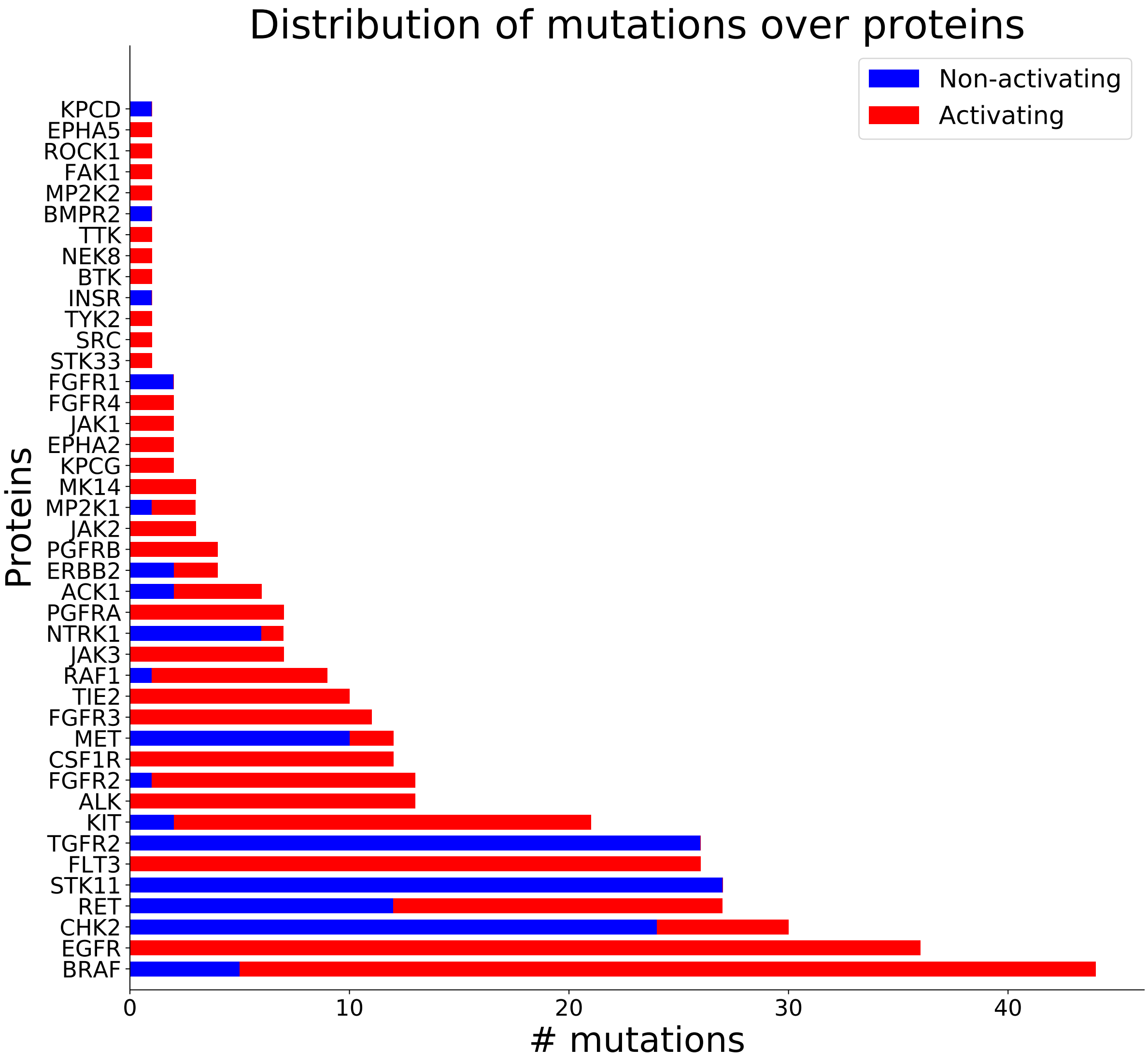
Lastly, the variations resource from the Ensembl project (HUBBARD, 2002) was also used as a source for mining data on missense variations. The Ensembl project is a joint effort by the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI) intending to provide an extensive resource of high-quality genomic data such as gene annotations, multiple sequence alignments, and whole genome variation, alongside tools for cross-species genomics analysis, at both sequence and gene levels. The Ensembl variation resource integrates data from a set of databases, such as dbSNP (SHERRY, 2001) and COSMIC (FORBES et al., 2009), as well as information on associated diseases and phenotype information when available. Only SNVs classified with clinical significance as benign or pathogenic were extracted for analysis. SNVs marked as benign were then classified as non-activating mutations, and those marked as pathogenic were grouped into activating mutations. This identified 44 missense mutations, and all of them could be mapped onto experimental structures in the Protein Data Bank.

Regarding the mapping of the mutations to experimental 3D structures in the PDB, cross-reference entries from Uniprot (BAIROCH et al., 2005) were used to try to identify protein regions that had been structurally resolved, particularly for the kinase domain. Two questions were considered when looking for experimental structures. First, where no structure associated with the Uniprot accession number was identified containing the wild-type residue at the position specified in the mutation, a new attempt was performed with the position mapped in PDBSWS[2] (MARTIN, 2005). Second, when multiple structures mapped the wild-type residue the decision of which structure to use was based on the resolution of the experimental structure, which is a measurement of the level of detail available within the electron density. Lower resolution indicates better quality structures.

[2] <http://www.bioinf.org.uk/pdbsws/>

As discussed in section 1.1.4, protein kinases may present two conformational states which impact its activating status. For this work, no consideration regarding this conformational state was taken into account when selecting the 3D structure to be used as input for the *in-silico* analysis, even though this could have an impact in the evidence provided to train/test the machine learning algorithms described in section 3.2.2. This remains a perspective for further evaluation in future implementations of this methodology.

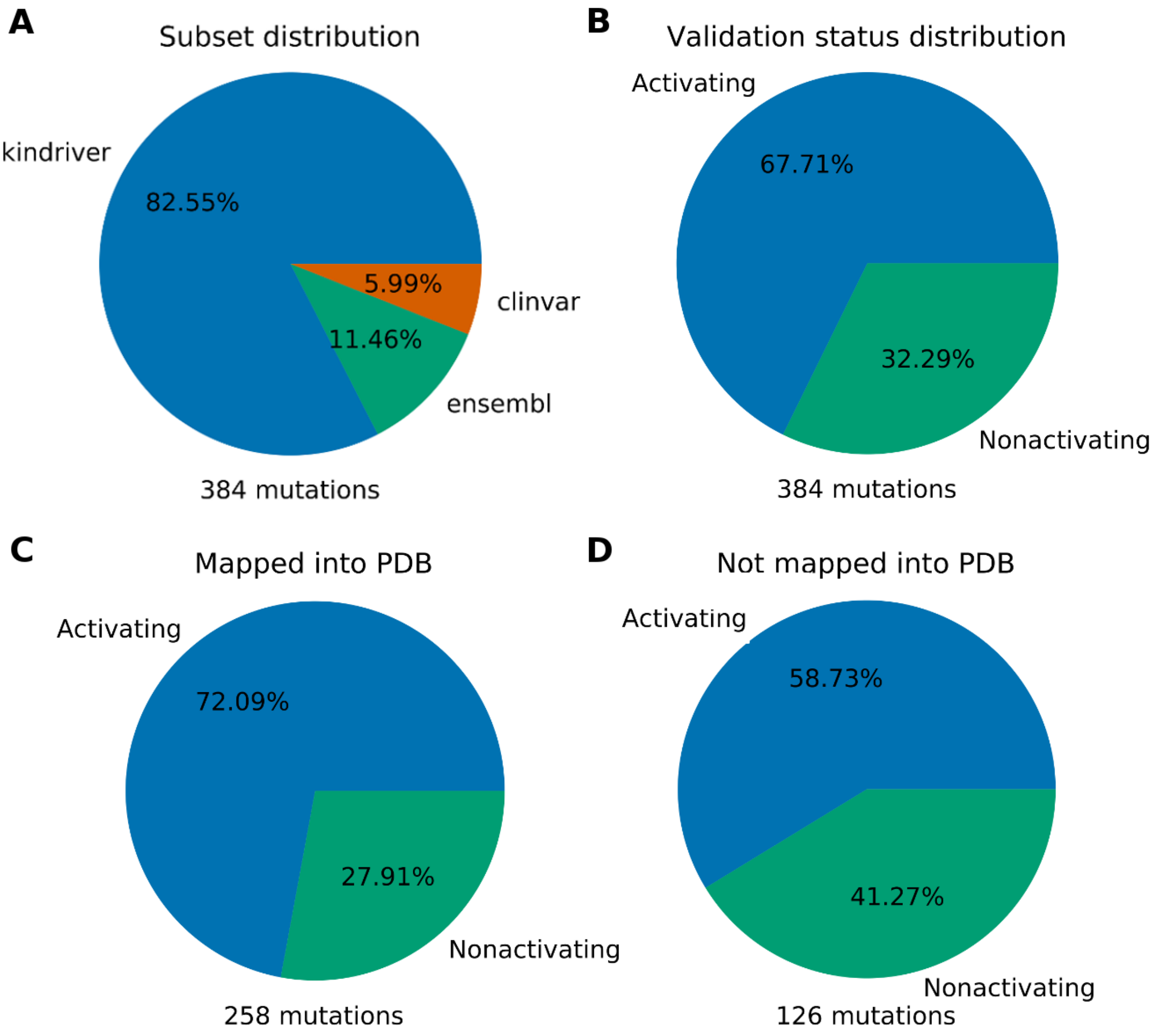
Figure 7 – Summary of number of mutations per protein within the data set.



The top 3 proteins with the highest number of mutations are BRAF, EGFR and CHK2 with 44, 36 and 30 mutations respectively. On the other hand, EPHA5, SRC, ROCK1 are among the proteins with only one mutation identified. Bars are colored according to the class of mutations: a blue bar indicates the amount of non-activating mutations and a red bar indicates the number of activating ones.

The complete data set comprising mutations collected from Kin-driver, Ensembl and Clinvar contained 384 mutations distributed across 42 proteins, of which 258 could be mapped onto experimental structures. Figure 7 depicts the distribution of mutations across the different proteins in the data set, of which BRAF, EGFR and CHK2 are the top 3 most mutated proteins with 44, 36 and 30 mutations respectively. Figure 8 summarizes other aspects of the data set, such as its composition and the class distribution over the full set of mutations, and broken down into those mutations who could be mapped onto protein structures, and those with no experimental structural information.

Figure 8 – Distribution of mutations and validations status.



A) shows the distribution of mutations regarding the subsets of origin (Kin-driver, Clinvar and Ensembl) that were used for data collection. Most of the data was obtained from Kin-driver followed by Ensembl and Clinvar. B) depicts the distribution of validation status (activating, non-activating) of mutations across the entire data set. C) shows the class distribution of mutations that have their region mapped onto 3D structures. These comprise 258 missense mutations from all the data collected. Finally, D) displays the class distribution over the mutations that were not structurally mapped into solved structures.

The distribution of mutations by the origin of the data, shows that most mutations in the final data set are from Kin-driver, followed by Ensembl and Clinvar, with 317, 44 and 23 mutations respectively, as shown in Figure 8A. Regarding the distribution of the functional effects of these mutations, there is a predominance of activating mutations over non-activating (Figure 8B). This uneven distribution is consistent over mutations that had experimental structures available. On the other hand, for mutations without structural information, the difference between the two classes is not as substantial (Figures 8C and 8D). This difference indicates that there is a selection bias in those mutations chosen for experimental validation and poses a challenge in building a predictor for unbalanced classes (SUN; WONG; KAMEL, 2009). Such bias can be justified by the role dominant activating mutations play on driving the development of metastasis of many cancers and other metabolic diseases, as discussed in section 1.2.

For mutations that did not have their region mapped into PDB structures, comparative modeling methods for predicting the 3D structure could greatly benefit future implementations of this study. These methods perform predictions for a given sequence (target) based primarily on its alignment to one or more protein of known structure (templates) (MISURA; BAKER, 2005; FISER, 2004; WEBB; SALI, 2016). This approach would benefit this work in future implementations by increasing the number of instances on the sample set of mutations that could be included in feature engineering analysis, especially due the highly conserved structures discussed in section 1.1.4. In addition, it could also increase the amount of evidence to be used for training the machine learning algorithms.

The final data set with all mutations mapped into PDB is available at the data section of the Kinact web server[3].

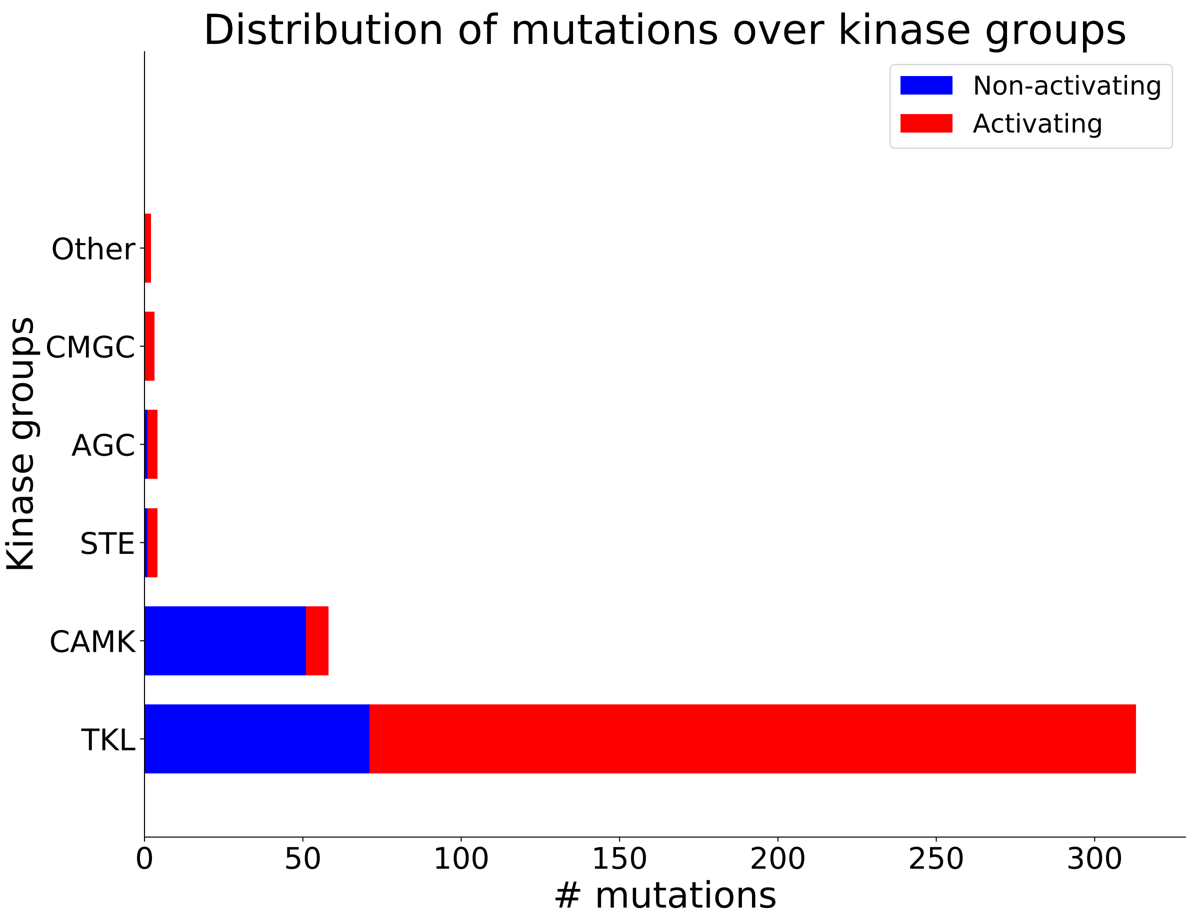
3.1.2 Data preparation

After data collection, all proteins were then assigned to a kinase super-group according to the classification described in Table 1 by Kinannote 1. This is a computer program that identifies and classifies members of the eukaryotic protein kinase superfamily. It uses a Hidden Markov Model (HMM) in combination with a position-specific scoring matrix for kinase identification, and subsequently a BLAST comparison with KinBase. Figure 9, summarizes the distribution of mutations of the dataset over the kinase groups, and also highlights the class distributions for each group. TKL (Tyrosine Kinase-Like) is the most predominant group of kinases across the data set with 313 mutations, followed by CAMK (Calmodulin/Calcium regulated kinases) with 58 mutations.

The abundance of mutations belonging to the TKL group is consistent with the fact that this is the most diverse and polymorphic group of kinases, and contains a large number of kinases that are difficult to classify. Among all the superfamilies in this group, two of them are involved in a set of molecular functions that can affect tumor progression, contributing directly or indirectly to cancer: MLKL (Mixed Lineage Kinases domain-Like)

[3] <http://biosig.unimelb.edu.au/kinact/data>

Figure 9 – Distribution of mutations over kinase groups.



Tyrosine Kinase-Like (TKL) kinases have the highest number of mutations followed by Calmodulin/Calcium regulated kinases (CAMK). The Other group presents only 2 mutations. Bars are colored according to the number of mutations on each class: red represents activating and blue non-activating mutations.

and RAF (Rapid Accelerated Fibrosarcoma). MLKL is a superfamily of pseudokinases that are involved in necroptosis along with proteins from the RIPK (Receptor-Interacting Protein Kinases) superfamily, also in this group. Proteins in the RAF superfamily are known for their role in embryogenesis, cell proliferation and differentiation (MANNING et al., 2002). When looking at the distribution of proteins, summarized in Figure 7, the protein with the largest number of characterised mutations is the B-RAF kinase, with 44 mutations, which is a member of the RAF superfamily. Despite the fact that by itself a mutation in B-RAF is not sufficient for cancer development, studies have reported that around 70% of human melanomas present mutations in this kinase (DHOMEN; MARAIS, 2007).

Since the distribution of classes of mutational effects over the full data set is notably unequal (more activating mutations than non-activating), as shown in Figure 8B, and in order to support model selection assessment and avoid biased estimations, described in more details in section 3.2, the data set was divided into training and testing subsets of mutations. This partitioning of the data is a crucial step in order to evaluate the machine learning algorithms separately across attributes, and also to assess the impact of each attribute in the final models, avoiding overfitting. Figure 10 summarizes the distribution of activating and non-activating classes of mutations used as input for training as well as a separate independent testing set, as described in section 3.2.3, for mutations with only sequence information (Figure 10A and 10B) and also for the ones with structural mapping in the PDB (Figure 10C and 10D).

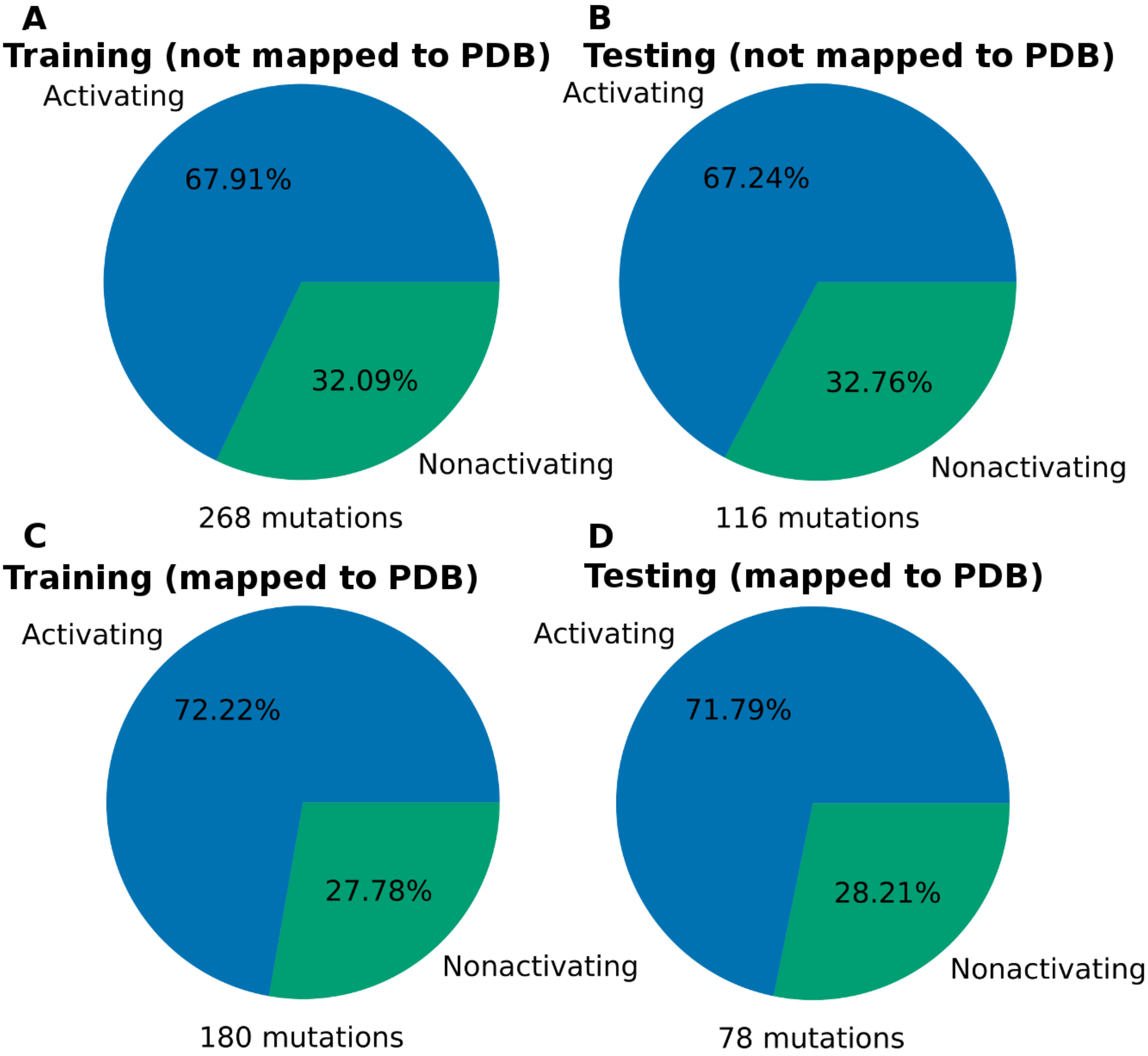
For both sample sets of mutations, with and without structural mapping into PDB, due the limited amount of data and for the sake of generating more robust predictive models when training the machine learning algorithms, described in section 3.2, a greater amount of data was reserved for the training (70%) than for the testing (30%) sets. In addition, when separating training and testing sets the uneven proportions of each class was preserved, about 70% of activating and 30% of non-activating mutations.

3.1.3 Attribute generation

After data collection, kinase group assignment and preparation structural and sequence attributes that will be used as evidence in the machine learning step were generated, as will be further described in section 3.2.

The task of predicting and understanding the effects of mutations in proteins can be accomplished using approaches that look at different biological attributes, each with their own assumptions and limitations, as described in section 3.1. Protein structural and sequence features have been the two most popular categories of attributes used by computational methods. Sequence-based features have focussed predominantly on the analysis of sequence residue conservation throughout a protein family and also of homologs proteins (NG; HENIKOFF, 2003). By contrast, previous studies have used a wide range of features that rely on structural data, including type of secondary structure, solvent accessibility and dihedral angles (CHASMAN; ADAMS, 2001; GUEROIS; NIELSEN; SERRANO, 2002). More sophisticated approaches, such as modeling the atoms of a protein structure as a graph and extracting distance patterns, also known as structural signatures, have also served as input for machine learning algorithms for predicting effects of mutations on protein stability and interactions, described on section 1.3.2. A combination of sequence and structural information has also been proven to be valuable when predicting damaging mutations (ADZHUBEI et al., 2010). Based on these assumptions, the attributes used in this work were categorized into six different groups. These categories are summarized in Table 2.

Figure 10 – Distribution of validations status class of mutations in the data set for training and testing sets.



Distribution of validations status class of mutations in the data set for training and testing sets. The complete set of mutations were divided into two groups. The first comprising all 384 mutations identified during data collection, as described in section 3.1.1. The second group contains only those mutations that had their region mapped into structures on the PDB. Each group is split into training and testing data for the machine learning algorithms as discussed in sections 3.2.3. A) displays the distribution of the two classes of mutations (activating, non-activating) over the set of mutations used in training for data without structural mapping. B) presents the distribution of the two classes of mutations for the testing on the same type of data. C) and D) introduces the class distribution for training and testing, respectively, for mutations that had their region mapped into 3D structures of PDB.

Table 2 – Description of categories of attributes generated presenting short summary of the attributes and the data and tools used for their calculation.

Category name	Attributes	Rely on	Tools	References
Wild-type residue environment	Type of secondary structure, solvent accessibility, residue depth, dihedral angles, flexibility, minimum distance to catalytic sites and relative b-factor	Structure	Biopython, ENCoM, CSA	Chapman e Chang (2000), Frappier, Chartier e Najmanovich (2015), Porter, Bartlett e Thornton (2004)
Wild-type residue interactions	clash, covalent, Van der Waalsvdw clash, vdw, proximal, hydrogen bond, weak hydrogen bond, halogen bond, ionic, metal complex, aromatic, hydrophobic, carbonyl, polar hydrogen bonds without angles, weak polar weak hydrogen bonds without angles	Structure	Arppegio	Jubb et al. (2017)
Structural signatures	pattern of distance among the atoms of the structure based on graph modeling	Structure	mCSM	Pires, Ascher e Blundell (2014b)
Stability change upon mutation	Variation of Gibbs Free Energy - $\Delta\Delta G$	Structure	SDM, mCSM and DUET	Topham, Srinivasan e Blundell (1997), Worth, Preissner e Blundell (2011), Pandurangan et al. (2017), Pires, Ascher e Blundell (2014b) and Pires, Ascher e Blundell (2014a)
Probability of damaging protein function	Tolerated or deleterious mutations that affects protein function and also pharmacophores calculations based on protein sequence	Sequence	Polyphen and SIFT	Adzhubei et al. (2010) and Ng e Henikoff (2001)
ΔPharmacophore	Pharmacophore difference based on protein sequence	Sequence	Biopython	Chapman e Chang (2000)

3.1.3.1 Residue Environment

Wild-type residue environment attributes are a set of conformational characteristics based on the structural data deposited into the PDB. To assist the calculation of features associated with the amino acid residue mutated in this category, such as type of secondary structure (α-helix, β-sheet, bend, turn, etc), relative solvent accessibility, residue depth, dihedral angles, and also a measure of the relative value for the isotropic b-factor among all the atoms on the wild-type residue of the mutation, the Biopython library was used. Biopython (CHAPMAN; CHANG, 2000) is a set of open source bioinformatics tools written in Python, an object-oriented scripting language, based on the highly successful Bioperl project (STAJICH, 2002).

The minimum distance from the mutated residue among all residues in the catalytic site of the molecule was calculated using the Catalytic Site Atlas (CSA) (PORTER; BARTLETT; THORNTON, 2004). CSA is a database documenting enzyme active sites and catalytic residues mapped to 3D structures. It defines a classification of catalytic residues which includes only those residues thought to be directly involved in some aspects of the of the reaction catalyzed by the enzyme. It contains 2 types of entries: original hand-annotated entries, derived from the primary literature and homologous entries, found by sequence comparison methods to one of the original entries. In case of no entry identified by CSA we added a default value of 30Å meaning that the residue is far away from any catalytic site, given that studies assessing the degree of conservation of residues close to the catalytic site have considered a maximum range of 12Å (BARTLETT et al., 2002), and also a study about the anatomy of enzyme have also shown that buried active site usually present channels, which helps access the site, with typical length $\geqslant$ 15Å (PRAVDA et al., 2014).

Lastly, the molecule flexibility feature was added to this category of attributes and defined as the entropy value calculated by ENCoM. This tool uses a coarse-grained normal mode analysis method that adds a layer information on its calculation which is the nature of amino acids for predicting the effect of a single point mutation on protein dynamics and thermostability resulting from vibrational entropy changes (FRAPPIER; CHARTIER; NAJMANOVICH, 2015).

3.1.3.2 Residue Interactions

The second category of features also rely on the 3D structure of the protein and provides information on the interatomic interactions that the mutated residue establish with other residues nearby, for instance hydrogen and covalent bonds, and hydrophobic interactions. Such attributes are calculated with Arpeggio which is an application and also a web server for calculating interactions within and between proteins and protein, DNA, or small-molecule ligands, including van der Waals', ionic, carbonyl, metal, hydrophobic, and halogen bond contacts, hydrogen bonds and specific atom-aromatic ring (e.g., cation-pi) (JUBB et al., 2017).

3.1.3.3 Structural Signatures

Structural signatures based on mCSM (PIRES; ASCHER; BLUNDELL, 2014b), described in subsection 1.3.2, were also clustered into a category of attributes. These were thought to be in a separated group due the great amount of signatures (784) generated for each mutation. These signatures used in this work can be divided into two major components. First of all, the graph based atom distance patterns using a minimum cutoff of 1Å and a maximum distance of 10Å with a step variation of 0.5Å. It also includes atom pharmacophoric changes between the wild-type and mutant residue residue using PMapper classification that comprises eight possible classes: hydrophobic, positive, negative, hydrogen acceptor, hydrogen donor, aromatic, sulphur and neutral.

3.1.3.4 Stability Changes Upon Mutations

Stability changes upon mutations is the last category that relies only on 3D protein structure data and provides information on the variation of stability on protein molecules caused by mutations that can lead to malfunction or even result on disease based on variation of Gibbs Free Energy ($\Delta\Delta$G) value. Here mCSM was also used, given the fact that it also produces $\Delta\Delta$G change predictions based on its signatures, alongside with the outputs of SDM and DUET that were described earlier as well.

3.1.3.5 Probability of Damaging Function and Pharmacophores

Finally, aiming to add a complementary group of features to amplify the search space of the supervised learning algorithms, described in following sections, the probability estimations of effects of mutations on protein function of SIFT and Polyphen were included as a different category. These are said to be complementary due the fact that they based their calculations on sequence information of proteins. In addition, each one of the 20 amino acid residues were represented by a vector with eight pharmacophores types: hydrophobic, positive, negative, hydrogen acceptor, hydrogen donor, aromatic, sulphur and neutral. These are known to be a collection of steric and electrostatic features required to ensure optimal interactions between groups of compounds and its biological target structure, according to the International Union of Pure and Applied Chemistry (IUPAC). For this work, wild-type and mutant residue pharmacophores are compared, and a vector with the difference between the two is also added in this category of attributes. These pharmacophores attributes are similar to the ones added to the graph signature of mCSM based in the descriptions of PMapper.

3.2 Machine Learning

3.2.1 Attribute selection

Given the high-dimensional feature space being investigated in this work, precautions were necessary in order to avoid the curse of the dimensionality and also model overfitting during supervised learning. These issues can have a negative effect not only on model generalization but also on performance during training, which could make the classification task unfeasible (ZAKI; MEIRA, 2014).

Principal Component Analysis (PCA) (ABDI; WILLIAMS, 2010) was performed over the data set of mutations to reduce the number of attributes from the category of structural signature features as an attempt to extract the most relevant and descriptive ones to be used alongside with the attributes of the other categories. This approach was implemented only on this category due to the great number of features in it (784). PCA is a mathematical algorithm that reduces the dimensionality of the data while maintaining the best variation in the data set. It does that by identifying directions, known as principal components, in which the variation in the data is maximal. Such direction is also the one that minimizes the mean squared error (JOLLIFFE, 2002).

An implementation of the algorithm execution is available in the Weka toolkit (HALL et al., 2009), and was used in this work. It is possible to use attributes from all groups detailed in Table 2 without compromising model scalability and generalization in the supervised learning step. Table 3 summarizes the number of attributes for each class of attributes that were used as evidence for training the supervised learning algorithms, described in section 3.2.2, after the reduction of the number of attributes within the structural signatures category. The total number of features is 86 in which 76 of these are structural-based and 10 sequence-based.

Table 3 – Number of attributes used as evidence for training supervised learning for each class of attributes after dimensionality reduction.

Category name	# Attributes
Wild-type residue environment	8
Wild-type residue interactions	15
Structural signatures	50
Stability change upon mutation	3
Probability of damaging protein function	2
ΔPharmacophores	8

3.2.2 Supervised Learning

Supervised learning is a machine learning task that aims to infer a function $f(x)$ from a set of labeled training data, also named training examples. The training data takes the form of a collection of (x, y) pairs, in which x is usually represented by a vector of features and y is the known output (or class) for a given x. The goal is to produce a prediction in response to a query with an unknown label, based on all the information extracted from the training step (JORDAN; MITCHELL, 2015). A variety of supervised learning algorithms has been proposed to estimate this type of mapping and this study will focus on four of them.

Multi-Layer Perceptron, also known as MLP, is a feed-forward neural network, consisting of a number of units, called neurons, which are connected by weighted links. The units are organized in several layers, namely an input layer, one or more hidden layers, and an output layer. The input layer receives an external activation vector, and forwards it via weighted connections to the units in the first hidden layer. These compute their activations and pass them to neurons in succeeding layers. From a distal point of view, an arbitrary input vector is propagated forward through the network, finally causing an activation vector in the output layer (RIEDMILLER, 1994). The entire network function, that maps the input vector onto the output vector is determined by the connection weights of the network.

Classification via regression handles the discrete classes (nominal) of the data set as continuous labels (probability) in a probabilistic classification manner (FRANK et al., 1998). The classification is achieved by defining a threshold, for example a prediction with a probability $\hat{y} < 0.5$ indicates non-activating and consequently $\hat{y} \geqslant 0.5$ results in activating output prediction, also known as linear decision boundary. Thus, algorithms that use this type of classification seeks for a model that generates the greatest approximate probability function that separates the classes in the dataset. In this sense, for the scope of this work, two algorithms were used Decision Trees M5P (KOTSIANTIS et al., 2007) and Gaussian Process (GP) with Radial Basis Functions (RBFs) (MACKAY, 1998).

A decision tree is an algorithm that simulates trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on the feature values. The basic assumption made in the decision trees is that instances with different classes have different values in at least one of their features. One of the most useful characteristics of such algorithm is their comprehensibility. One can easily understand why the algorithm classifies an instance as belonging to a specific class by just looking at the generated tree and analysing its rules (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2006). In this sense, the M5P algorithm uses model trees, which are binary decision trees with linear regression functions at the leaf nodes, that can represent any piecewise linear approximation to an unknown function. It usually builds and ordinary decision tree, using splitting criterion the maximization of the intra-subset variation of the target value and after that it prunes this tree back by replacing subtrees with regression functions wherever this seems appropriate (KOTSIANTIS et al., 2007).

Gaussian Process provides an alternative way of characterizing functions that does not require committing to a particular function class, but instead to the relation that different points on the function have to each other. It states that all uncertainty about any input variables, or combination of variables, is characterized by Gaussian distributions. GPs parameterize the probability in terms of a NxN covariance matrix, which is generated based on distance functions, also known as kernel, calculated for every pair or the N total observed points. In this work we used the RBF as kernel for the GP (RASMUSSEN; WILLIAMS, 2006; MACKAY, 1998).

The last classification algorithm used in this work was Random Forest (RF). This algorithm uses a combination of decision tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the "forest". The generalization error for forests converge to a limit as the number of trees in the forest become large (BREIMAN, 2001). It is a fast and relatively easy to implement algorithm, produce highly accurate predictions and can handle a very large number of input variables without overfitting, given that all the trees are built from scratch without any previous information on the other trees in the forest and also the final prediction is the average of all the predictions for each tree. In fact, it is considered to be on of the most accurate general-purpose learning techniques available.

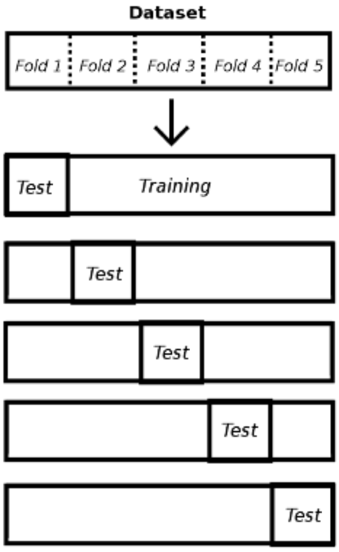
All the algorithms described previously are implemented and available through the Waikato Environment for Knowledge Analysis, also known as Weka, Toolkit (HALL et al., 2009) which allows researchers easy access to state-of-art techniques in machine learning such as the ones presented so far in this work. For the scope of this work the version 3.6 of Weka was used and for each algorithm the default parameters were preserved during training. In this sense, this work could also benefit from further investigation over parameter tunning for the supervised learning algorithms.

3.2.3 Cross-validation training

Evaluating classifiers induced by supervised learning algorithms is crucial for selecting the best performing a classifier, capable of generalization also reducing chances of overfitting from a given set and also to estimate its future prediction performance. The estimation method used in this work is k-fold cross-validation. In k-fold cross-validation the data set (D) is randomly split into k mutually exclusive subsets, known as folds, D_1,D_2,...,D_k of approximately equal size. The classifier is trained and tested k times; each time $t \in \{1, 2, ..., k\}$, it is trained on $\dfrac{D}{D_t}$ and tested on D_t. The cross-validation estimate of accuracy is the overall number of correct classifications, divided by the number of instances in the data set.

The value of k is usually chosen to be 5 or 10. There is a special case, when k is equal to the total number of instances in the data set, that is called leave-one-out cross-validation, where the testing set comprises a single point and the remaining data is used for training purposes (ZAKI; MEIRA, 2014). Figure 11, summarizes the k-fold cross-validation with a k value of 5. Despite its almost unbiased estimates it has high variance leading to unreliable estimates (EFRON, 1983). The results for each classifier training phase presented in further section are based on k-fold cross-validation with k=10.

Figure 11 – k-fold cross-validation with k=5.



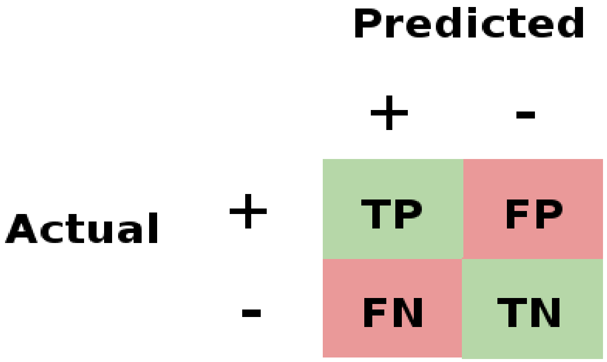
The data set is divided into 5 folds. The classifier is then trained and test 5 times varying its training and test subsets. The final result is the average of all the 5 train and test.

For the scope of this work, the supervised learning algorithms were trained with the full subset of training data, described in section 3.1.2, and the metrics to evaluate these algorithms during their training step were generated based on 10-fold cross-validation. In this sense, the model that is actually used for validation with the testing subset was built using the full training set.

3.2.4 Evaluation Metrics for Classification Algorithms

For model evaluation, four metrics were used due the fact that each metric presents its own limitations and a broader analysis of all of them together is better suited for evaluating the models described in this work. The metrics are precision, recall, f-measure (also known as f-score) and area under the ROC curve (AUC). These are well established and broadly used metrics for assessing the results of binary classification algorithms. Such measurements are expressed based on the values of a binary contingency table, also known as confusion matrix, where the classes are represented by convention with + (positive) and - (negative) signs. This 2x2 matrix (actual versus predicted class) uses the raw counts of the number of times each predicted label is associated with each real class. Figure 11, presents an example of confusion matrix.

Figure 12 – Confusion matrix (actual x predicted).



		Predicted	
		+	-
Actual	+	TP	FP
	-	FN	TN

True and False Positives (TP and FP) indicate the number of predicted positives that were correct and incorrect, respectively. Similarly, True and False Negatives (TN and FN) refer to correct and wrong predictions for negative class. The sum TP+FP+TN+FN is equal to the total amount number of instances in the data set being used.

Precision denotes the proportion of Predicted Positive cases that are Actual Positives. It is defined by $\frac{TP}{TP+FP}$. On the other hand, Recall is defined as the proportion of Predicted Positives cases that are Actual Positives over all Predicted Positives. Using the convention described in Figure 12, it is defined as $\frac{TP}{TP+FN}$. F-measure is a combination of Precision and Recall in a harmonic mean between them. This measure is defined by the square of the geometric mean divided by the arithmetic mean. All of these metrics present biases towards the predictions of positive class and ignore the performance in correctly predicting the negative class. This is particularly true for data with classes that are not balanced, such as the ones presented in this work (POWERS, 2011).

For a different perspective of analysis, given the bias problem with precision,recall and f-measure, the measure of Area Under the ROC Curve (AUC or AUROC) was also used. AUC considers the True Positive Rate (TPR), also known as sensitivity, that corresponds to the proportion of positive data points that are correctly considered as positive; and also the False Positive Rate (FPR) that corresponds to the proportion of negative data that are wrongly considered as positive, regarding all negative data points. A Receiver Operating Curve (ROC) is then plotted using TPR versus FPR and the AUC is the area under such curve (ZAKI; MEIRA, 2014). Like precision, recall and f-measure, AUC has its best result is 1 and the worse is 0. A random binary classifier would generate an AUC of 0.5.

3.3 Predictive Models (Structural, Sequence, Structural+Sequence)

Based on the classes of attributes and the supervised learning algorithms described previously, three different predictive models were generated as an attempt to evaluate the influence on each type of attributes on the final predictions. The first model uses only the attributes that rely on protein sequence information and is mainly comprised by PolyPhen-2 and SIFT predictions as well as pharmacophores calculations comprising 10 features. The second model uses only the attributes that were calculated based on structural data, such as relative solvent accessibility, graph based signatures and wild-type residue interactions (76 features in total). Finally, the third model was constructed based on a combination of all attributes, relying on structure and sequence data which represent the 86 features described in section 3.2.1. The results are present in the following section.

4 Results and Discussions

4.1 Data Summary of Mutations Sample

The performance of a classifier is limited to the descriptive features used as evidence during the training and testing processes. In this sense, one of the most important steps in machine learning is identifying a set of predictive attributes to be used to train a model. For the scope of this work, since it is dealing with binary classification (whether a mutation is activating or non-activating), the attributes selected should help the classifier to differentiate between the two classes. Figures 13 to 16 display the summary of attributes distribution for each of the classes in the data set of mutations using boxplot representation. Since the classes are not equally balanced in the data set, t-test is applied with a 95% confidence interval for comparison (JAIN, 2015). Attributes with statistically significant differences between the two classes are marked with a red asterisk (*).

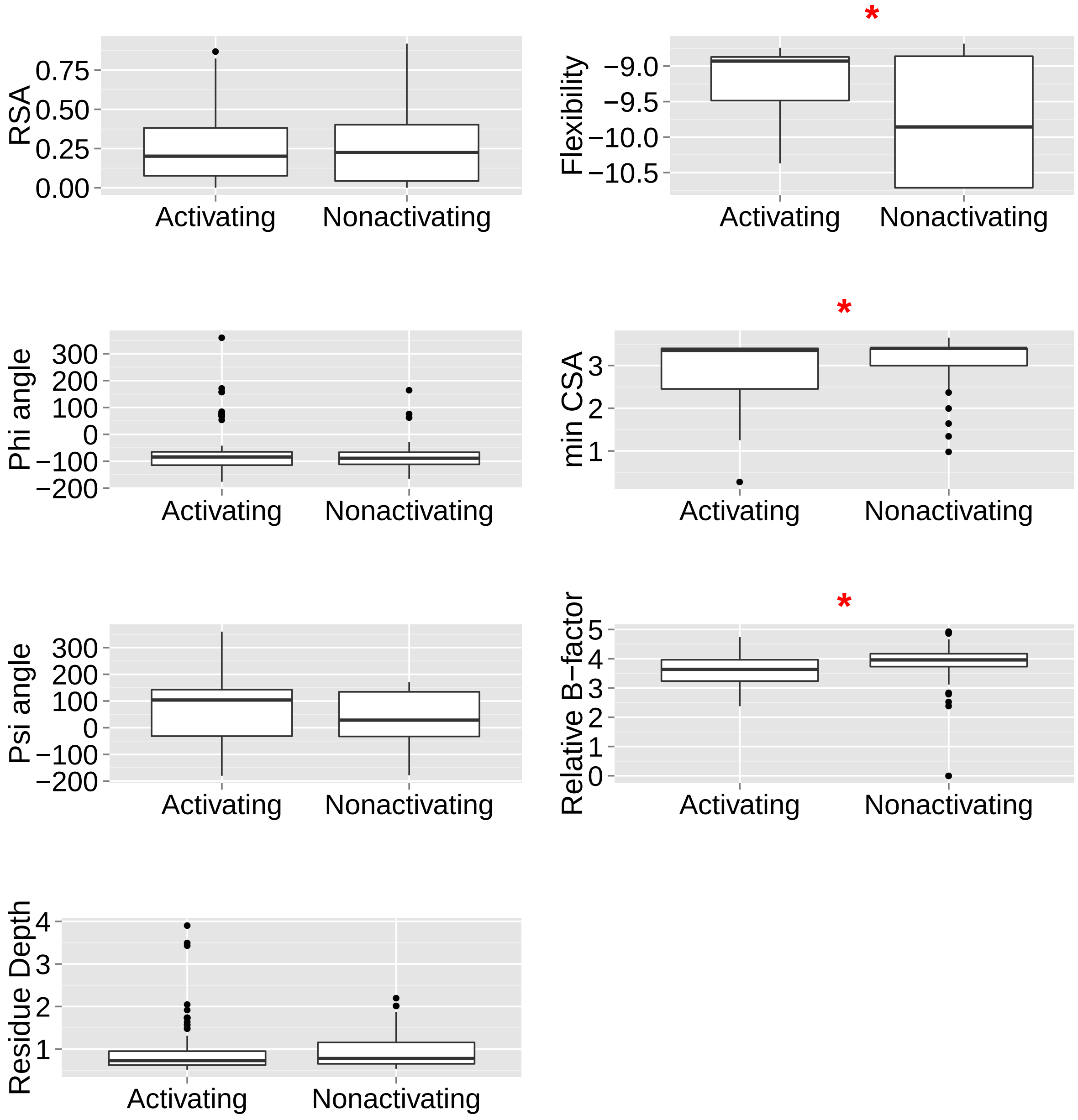
Protein kinase transition state from active to inactive and vice-versa require that the protein presents a minimum flexibility no matter if the kinase is activated or not. The mechanisms by which the activating mutations affect kinases are associated with a restriction in the transition from active to inactive, resulting in one conformational state being favoured (WAN et al., 2004; SUTTO; GERVASIO, 2013). This transitional state is directly affected by the molecule flexibility and here we address this with two features: Entropy energy prediction and relative B-factor of the 3D structure.

Entropy energy predictions calculated by ENCoM show significantly differences for the two classes of mutations. Wild-type 3D structures of proteins with evidence of activating mutations have, in general, higher values of entropy and are consequently more flexible than the structures with evidence of non-activating mutations (Figure 13).

On the other hand, the measure of relative B-factor indicates the relative vibrational motion of the structure of the protein as whole. This value is calculated based on an simple mean of the b-factor of all atoms within the PDB structure. Each atom has a b-factor value which was experimentally measured during the process of refinement of the structure and shows the amplitude of oscillation of each atom (WLODAWER et al., 2008). Higher values of this feature imply more flexible structures.

Even though, both attributes describe similar characteristic of the molecule and show significantly difference between the two classes of mutations, the results for the relative B-factor points into a completely opposite direction when compared to entropy energy predictions of ENCoM. This may be due to the set of extra attributes used by ENCoM algorithms, such as the nature of the amino acids in the structure, since it also

Figure 13 – Boxplots comparing the distribution of values on the classes activating and non-activating for Wild-type residue environment attributes.



A red asterisk denotes a significantly difference between mutation types (p-value < 0.05) assessed via a t-test. In this case ENCoM entropy, relative B-factor and minimum distance to catalytic sites calculated by CSA presented a significant difference.

uses the b-factor information within the structure of the protein in its prediction. Thus some non-explicit relationship between b-factor and other attributes might be inversing the relationship observed for both features in this work. Further investigation of this relationship would benefit this work in future implementations.

Surprisingly, none of the stability changes predicted by the three methods discussed here (mCSM, SDM and DUET) presented significant differences between classes. The average predictions of both classes was slightly negative, indicating that the mutations would likely lead to mild destabilisation of the protein structure. While not statistically significant, in part due to the limited data for analyses, the activating mutations tended to have more negative predicted changes in the Gibbs Free Energy of folding and stability, suggesting that these mutations were more likely to lead to disruption of the local structure. This is the opposite of what was expected given that recent studies with the EGFR kinase have shown that activating mutations are more likely to lead to a more significant variation on stability when compared to neutral mutations in which no apparent change was observed (SUTTO; GERVASIO, 2013).

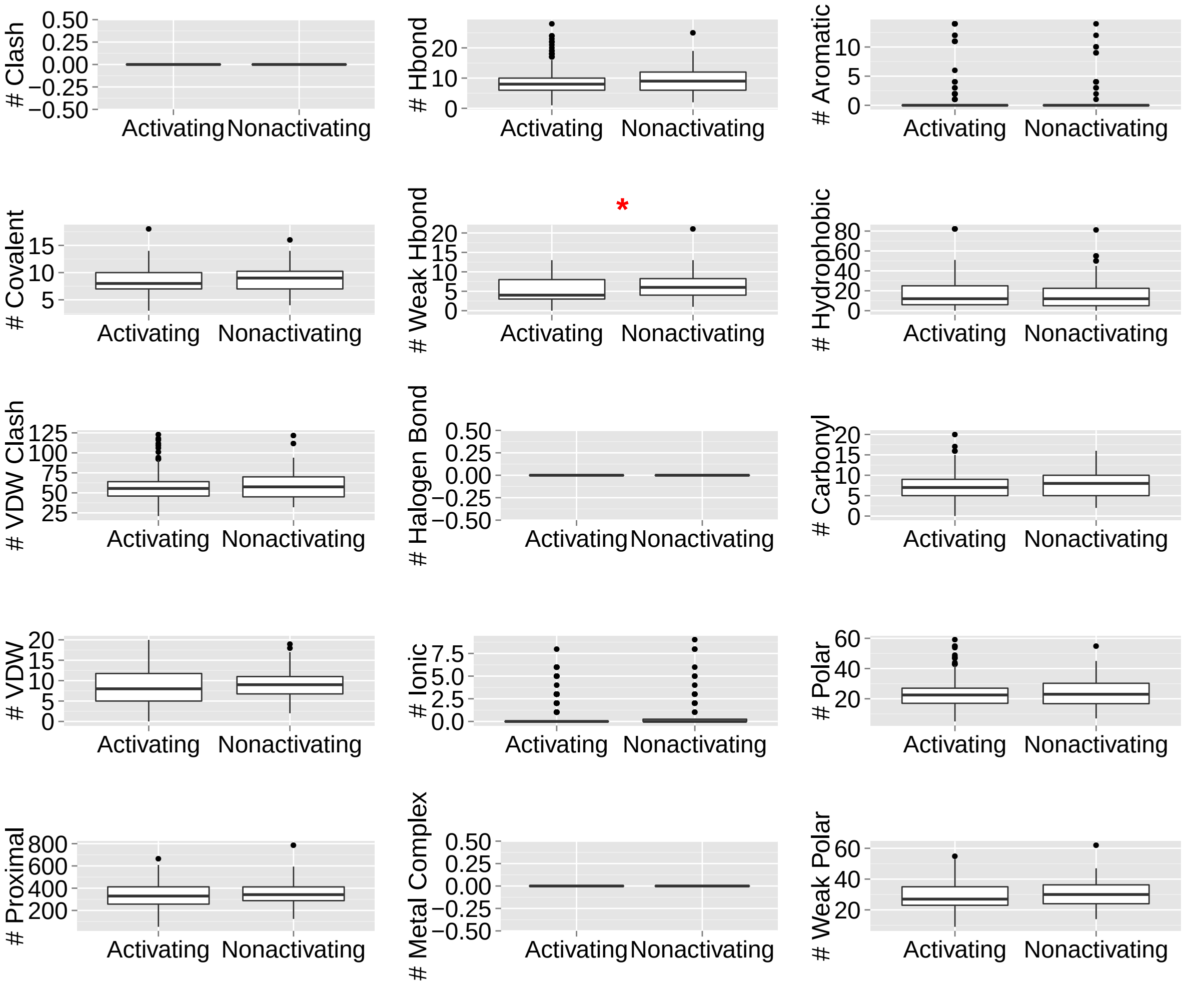
Regarding the distance to the catalytic site of the protein, both classes of mutations occur close to the catalytic site on average (3Å or less), however activating mutations are more densely distributed closer to the catalytic site, while non-activating ones values are more condensed above the distance of 3Å. Both results are expected due the fact that some studies (WAN et al., 2004; SUTTO; GERVASIO, 2013; BOSE et al., 2013) indicate that mutations that somewhat alter the function of this type of protein and may also confer resistance to kinase inhibitors occur in regions that are related or close to the catalytic site of kinases. Such mutations can help the protein to be stabilized in its active state conformations by making them more rigid.

For the interatomic interactions of the residue in the wild-type structure (Figure 14, only the number of weak hydrogen bonds in which the residue participates presented a significant difference for protein structures of both classes of mutations. This type of interaction is important for the stability of the protein during the folding process alongside with hydrophobic and Van der Walls for example, but none of them show any explicit difference.

Figure 15 depicts the value distributions of graph-based signature attributes generated with mCSM and selected with PCA. In this case, only the distance pattern represented by atoms labelled as negative associated with sulphur atoms closer than 7.5Å, showed significant difference. The distribution for both classes are centered in zero with a few outliers that could be the reason why this specific signature presented such explicit difference.

According to Figure 16, activating mutations are more likely to be classified as pathogenic, which is expected due the bias problem towards that class already mentioned in previous section, given that the distribution of values is closer to 1. Regarding non-

Figure 14 – Boxplots comparing the distribution of values on the classes activating and non-activating for wild-type residue interactions attributes.

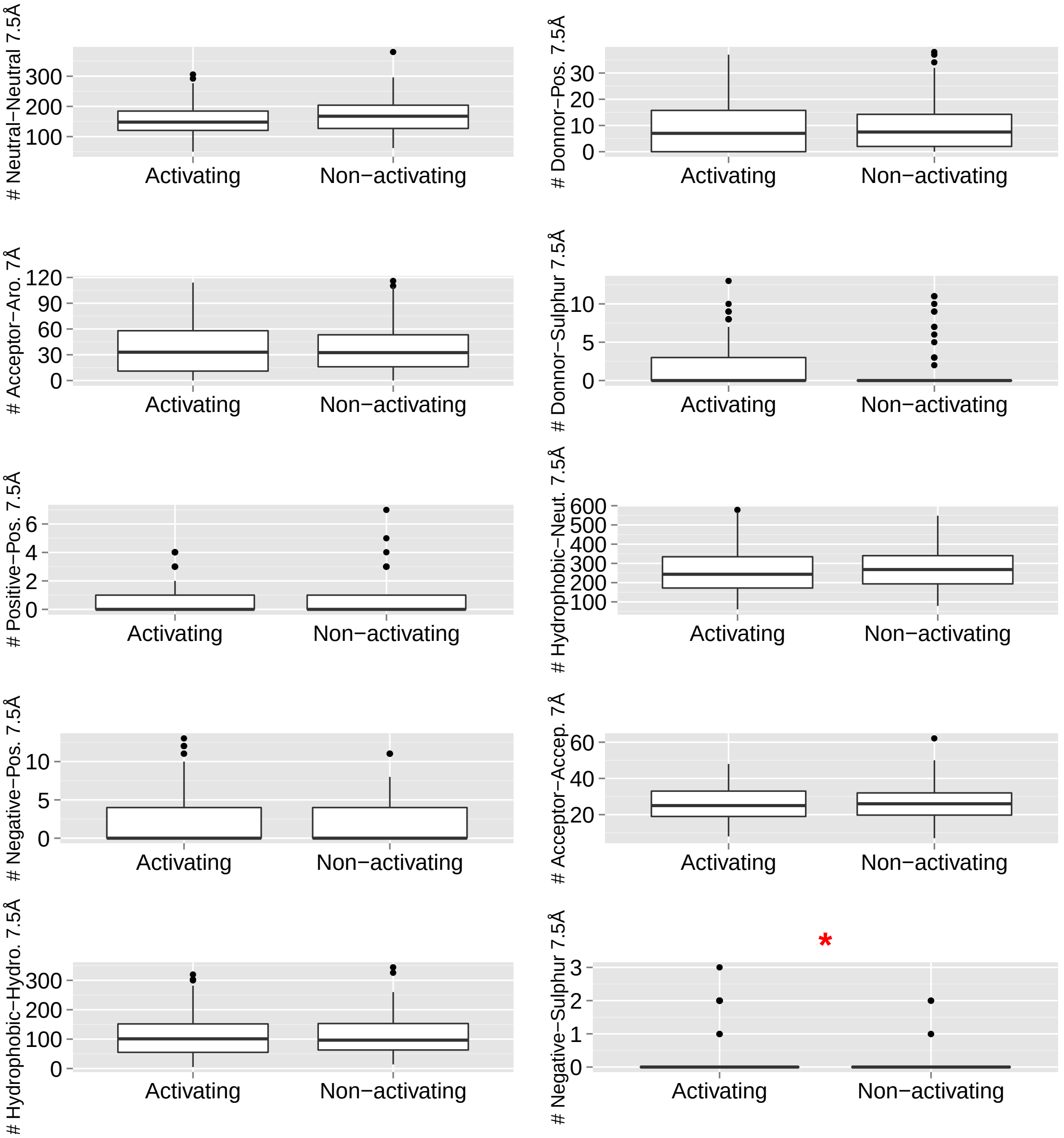


A red asterisk denotes a significantly difference between mutation types (p-value < 0.05) assessed via a t-test. In this case only the number of weak hydrogen bonds presented a significant difference.

activating mutations, the mean is as close as the mean for the activating class, but the values are more spread towards zero, which is also expected given that non-activating mutations, as discussed in section 3.2.1, are in this study a combination of inactivating and neutral mutations. In other words, some of the non-activating mutations can indeed be pathogenic if they act as tumor suppressors for example, and some of them can also be classified as benign by PolyPhen-2 (scores closer to 0) given that they are neutral mutations.

Notably, SIFT did not show the same difference even though both methods rely mainly on the same type of data. This can be explained due the extra 3 structural attributes used by PolyPhen-2, when available, as evidence for its model. In this case both classes presented a distribution concentrated close to zero indicating that these are mutations that

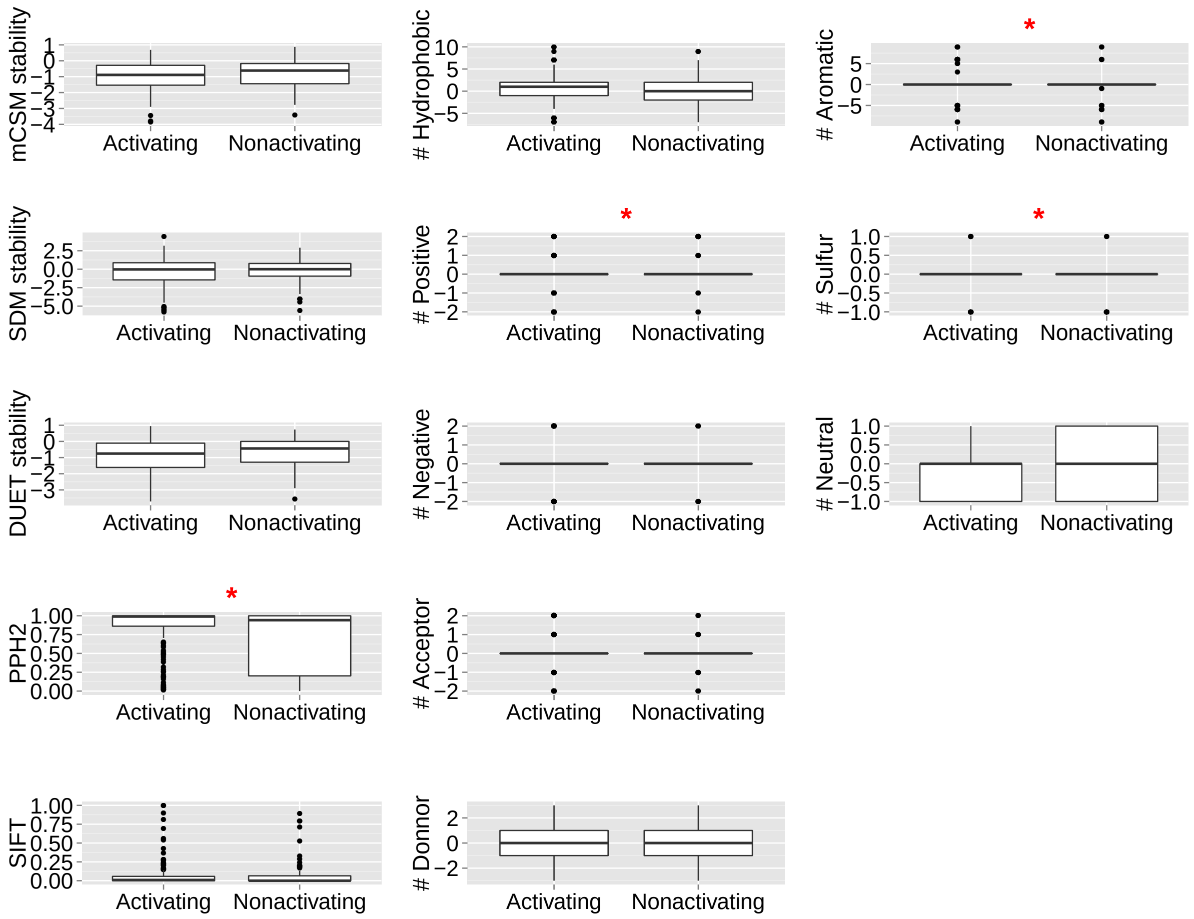
Figure 15 – Boxplots comparing the distribution of values on the classes activating and non-activating for structural signatures attributes.



A red asterisk denotes a significantly difference between mutation types (p-value < 0.05) assessed via a t-test. In this case only one of the signatures in this image presented a significant difference. Only 10 signatures are displayed in this figure due to visual purposes.

might affect protein function and consequently would not be tolerated. Alongside with PolyPhen-2 and SIFT the features representing the pharmacophore difference between the wild-type residue and the mutated residue are also summarized in Figure 16. For this class of features the hydrogen bond donor and positive pharmacophores presented a significant difference between mutation classes. However, for the positive pharmacophore the distribution is very scattered and no apparent information can be inferred, and the difference identified by the t-test might be an artifact. In the case of hydrogen bond donor, for both classes, the values are distributed over zero with mutations in which there is a gain or loss of less than two hydrogen bond donor for both classes. Overall, there is no substantial difference for the other pharmacophores.

Figure 16 – Boxplots comparing the distribution of values on the classes activating and non-activating for stability change upon mutations and also probability of damaging protein function attributes.



A red asterisk denotes a significantly difference between mutation types (p-value < 0.05) assessed via a t-test. In this case PolyPhen-2 probability and 3 pharmacophores presented a significant difference.

From the boxplots showed above, many features seem to be "useful" for classification task, specially the ones that present significant differences statistically, but none of them perfectly separates the classes in the data set. This assumption indicates that multiple features should be used as evidence for the machine learning algorithms described in section 3.2.

Most features present no significant difference between the two classes indicating that they do not directly correlate to the classes in the data set. However, providing different types of data to algorithms of classification can have a positive impact in their results due to the fact that distinct features capture more diverse types of information. Moreover, during the process of learning, features are combined and transformed as an attempt to identify implicit and non-trivial correlations. None of the features were removed from the selected features before training the supervised learning algorithms, even though some of them do not show apparent significant distinction between the two classes.

4.2 Model Assessment

The four supervised learning algorithms discussed in section 3.2.2 were evaluated for each type of features with the sample set of mutations described in section 3.1. The assessment was based on the four metrics of evaluation: precision, recall, f-measure and AUC. The results are evaluated separately for each type of data and a final section was dedicated to comparisons among the best performing results and also with the well established methods: SIFT and PolyPhen-2

4.2.1 Sequence-Based

For the models resulted from training the supervised learning algorithms with features calculated based on the sequence data available for proteins, the Random Forest algorithm presented the best results for most of the metrics used for evaluation. The AUC of 0.77, for instance, outperforms MLP and the algorithms that use classification via regression, Decision Tree with M5P algorithm and Gaussian Process, which achieved AUCs of 0.61, 0.69 and 0.64, respectively. The same is observed for precision and F-measure for both classes. The results are summarized in Table 4.

The algorithms trained and evaluated on 10-fold cross-validation with only sequence-based features as evidence, in general, struggle when dealing with mutations that are instances of the class non-activating, similar behaviour from what was discussed in section 1.3.2 for SIFT and PolyPhen-2. For example a precision of only 44% for Multi-Layer Perceptron algorithm and a Recall of 20% for classification via regression with Gaussian Process were achieved. Feature selection did not show any apparent improvement of the performance for this type of data, which indicates that more information is necessary for the algorithms to improve performance.

Table 4 – Results for all classifiers trained with sequence-based features in each one of the mutation classes.

Classifier	Precision	Recall	F-Measure	AUC	Class
MLP	0,730	0,758	0,744	0,608	Activating
MLP	0,443	0,407	0,424	0,608	Nonactivating
M5P	0,754	0,857	0,812	0,697	Activating
M5P	0,643	0,407	0,429	0,697	Nonactivating
Gaussian Process	0,679	0,833	0,748	0,636	Activating
Gaussian Process	0,367	0,196	0,256	0,636	Nonactivating
Random Forest	**0,775**	**0,877**	**0,823**	**0,769**	**Activating**
Random Forest	**0,659**	**0,482**	**0,557**	**0,769**	**Nonactivating**

Best performing model is highlighted.

Blind tests were performed to further validate the train models and assess their generalization ability. The results, summarized in Table 5, show that Random Forest has the best performance when classifying activating mutations. Even though, it does not have the best score for AUC, 0.70 achieved by classification via regression with Gaussian Process, the Random Forest algorithm presented better Recall and Precision for non-activating mutations, being a more balanced predictor.

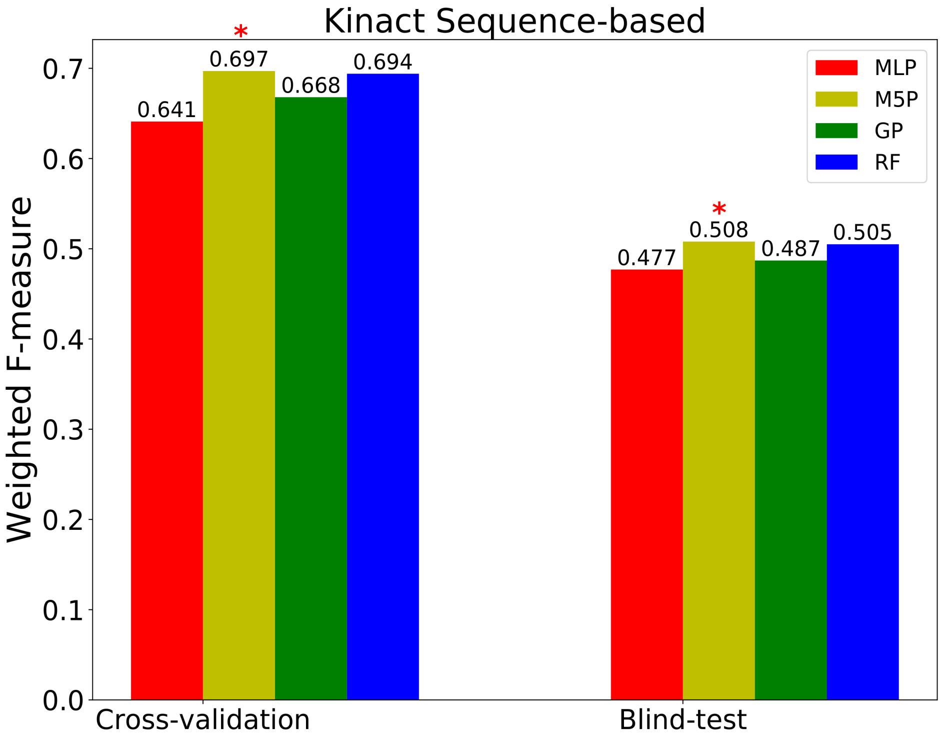
Table 5 – Results for all classifiers tested with sequence-based features in each one of the classes.

Classifier	Precision	Recall	F-Measure	AUC	Class
MLP	0,706	0,681	0,684	0,617	Activating
MLP	0,462	0,500	0,480	0,617	Nonactivating
M5P	0,737	0,875	0,800	0,560	Activating
M5P	0,333	0,167	0,222	0,560	Nonactivating
Gaussian Process	0,763	0,706	0,729	0,633	Activating
Gaussian Process	0,500	0,205	0,333	0,633	Nonactivating
Random Forest	**0,771**	**0,844**	**0,806**	**0,668**	**Activating**
Random Forest	**0,444**	**0,333**	**0,381**	**0,668**	**Nonactivating**

Best performing model is highlighted.

Again, the unbalanced characteristic of the data set might be the main reason for lower scores, specially for the less frequent class (non-activating). This is also observed for the weighted average of F-measure for all classifiers (Figure 17). Even though, Classification via regression using M5P and Random Forest performs slightly better than the other classifiers on training (70% for both) and also on blind-test (51% and 50%, respectively), all algorithms performed similarly regarding this metric.

Figure 17 – Comparison of weighted average for F-measure of all algorithms that used only sequence-based features as evidence.



Classification via regression with M5P algorithm has the best performance in training (left) and also on Blind-test (right), achieving a score of 70% and 51%, respectively, and it is closely followed by Random Forest with also 70% on training and 50% on blind-test. However, all algorithms present similar F-measure values, mostly due the poor performance when dealing with instances of non-activating class. A red asterisks identifies the best performing algorithm.

4.2.2 Structure-Based

The algorithms trained on 3D structural-based features performed in general better than the ones that used only sequence-based data. The results over training are compiled in Table 6 and show that classification via regression of Decision Trees with the M5P algorithm presented the best results for all the four metrics for both classes. The lowest score for this algorithm was a Recall of 72% for non-activating class, even though it confers a significantly increase in comparison with the sequence-based predictor. With this set of features, the least balanced predictor was the Classification via regression using Gaussian Process that produced Recall and F-measure of only 23% and 38%, respectively.

Table 6 – Results for all classifiers trained with training set of mutations that had their region mapped to PDB with structure-based features.

Classifier	Precision	Recall	F-Measure	AUC	Class
MLP	0,861	0,929	0,894	0,853	Activating
MLP	0,771	0,617	0,685	0,853	Nonactivating
M5P	**0,899**	**0,987**	**0,941**	**0,895**	**Activating**
M5P	**0,956**	**0,717**	**0,819**	**0,895**	**Nonactivating**
Gaussian Process	0,769	0,994	0,867	0,836	Activating
Gaussian Process	0,933	0,233	0,373	0,836	Nonactivating
Random Forest	0,876	0,968	0,920	0,868	Activating
Random Forest	0,806	0,550	0,700	0,868	Nonactivating

Results are presented for both classes of mutations. Best performing model is highlighted.

Aiming to validate the training step described in this section, the subset for blind-test (described in subsection 3.1.2) was used for further testing the four trained models. Table 7 presents the results for all the classification algorithms validated with the test set using structural features. The best results are highlighted in bold.

Table 7 – Results for all classifiers validated with test set with only structure-based features.

Classifier	Precision	Recall	F-Measure	AUC	Class
MLP	0,806	0,781	0,794	0,615	Activating
MLP	0,462	0,500	0,480	0,615	Nonactivating
M5P	**0,810**	**0,875**	**0,836**	**0,704**	**Activating**
M5P	**0,667**	**0,417**	**0,541**	**0,704**	**Nonactivating**
Gaussian Process	0,730	0,844	0,783	0,464	Activating
Gaussian Process	0,286	0,167	0,211	0,464	Nonactivating
Random Forest	0,769	0,893	0,826	0,619	Activating
Random Forest	0,538	0,318	0,400	0,619	Nonactivating

Results are presented for both classes of mutations in each one of the classes. Best performing model is highlighted.

Based on the achieved AUC values, classification via regression with M5P algorithm was the best performing model, closely followed by MLP with 0.70 and 0.61, respectively. Random Forest is ranked only as the third best performance with lower scores for Recall and F-Measure for the non-activating mutations. Even though, Random Forest presented the best AUC, the other metrics presented lower values than the other two predictors (MLP and classification via regression with M5P) suggesting that a higher AUC in Random Forest is very influenced by how well it performed on the instances of the Activating class. Between MLP and Classification via regression with M5P, the metrics have a slightly variation, but overall MLP performs better.

However, due the fact that Decision Trees with M5P performed better in training and has comparable performance with the best results in blind-test, it remains the best candidate for best model.

The greater amount and diversity of features for the sample set of mutations with structural data provided to the algorithms corroborates the better performance of models trained with this type of data. There are 4 different classes of attributes that rely on structural data with 76 attributes against 2 different classes of attributes sequence based with 10 attributes (Tables 2 and 3 in section 3.1.3).

Complementary analysis of the weighted average F-measure, for all algorithms using only structural-based features as evidence, confirm M5P as the best candidate for model generation (Figure 18). Classification via regression with M5P algorithm is the best performing among all the others on traninng and also blind-test with 89% and 77%, respectively. MLP also presented equivalent score on training (86%), but on blind-test the performance drops to 62%.

4.2.3 Structural + Sequence Based

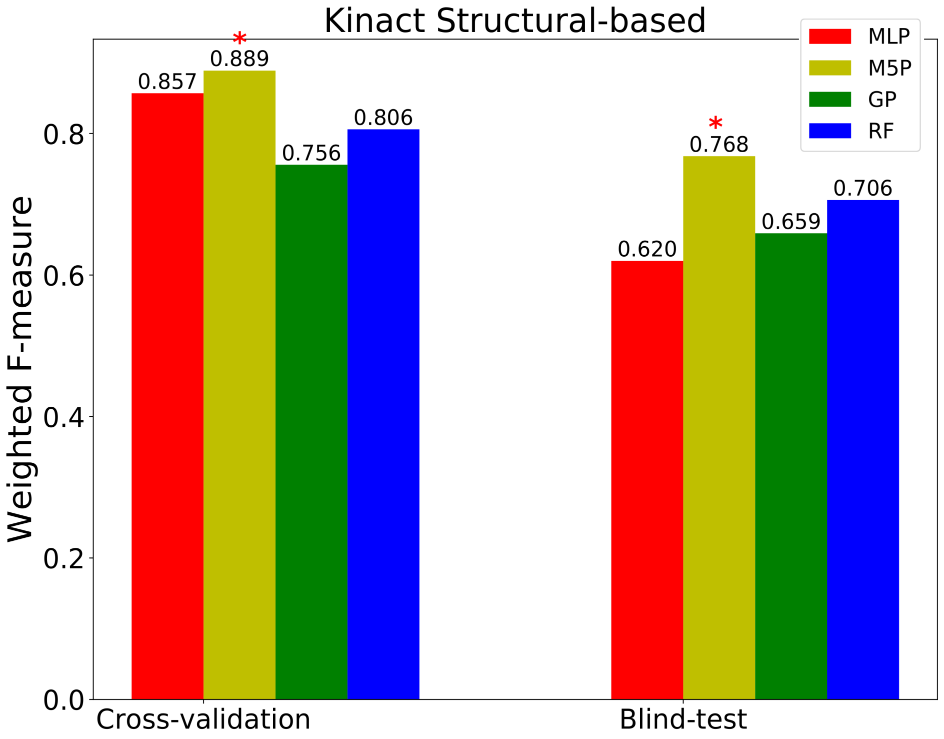
In this section the performance of the classification algorithms is evaluated when using both types of data from the sample set of mutations, structural and sequence based features, as evidence for training and testing. The results for the predictors using the two sets of features on training are summarized on Table 8.

Classification via regression with M5P algorithm showed the best performance in comparison with the remaining algorithms. The results are similar, but slightly better than those observed when using structural data only (depicted in Table 4). When evaluating the metrics for non-activating class, there is still a slightly worse performance than for Activating class. However, this is expected due the characteristics of the data set of mutations extensively described throughout this work.

Table 9 presents the results for all the classification algorithms validated with the blind-test set described in section 3.1.3, which comprises structural and sequence data from the mutations sample set. Best results are highlighted.

Again, classification via regression with M5P algorithm presented the best performance (AUC:0.89) on blind-test with highest values for all metrics when handling instances

Figure 18 – Comparison of weighted average for F-measure of all algorithms that used only structural-based features as evidence.



Classification via regression with M5P algorithm has the best performance in training (left) and also on Blind-test (right), achieving a score of 89% and 77%, respectively. All algorithms presented values above 75% on training. However, the performance drop to below 66% for MLP and classification via regression with Gaussian Process.

Table 8 – Results for all classifiers trained with the training test of mutations with structural and sequence-based features.

Classifier	Precision	Recall	F-Measure	AUC	Class
MLP	0,864	0,887	0,875	0,835	Activating
MLP	0,687	0,639	0,662	0,835	Nonactivating
M5P	**0,908**	**0,985**	**0,945**	**0,964**	**Activating**
M5P	**0,949**	**0,736**	**0,811**	**0,964**	**Nonactivating**
Gaussian Process	0,769	0,994	0,867	0,836	Activating
Gaussian Process	0,933	0,233	0,373	0,836	Nonactivating
Random Forest	0,876	0,968	0,920	0,888	Activating
Random Forest	0,886	0,650	0,750	0,888	Nonactivating

Results are presented for both classes of mutations. Best performing model is highlighted.

Table 9 – Results for all classifiers tested with structural data in each one of the classes

Classifier	Precision	Recall	F-Measure	AUC	Class
MLP	0,774	0,732	0,752	0,603	Activating
MLP	0,400	0,445	0,426	0,603	Nonactivating
M5P	**0,810**	**0,911**	**0,857**	**0,884**	**Activating**
M5P	**0,667**	**0,583**	**0,609**	**0,884**	**Nonactivating**
Gaussian Process	0,720	0,964	0,824	0,504	Activating
Gaussian Process	0,333	0,145	0,100	0,504	Nonactivating
Random Forest	0,800	0,857	0,828	0,655	Activating
Random Forest	0,556	0,455	0,500	0,655	Nonactivating

Best result is highlighted

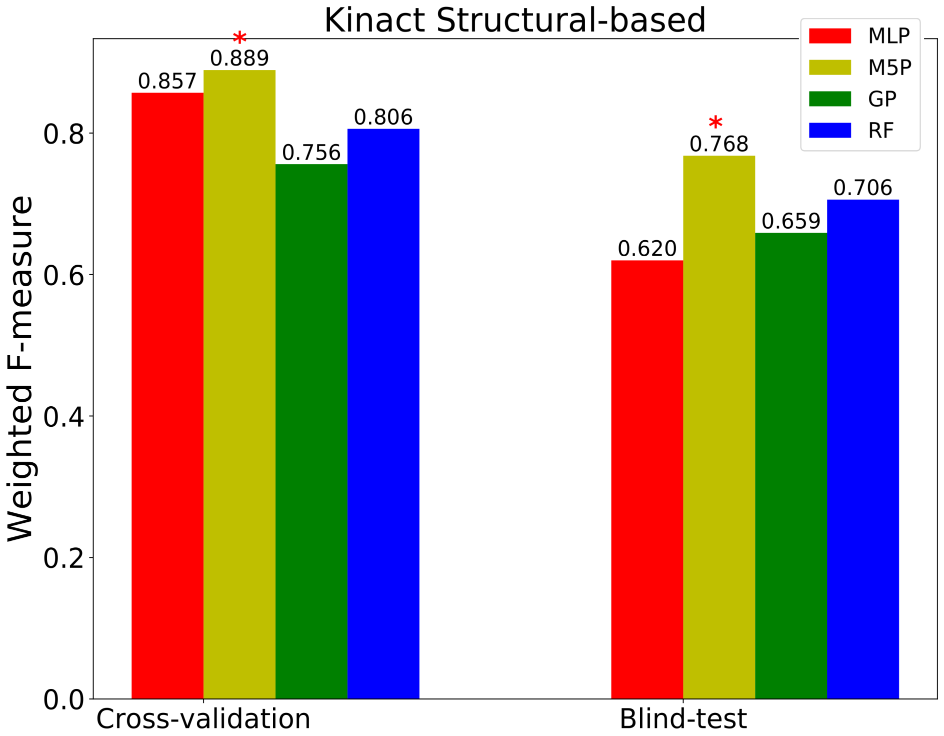
of non-activating class. Classification via regression using Gaussian Process, on the other hand, had the worst performance with Recall of 15% and AUC of 0.50. Comparison of the weighted average for F-measure for all classifiers corroborates for the choosing of M5P as the best candidate for further model generation (Figure). M5P performs better than all the other classifiers on training and also blind test, with 91% and 77%, respectively.

4.2.4 Performance Comparisons

For comparison purposes and given that AUC is the most unbiased evaluation metric, as discussed in section 3.2.4, the ROC curve and AUC value for the best classifier in each type of data were analyzed. The results are shown in Figure 16. On training, Kinact has its worst performance when using only sequence-based data (AUC: 0.77). The performance presents a significant improvement, with a p-value < 0.01 (HANLEY; MCNEIL, 1982), when using only structural-based data on mutations. One of the reasons for such difference is the greater number of features used for training the algorithms when using only sequence-based features, providing a certain heterogeneity of the input data that contributes for better predictions as discussed in section 4.1. Nevertheless, the best performance (AUC: 0.97) is obtained when both types of data are used as input and it is significantly different from the performances for the predictors that used only sequence-based or structural-based data (p-value < 0.01).

On blind-tests, similar behaviour is observed and Kinact also has its best when both types of data are used (AUC: 0.89) followed by the version that uses only structural-based data (AUC: 0.70) and the one that uses only sequence-based features (AUC: 0.66).

Figure 19 – Comparison of weighted average for F-measure of all algorithms that used both type of attributes (structural and sequence-based features) as evidence.



Classification via regression with M5P algorithm has the best performance in training (left) and also on Blind-test (right), achieving a score of 91% and 77%, respectively. All algorithms presented values slightly better than the F-measures observed for training and testing with only structural-based features, except for classification via regression with Gaussian Process that present even lower values for this metric.

4.3 Web Server - Kinact

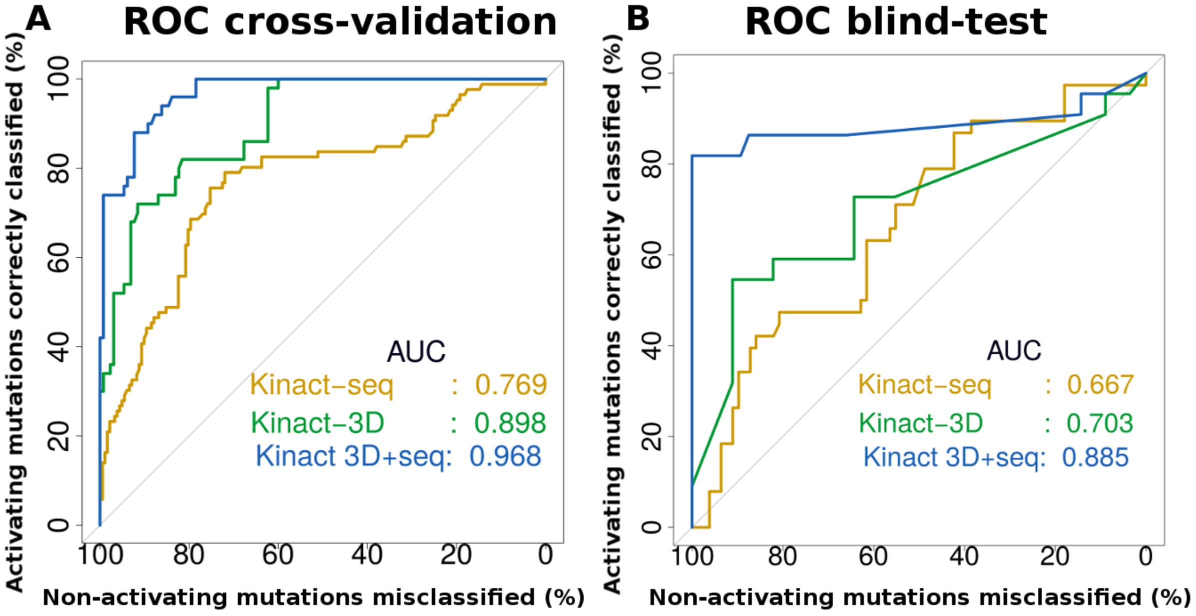
A web server was developed based on the best models obtained for predicting the activating missense mutations in protein kinases.

Kinact was implemented via a user-friendly web server freely available[1]. All resources used for building Kinact are open source and only a few are discussed here for the sake of simplicity. For a complete list of dependencies used to build this web server check Appendix B or visit [2].

The server front-end was built using Bootstrap-3.3.7 [3] which is one of the most popular front-end frameworks and open source projects. Such frameworks make it easier to create client-side design by providing predefined CSS classes, each of which indicates the width of the column you want to create, the type of element you are using, or the color and style you want to use. On the server side, the back-end was built in Python

[1] <http://biosig.unimelb.edu.au/kinact>
[2] <http://biosig.unimelb.edu.au/kinact/components>
[3] <http://getbootstrap.com/>

Figure 20 – ROC curves for comparison of the three versions of Kinact (sequence-based, structure-based and sequence+structure-based).



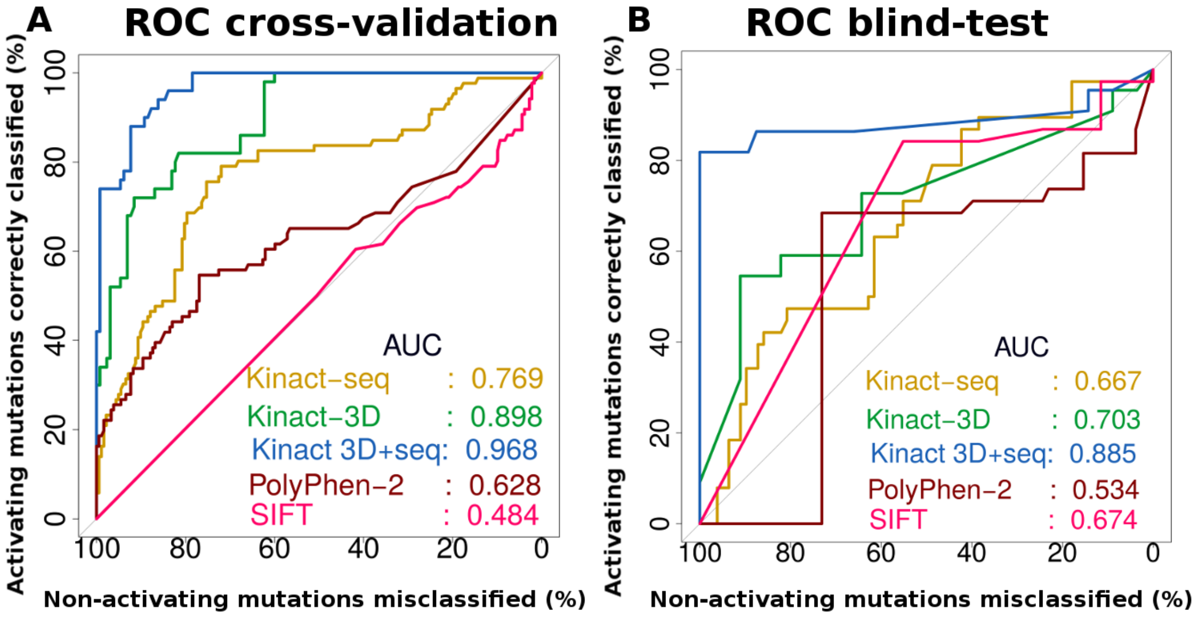
A) shows the ROC curves for the three versions of Kinact: using only sequence-based features (yellow), only structural-based features (green) and using both type of features (blue) on training. The best performance is observed for the predictor that uses both types of data (AUC: 0.97) followed by the one that uses only structural-based features (AUC: 0.90) and the one that uses only sequence-based data. B) summarizes the results on the blind-test set of mutations. The predictors are ranked similarly, the best performance is obtained when using both types of features (AUC: 0.89). The predictor that uses only structural-based features (AUC: 0.70) still outperforms the one that uses only sequence-based features (AUC: 0.67).

via the Flask framework. Python[4] is a general-purpose, high-level programming language whose philosophy emphasizes code readability. Python's syntax allows programmers to express concepts in fewer lines of code when compared with languages such as C and Java (MARTELLI; RAVENSCROFT; ASCHER, 2005). Flask[5] is a small framework by most standards, small enough to be also known as a "micro-framework" based on Python. It was designed as an extensible framework providing a solid core with the basic services, while extensions provide the rest. Such flexibility, allows developers to have a lean stack that has no bloat with only what is necessary for the application to work properly, in contrast with larger frameworks, where most choices have already been made and are, usually, hard or sometimes impossible to customize (GRINBERG, 2014). The application is hosted on a Linux server running Apache. Figure 22 shows a screenshot of the home page of Kinact.

[4] <https://www.python.org>

[5] <http://flask.pocoo.org/>

Figure 21 – ROC curve for comparison of Kinact best predictor with SIFT and PolyPhen-2.



A) shows the ROC curves for Kinact best predictor (blue), PolyPhen-2 (red) and SIFT (pink) on training. Kinact (AUC: 0.97) significantly outperformed both SIFT (AUC: 0.49) and PolyPhen-2 (AUC: 0.63). B) summarizes the results on the blind-test set of mutations. SIFT (AUC: 0.67) performs better than PolyPhen-2 (AUC:0.53), however Kinact best predictor was again the best performing method (AUC: 0.90). Kinact version that use only sequence-based data is shown (yellow) as well as the version that uses only structural-based data (green) for comparison purposes.

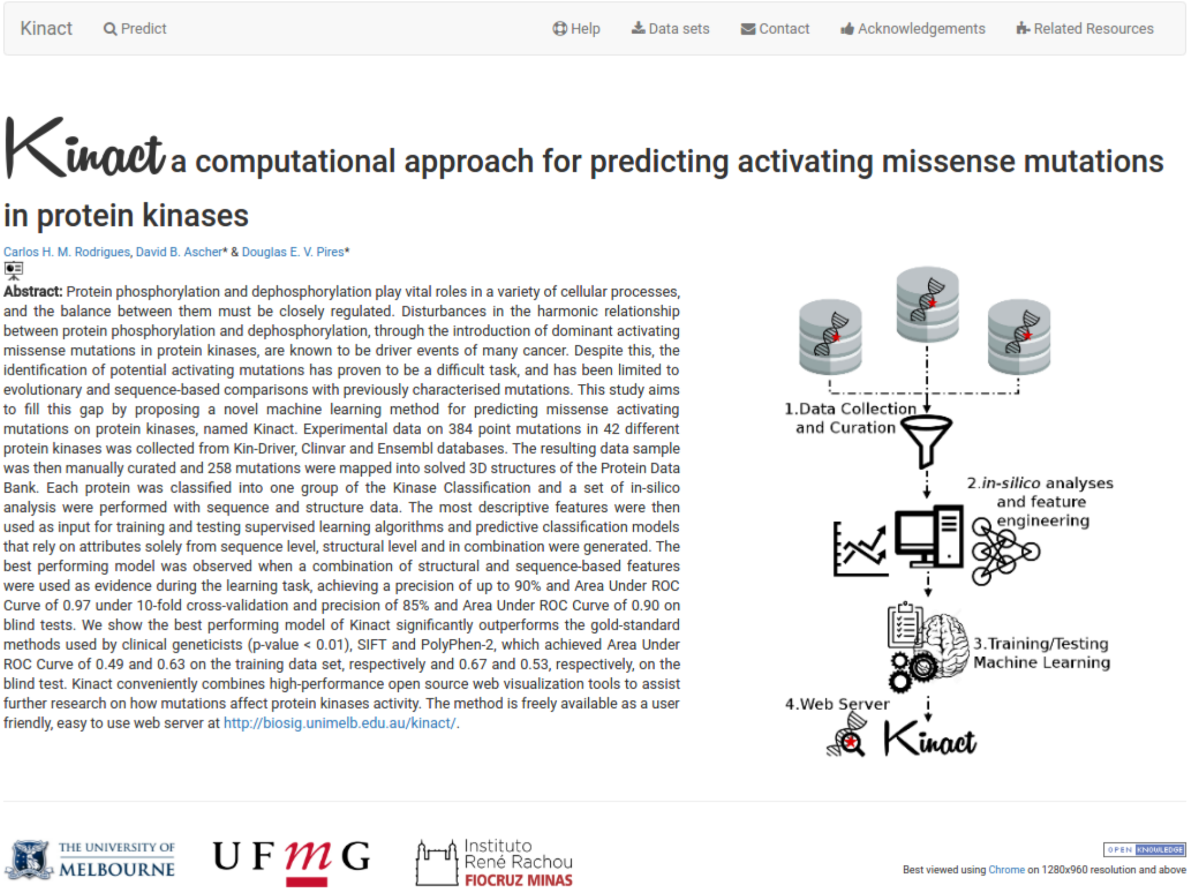
4.3.1 Input

Kinact provides two different input options for users, as shown in Figure 23. The “Single mutation” option allows users to predict whether a single mutation in a kinase is Activating. The information required includes a PDB file and also the sequence in fasta format of the protein, alongside with the mutation code specified as a string consisting of a single letter code of the wild-type residue in the protein, its corresponding residue number and the single letter code of the mutant residue.

Alternatively, the “Mutation list” option allows users to submit a file with a list of mutations and chain identifiers (similar to described previously for single mutation prediction) to be evaluated for a specific PDB file and/or sequence that are also required as input.

Given that the best model generated by this work relies on the combination of both types of attributes (structural and sequence based), the users are encouraged to provide both sequence and structural information so that the prediction of Kinact uses all features necessary for a more reliable outcome prediction. However, if only the PDB structure or only the sequence is provided Kinact still runs the prediction with only the

Figure 22 – Kinact home page


Kinact
Predict
Help
Data sets
Contact
Acknowledgements
Related Resources
1.Data Collection and Curation
2.*in-silico* analyses and feature engineering
3.Training/Testing Machine Learning
4.Web Server
Kinact
THE UNIVERSITY OF MELBOURNE
UFMG
Instituto René Rachou FIOCRUZ MINAS
OPEN KNOWLEDGE

Kinact a computational approach for predicting activating missense mutations in protein kinases

Carlos H. M. Rodrigues, David B. Ascher* & Douglas E. V. Pires*

Abstract: Protein phosphorylation and dephosphorylation play vital roles in a variety of cellular processes, and the balance between them must be closely regulated. Disturbances in the harmonic relationship between protein phosphorylation and dephosphorylation, through the introduction of dominant activating missense mutations in protein kinases, are known to be driver events of many cancer. Despite this, the identification of potential activating mutations has proven to be a difficult task, and has been limited to evolutionary and sequence-based comparisons with previously characterised mutations. This study aims to fill this gap by proposing a novel machine learning method for predicting missense activating mutations on protein kinases, named Kinact. Experimental data on 384 point mutations in 42 different protein kinases was collected from Kin-Driver, Clinvar and Ensembl databases. The resulting data sample was then manually curated and 258 mutations were mapped into solved 3D structures of the Protein Data Bank. Each protein was classified into one group of the Kinase Classification and a set of in-silico analysis were performed with sequence and structure data. The most descriptive features were then used as input for training and testing supervised learning algorithms and predictive classification models that rely on attributes solely from sequence level, structural level and in combination were generated. The best performing model was observed when a combination of structural and sequence-based features were used as evidence during the learning task, achieving a precision of up to 90% and Area Under ROC Curve of 0.97 under 10-fold cross-validation and precision of 85% and Area Under ROC Curve of 0.90 on blind tests. We show the best performing model of Kinact significantly outperforms the gold-standard methods used by clinical geneticists (p-value < 0.01), SIFT and PolyPhen-2, which achieved Area Under ROC Curve of 0.49 and 0.63 on the training data set, respectively and 0.67 and 0.53, respectively, on the blind test. Kinact conveniently combines high-performance open source web visualization tools to assist further research on how mutations affect protein kinases activity. The method is freely available as a user friendly, easy to use web server at http://biosig.unimelb.edu.au/kinact/.

Best viewed using Chrome on 1280x960 resolution and above

Available at <http://biosig.unimelb.edu.au/kinact>.

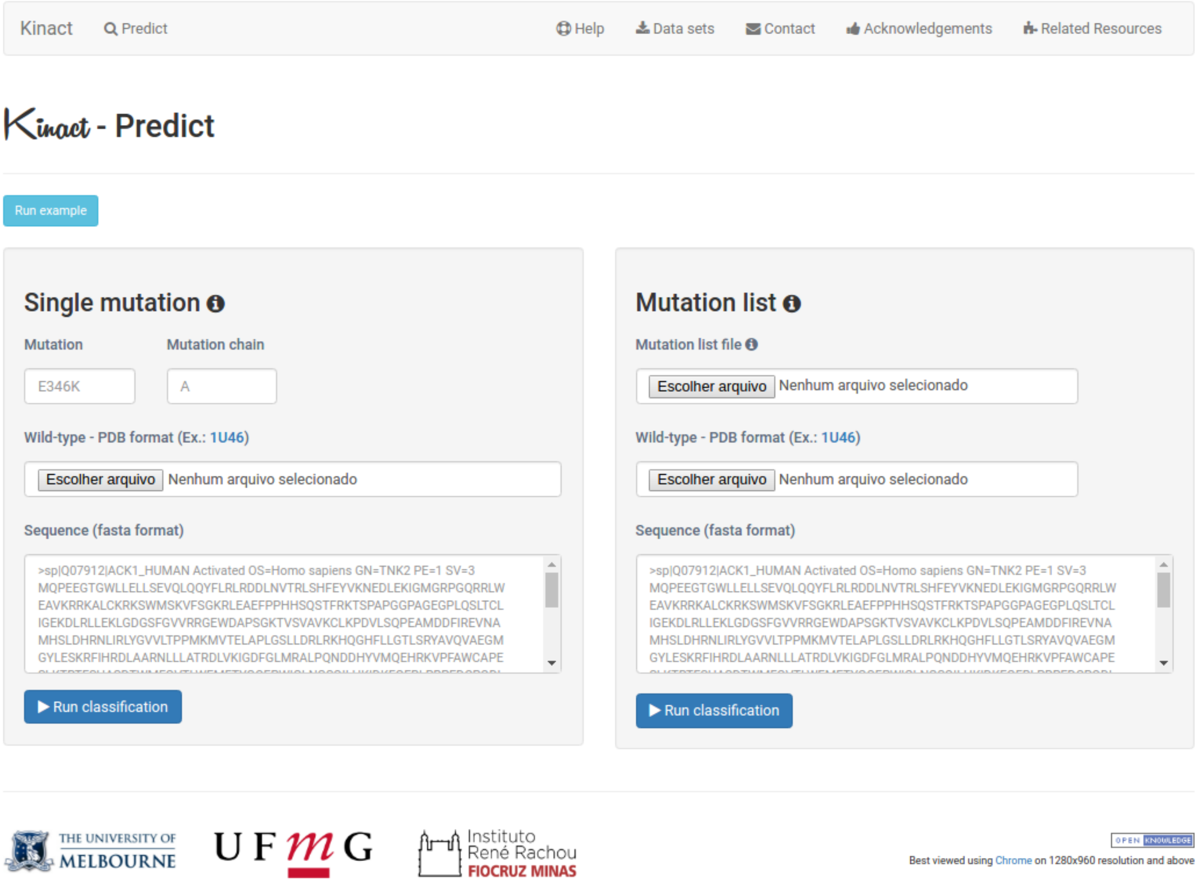
features available based on the data provided. The more data supplied to Kinact, the more trustworthy its prediction outcome.

4.3.2 Data Preparation

As soon as the mutation is submitted by the user a set of validation steps are executed before proceeding with the classification, such as checking if the PDB file format is correct and whether the residue number of mutation it is actually mapped into the 3D structure. If the input is not valid an error message is displayed to inform the user about the problem. Figure 24 shows the error message for a submission where no PDB file was provided.

If the data provided is valid, the attributes are calculated and then serve as input for the prediction model. Again, at this point the attributes are calculated based on the type of data submitted. For instance, if the user provide a PDB file and also the fasta sequence, Kinact will generate structure and sequence-based attributes. If only the PDB structure data is provided, only structure based attributes are calculated before the

Figure 23 – Kinact input page



Kinact Predict Help Data sets Contact Acknowledgements Related Resources

Kinact - Predict

Run example

Single mutation

Mutation: E346K

Mutation chain: A

Wild-type - PDB format (Ex.: 1U46)

Escolher arquivo Nenhum arquivo selecionado

Sequence (fasta format)

```
>sp|Q07912|ACK1_HUMAN Activated OS=Homo sapiens GN=TNK2 PE=1 SV=3
MQPEEGTGWLLELLSEVQLQQYFLRLRDDLNVTRLSHFEYVKNEDLEKIGMGRPGQRRLW
EAVKRRKALCKRKSWMSKVFSGKRLEAEFPPHHSQSTFRKTSPAPGGPAGEGPLQSLTCL
IGEKDLRLLEKLGDGSFGVVRRGEWDAPSGKTVSVAVKCLKPDVLSQPEAMDDFIREVNA
MHSLDHRNLIRLYGVVLTPPMKMVTELAPLGSLLDRLRKHQGHFLLGTLSRYAVQVAEGM
GYLESKRFIHRDLAARNLLLATRDLVKIGDFGLMRALPQNDDHYVMQEHRKVPFAWCAPE
```

▶ Run classification

Mutation list

Mutation list file

Escolher arquivo Nenhum arquivo selecionado

Wild-type - PDB format (Ex.: 1U46)

Escolher arquivo Nenhum arquivo selecionado

Sequence (fasta format)

```
>sp|Q07912|ACK1_HUMAN Activated OS=Homo sapiens GN=TNK2 PE=1 SV=3
MQPEEGTGWLLELLSEVQLQQYFLRLRDDLNVTRLSHFEYVKNEDLEKIGMGRPGQRRLW
EAVKRRKALCKRKSWMSKVFSGKRLEAEFPPHHSQSTFRKTSPAPGGPAGEGPLQSLTCL
IGEKDLRLLEKLGDGSFGVVRRGEWDAPSGKTVSVAVKCLKPDVLSQPEAMDDFIREVNA
MHSLDHRNLIRLYGVVLTPPMKMVTELAPLGSLLDRLRKHQGHFLLGTLSRYAVQVAEGM
GYLESKRFIHRDLAARNLLLATRDLVKIGDFGLMRALPQNDDHYVMQEHRKVPFAWCAPE
```

▶ Run classification

THE UNIVERSITY OF MELBOURNE UFMG Instituto René Rachou FIOCRUZ MINAS OPEN KNOWLEDGE

Best viewed using Chrome on 1280x960 resolution and above

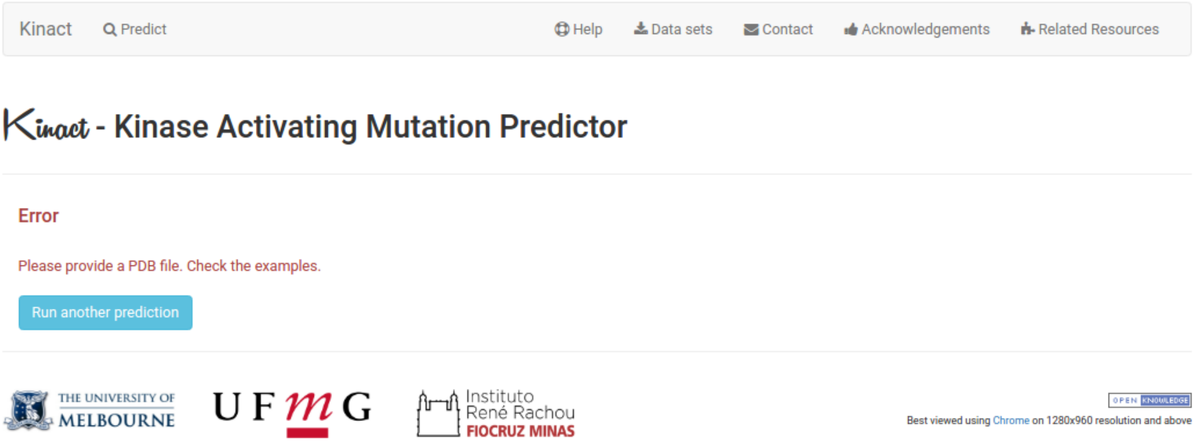
Kinact allows users to submit one single mutation for prediction (left) and also a file with a list of mutations and chain identifiers (right) to be evaluated for a specific PDB file and/or sequence also required as input. Users are encouraged to provide both sequence and structural information so that the prediction of Kinact uses all types of features as evidence in a more reliable outcome prediction, as discussed in section 4.2. Available at <http://biosig.unimelb.edu.au/kinact/prediction>.

classification and so forth.

Due the fact that Kinact uses a considerable amount of tools for generating attributes (mCSM, DUET, SDM, CSA, PolyPhen-2, SIFT, for example) and given the time that takes for each of these tools to be executed can vary depending on the data submitted, the calculations are packed into a job and moved to a queue for asynchronous processing, freeing up the web application to respond to other requests. The jobs in the queue are executed by a group of processes in the background called workers. Meanwhile a web page that automatically refreshes 10 seconds verifies whether the job is already processed so that the prediction results can be displayed. Such queue implementation is achieved by the open source project Redis[6] which is an in-memory data structure store, used as a database, cache and message broker.

[6] <https://redis.io/>

Figure 24 – Kinact error page



Kinact - Kinase Activating Mutation Predictor

Error

Please provide a PDB file. Check the examples.

Best viewed using Chrome on 1280x960 resolution and above

Message is displayed when the validation of user data submission is not valid. In this case, no PDB file was provided.

4.3.3 Output

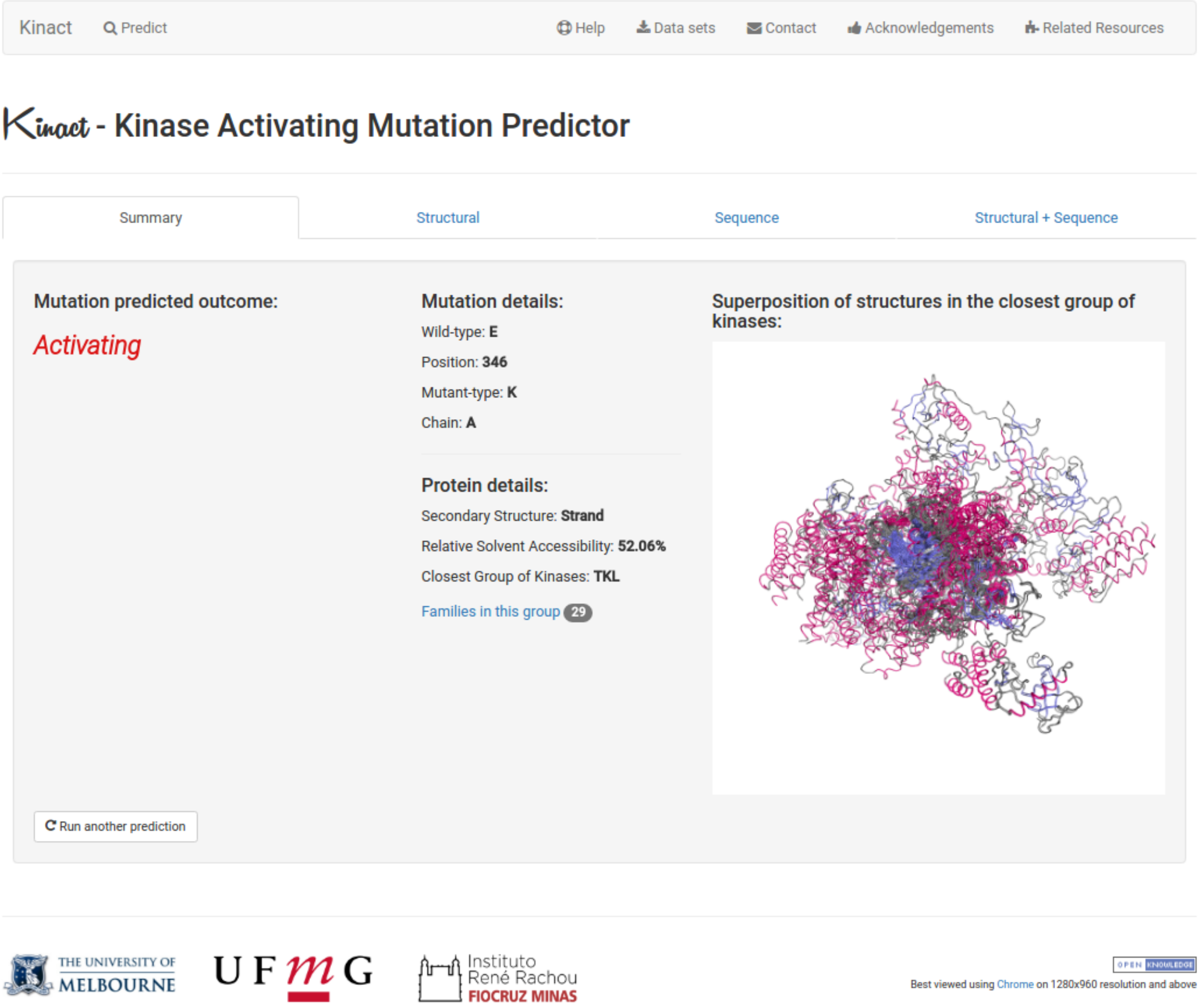
The primary step, after the attribute calculation, mentioned previously, is the assignment of the submitted molecule into one of the kinase families. This is accomplished by the help of Kinannote, also described earlier. This step is crucial for Kinact output, since the prediction outcome consists not only of the actual prediction based on the machine learning model (activating or non-activating), but it also provides correlated information with mutation data collected for all proteins in the same kinase group. If Kinannote outputs results indicates that the molecule submitted might not be a kinase the prediction of Kinact is still executed, but a message is displayed to inform the user.

For the “Single Mutation” option, and assuming that PDB structure and sequence is provided, Kinact outputs the model prediction (activating or non-activating) aside with mutation details and information on the kinase group in which the submitted molecule was assigned by Kinannote. All the families of protein kinases comprised in the group are also displayed. Figure 18, shows an example of such prediction summary page.

Alongside with the summary of the prediction, Kinact also provides a set of analyses depending on the data submitted by the user. These analyses are separated by tabs in the results page as displayed in Figure 25.

Structural analysis is performed with the help of 3Dmol.js (REGO; KOES, 2015) which is a powerful object-oriented Javascript library that provides interactive, hardware-accelerated three-dimensional representations of molecular data without the need to install browser plugins or Java. By default, the molecule is represented as cartoon with the mutated residue highlighted as stick. The residues that surround the mutated residue and make interatomic interactions, according to Arpeggio, are also highlighted and labeled and the interactions are colored according to their type. A legend for the binding type

Figure 25 – Summary page of Kinact prediction



Kinact - Kinase Activating Mutation Predictor

Mutation predicted outcome:

Activating

Mutation details:

Wild-type: **E**

Position: **346**

Mutant-type: **K**

Chain: **A**

Protein details:

Secondary Structure: **Strand**

Relative Solvent Accessibility: **52.06%**

Closest Group of Kinases: **TKL**

Families in this group 29

Superposition of structures in the closest group of kinases:

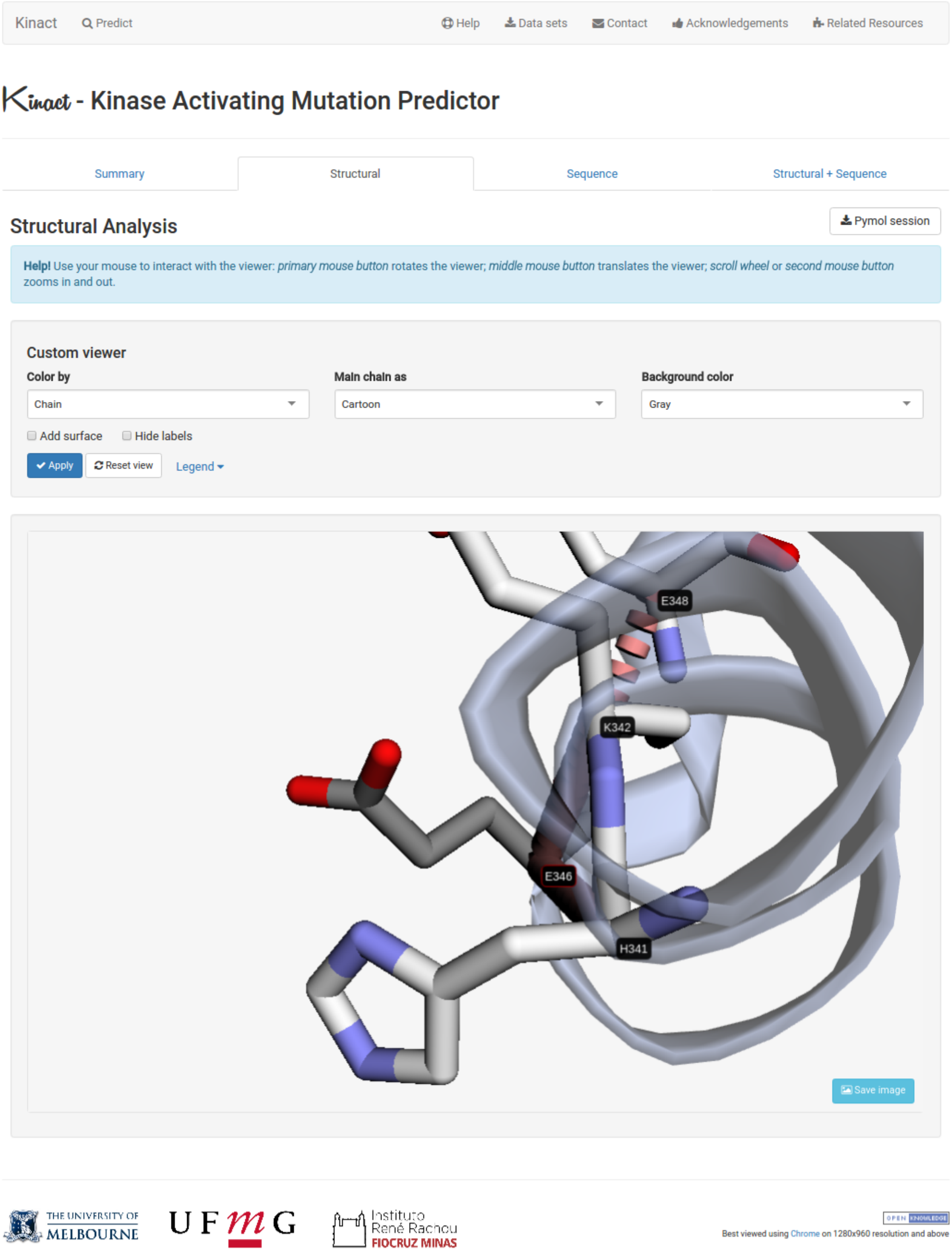
Best viewed using Chrome on 1280x960 resolution and above

The prediction outcome is displayed in the left side. Mutations details and protein details are showed in the center. Lastly, a image with the superposition of structures in the closes group of kinases in which the submitted molecule was classified is displayed in the right side.

is also provided. Furthermore a set of options for customization are provided, such as changing color, representation and adding surface. Users can also save the image of the viewer at any time through the "Save image" button at the bottom right corner of the viewer. Figure 26 displays an example of the structural analysis results page.

Sequence analysis is also provided in the form of a Multiple Sequence Alignment (MSA) with all the proteins in the same group of kinases according to Kinannote, as demonstrated in Figure 27. Such MSA, allows the user to visualize all the activating mutations (highlighted in with red background) present in every protein of the group with external links for the mutation evidence. All the residues are colored by its type: Polar as pink, Hydrophobic as light green, Charged as blue and Sulphuretted as orange. A legend is also provided for helping users understand such representations.

Figure 26 – Structural analysis results page



Kinact - Kinase Activating Mutation Predictor

Structural Analysis

Help! Use your mouse to interact with the viewer: *primary mouse button* rotates the viewer; *middle mouse button* translates the viewer; *scroll wheel* or *second mouse button* zooms in and out.

Custom viewer

Color by: Chain

Main chain as: Cartoon

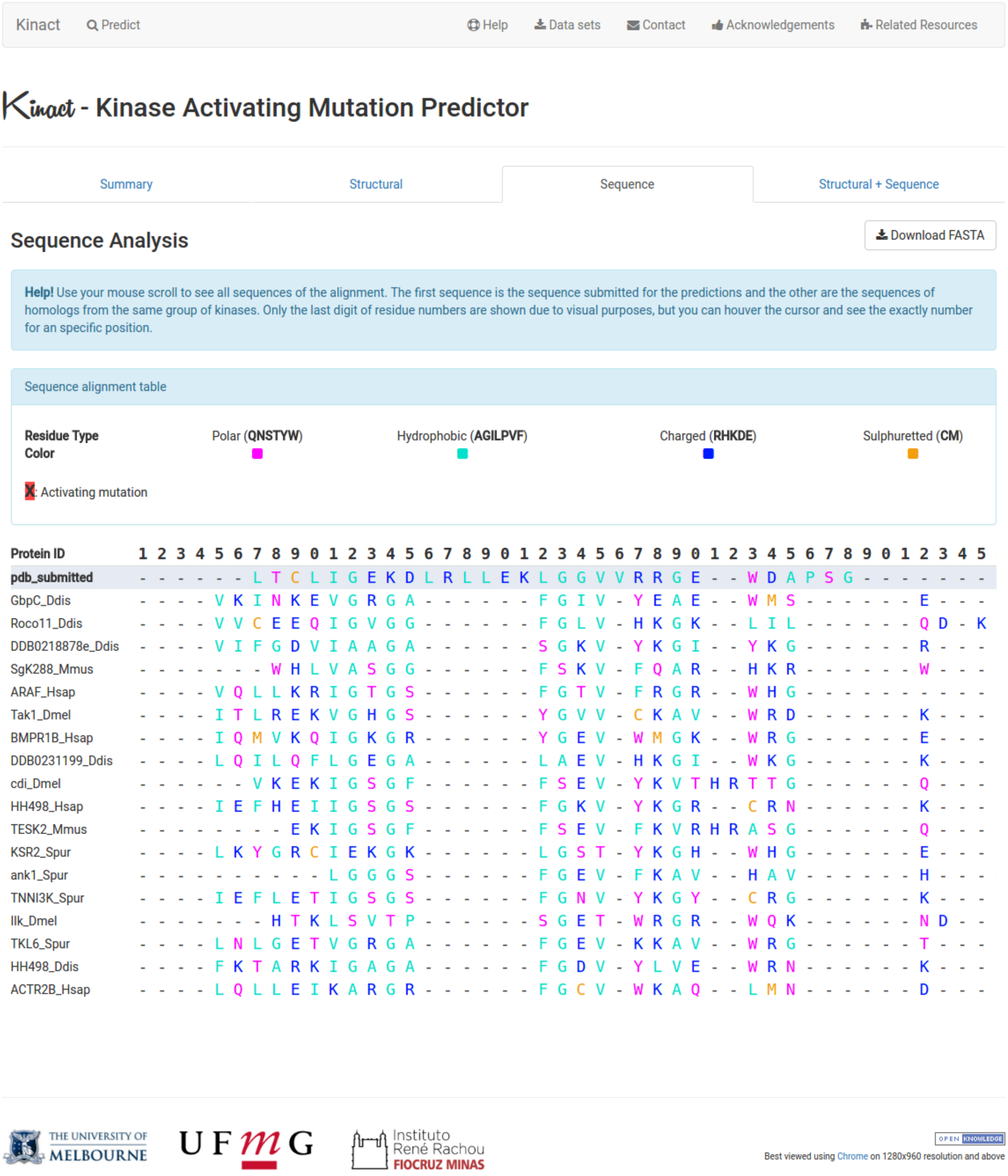
Background color: Gray

□ Add surface □ Hide labels

Best viewed using Chrome on 1280x960 resolution and above

By default the molecule is displayed using the cartoon representation with the mutated residue highlighted as stick and labeled as well as the surrounding residues that have interactions with it according to Arpeggio. On the top of the page, a set of options allow the user to customize the viewer and a legend is also provided for the binding types.

Figure 27 – Sequence analysis results page



Kinact - Kinase Activating Mutation Predictor

Sequence Analysis

Help! Use your mouse scroll to see all sequences of the alignment. The first sequence is the sequence submitted for the predictions and the other are the sequences of homologs from the same group of kinases. Only the last digit of residue numbers are shown due to visual purposes, but you can houver the cursor and see the exactly number for an specific position.

Sequence alignment table

Residue Type **Color**	Polar (**QNSTYW**)	Hydrophobic (**AGILPVF**)	Charged (**RHKDE**)	Sulphuretted (**CM**)
	■	■	■	■

X: Activating mutation

Protein ID	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5
pdb_submitted	-	-	-	-	-	-	L	T	C	L	I	G	E	K	D	L	R	L	L	E	K	L	G	G	V	V	R	R	G	E	-	-	W	D	A	P	S	G	-	-	-	-	-	-	-
GbpC_Ddis	-	-	-	-	V	K	I	N	K	E	V	G	R	G	A	-	-	-	-	-	-	F	G	I	V	-	Y	E	A	E	-	-	W	M	S	-	-	-	-	-	-	E	-	-	-
Roco11_Ddis	-	-	-	-	V	V	C	E	E	Q	I	G	V	G	G	-	-	-	-	-	-	F	G	L	V	-	H	K	G	K	-	-	L	I	L	-	-	-	-	-	-	Q	D	-	K
DDB0218878e_Ddis	-	-	-	-	V	I	F	G	D	V	I	A	A	G	A	-	-	-	-	-	-	S	G	K	V	-	Y	K	G	I	-	-	Y	K	G	-	-	-	-	-	-	R	-	-	-
SgK288_Mmus	-	-	-	-	-	-	-	W	H	L	V	A	S	G	G	-	-	-	-	-	-	F	S	K	V	-	F	Q	A	R	-	-	H	K	R	-	-	-	-	-	-	W	-	-	-
ARAF_Hsap	-	-	-	-	V	Q	L	L	K	R	I	G	T	G	S	-	-	-	-	-	-	F	G	T	V	-	F	R	G	R	-	-	W	H	G	-	-	-	-	-	-	-	-	-	-
Tak1_Dmel	-	-	-	-	I	T	L	R	E	K	V	G	H	G	S	-	-	-	-	-	-	Y	G	V	V	-	C	K	A	V	-	-	W	R	D	-	-	-	-	-	-	K	-	-	-
BMPR1B_Hsap	-	-	-	-	I	Q	M	V	K	Q	I	G	K	G	R	-	-	-	-	-	-	Y	G	E	V	-	W	M	G	K	-	-	W	R	G	-	-	-	-	-	-	E	-	-	-
DDB0231199_Ddis	-	-	-	-	L	Q	I	L	Q	F	L	G	E	G	A	-	-	-	-	-	-	L	A	E	V	-	H	K	G	I	-	-	W	K	G	-	-	-	-	-	-	K	-	-	-
cdi_Dmel	-	-	-	-	-	-	V	K	E	K	I	G	S	G	F	-	-	-	-	-	-	F	S	E	V	-	Y	K	V	T	H	R	T	T	G	-	-	-	-	-	-	Q	-	-	-
HH498_Hsap	-	-	-	-	I	E	F	H	E	I	I	G	S	G	S	-	-	-	-	-	-	F	G	K	V	-	Y	K	G	R	-	-	C	R	N	-	-	-	-	-	-	K	-	-	-
TESK2_Mmus	-	-	-	-	-	-	-	-	E	K	I	G	S	G	F	-	-	-	-	-	-	F	S	E	V	-	F	K	V	R	H	R	A	S	G	-	-	-	-	-	-	Q	-	-	-
KSR2_Spur	-	-	-	-	L	K	Y	G	R	C	I	E	K	G	K	-	-	-	-	-	-	L	G	S	T	-	Y	K	G	H	-	-	W	H	G	-	-	-	-	-	-	E	-	-	-
ank1_Spur	-	-	-	-	-	-	-	-	-	-	L	G	G	G	S	-	-	-	-	-	-	F	G	E	V	-	F	K	A	V	-	-	H	A	V	-	-	-	-	-	-	H	-	-	-
TNNI3K_Spur	-	-	-	-	I	E	F	L	E	T	I	G	S	G	S	-	-	-	-	-	-	F	G	N	V	-	Y	K	G	Y	-	-	C	R	G	-	-	-	-	-	-	K	-	-	-
Ilk_Dmel	-	-	-	-	-	-	-	H	T	K	L	S	V	T	P	-	-	-	-	-	-	S	G	E	T	-	W	R	G	R	-	-	W	Q	K	-	-	-	-	-	-	N	D	-	-
TKL6_Spur	-	-	-	-	L	N	L	G	E	T	V	G	R	G	A	-	-	-	-	-	-	F	G	E	V	-	K	K	A	V	-	-	W	R	G	-	-	-	-	-	-	T	-	-	-
HH498_Ddis	-	-	-	-	F	K	T	A	R	K	I	G	A	G	A	-	-	-	-	-	-	F	G	D	V	-	Y	L	V	E	-	-	W	R	N	-	-	-	-	-	-	K	-	-	-
ACTR2B_Hsap	-	-	-	-	L	Q	L	L	E	I	K	A	R	G	R	-	-	-	-	-	-	F	G	C	V	-	W	K	A	Q	-	-	L	M	N	-	-	-	-	-	-	D	-	-	-

THE UNIVERSITY OF MELBOURNE
UFMG
Instituto René Rachou FIOCRUZ MINAS
OPEN KNOWLEDGE

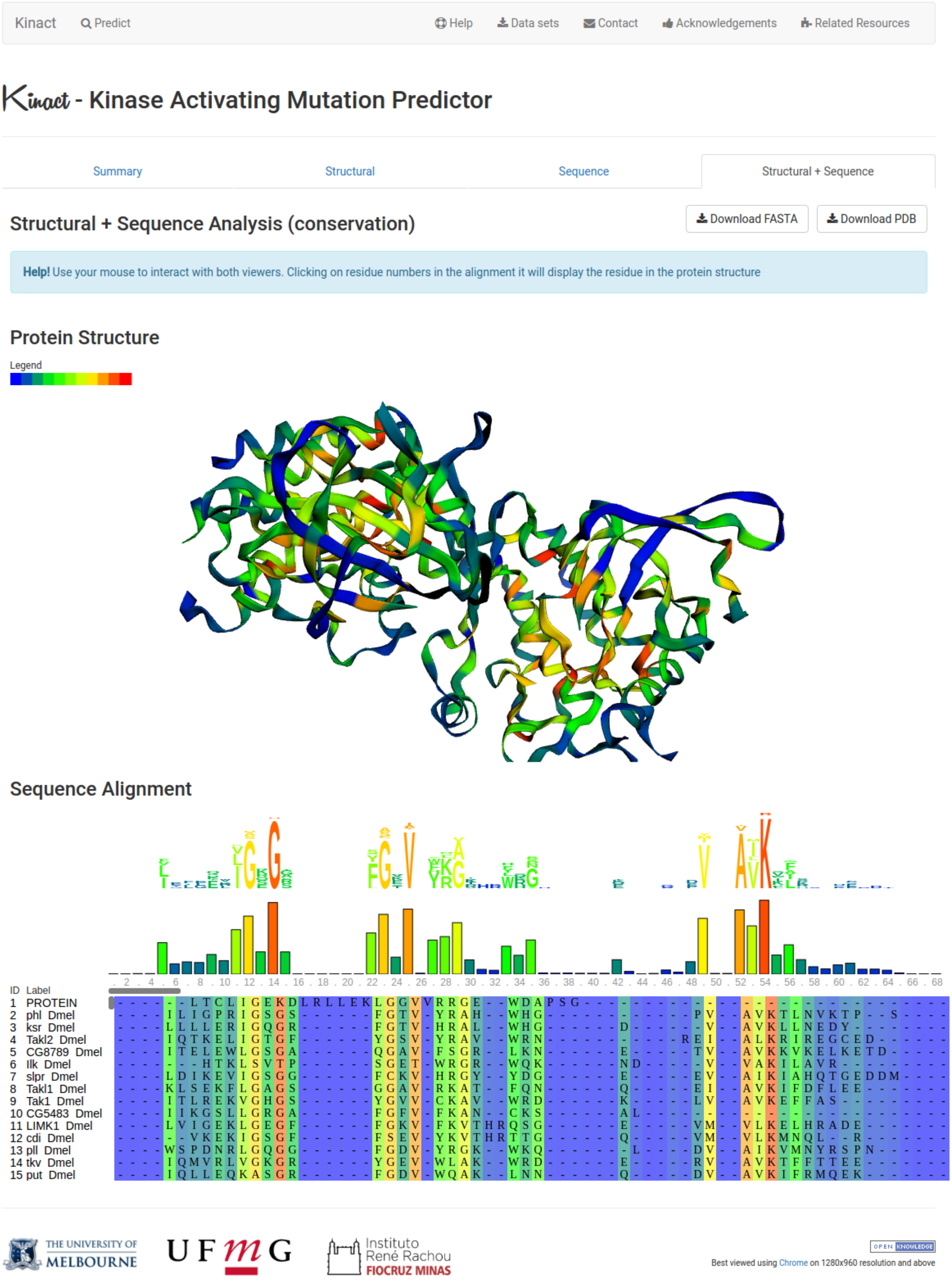
Best viewed using Chrome on 1280x960 resolution and above

MSA with the proteins of the group in which the submitted molecule was assigned according to Kinannote. All residues are colored by their type: Polar as pink, Hydrophobic as light green, Charged as blue and Sulphuretted as orange. Activating mutations are highlighted with red background and external links for the mutations evidence is provided. A legend is shown on top of the page.

Lastly, a combination of analysis is shown under the “Structural + Sequence” tab. Here both structure and sequence are displayed and the whole molecule in the 3D viewer and the MSA are colored according to their conservation in the group of kinase, from blue (not conserved) to red (conserved). 3Dmol.js is also used alongside with the MSAViewer (YACHDAV et al., 2016) which is a quick and easy to use visualization and analysis Javascript component for Multiple Sequence Alignment data. Like 3Dmol.js it does not require any specialized software to be installed. The MSAViewer is part of the BioJS collection of components (CORPAS et al., 2014). Figure 27 exhibits the “Structural + Sequence” tab on the results page. Protein Structure is displayed on top and MSA on the bottom. A sequence logo is also shown on top of the MSA to assist conservation analysis.

The results page for the option “Mutation list” is presented in a tabular format with a set of details about each mutation on the summary tab results page, such results can be downloaded as a comma separated file (csv). Structure and sequence analysis are performed similarly to what was described for the option “Single mutation”.

Figure 28 – Structural + Sequence analysis results page



Kinact - Kinase Activating Mutation Predictor

Structural + Sequence Analysis (conservation)

Help! Use your mouse to interact with both viewers. Clicking on residue numbers in the alignment it will display the residue in the protein structure

Protein Structure

Sequence Alignment

```
ID Label
1  PROTEIN      ------LTCLIGEKDLRLLEKLGGVVRRGE--WDAPSG------------------------------
2  phl  Dmel    ----ILIGPRIGSGS------FGTV-YRAH--WHG------------PV--AVKTLNVKTP--S----
3  ksr  Dmel    ----LLLLERIGQGR------FGTV-HRAL--WHG------D------V--AVKLLNEDY--------
4  Takl2  Dmel  ----IQTKELIGTGF------YGSV-YRAV--WRN-----------REI--ALKRIREGCED------
5  CG8789  Dmel ----ITELEWLGSGA------QGAV-FSGR--LKN------E-----TV--AVKKVKELKETD-----
6  Ilk  Dmel    -------HTKLSVTP------SGET-WRGR--WQK------ND-----V--VAKILAVR---------
7  slpr  Dmel   ----LDIKEVIGSGG------FCKV-HRGY--YDG------E-----EV--AIKIAHQTGEDDM----
8  Takl1  Dmel  ----KLSEKFLGAGS------GGAV-RKAT--FQN------Q-----EI--AVKIFDFLEE-------
9  Tak1  Dmel   ----ITLREKVGHGS------YGVV-CKAV--WRD------K-----LV--AVKEFFAS---------
10 CG5483  Dmel ----IIKGSLLGRGA------FGFV-FKAN--CKS------AL-------------------------
11 LIMK1  Dmel  ----LVIGEKLGEGF------FGKV-FKVTHRQSG------E-----VM--VLKELHRADE-------
12 cdi  Dmel    ------VKEKIGSGF------FSEV-YKVTHRTTG------Q-----VM--VLKMNQL--R-------
13 pll  Dmel    ----WSPDNRLGQGG------FGDV-YRGK--WKQ-------L----DV--AIKVMNYRSPN------
14 tkv  Dmel    ----IQMVRLVGKGR------YGEV-WLAK--WRD------E-----RV--AVKTFFTTEE-------
15 put  Dmel    ----IQLLEQKASGR------FGDV-WQAK--LNN------Q-----DV--AVKIFRMQEK-------
```

THE UNIVERSITY OF MELBOURNE
UFMG
Instituto René Rachou FIOCRUZ MINAS
OPEN KNOWLEDGE

Best viewed using Chrome on 1280x960 resolution and above

Protein 3D structure is shown on top and Sequence Alignment on bottom. Both 3D structure and columns in MSA are colored according to residue conservation varying from blue (not conserved) to red (conserved). A sequence logo diagram is also displayed on top of the MSA to aid the identification of residues conservation on an specific position.

5 Conclusion and Perspectives

Protein kinases catalyze phosphorylation, a key regulatory reaction across most signalling and biological pathways. Mutations in these proteins that lead to dysregulation of catalytic activity play important roles in many diseases, and therefore the ability to identify these mutations in genomic sequences has significant implications to help guide patient management and treatment. However, no robust computational method for identifying activating mutations in protein kinases have been developed; with the standard clinically used tools being of limited use in the identification of these mutations.

The aim of this work was to address this limitation through the development of Kinact, a novel robust machine learning method for predicting missense activating mutations in kinases that also provides a set of modern visualization tools to support analyses of such mutations from both sequence and structural perspectives.

The inclusion of both sequence and structural information in the final model outperformed the use of either alone, and the current standard clinical tools SIFT and PolyPhen-2. Models trained on either sequence or structural attributes separately provided plausible results, but the performance was greatly improved when the algorithms were trained using all the attributes, specially for the identification of non-activating mutations, making the predictions more balanced between mutation classes. Amongst potential future steps for improvement are the inclusion of new sequence-based attributes into the method, in particular more explicit consideration of residue conservation within the specific sub-group of kinases, and incorporation of information of the location of known characterised mutations within the kinase sub-group, as well as assessing the impact of different kinase conformations (active/inactive) on predictive performance.

Despite the small amount of available data, and the biased distribution of the data set, a set of careful validation steps were performed to ensure the predictions reliability and robustness of the method.

In order to facilitate and streamline the routine collection, analysis and incorporation of new data into Kinact, all scripts for parsing and attribute generation are versioned with Git[1] and hosted on Bitbucket [2]. This greatly improve the ability to expand this methodology across a broader range of databases of mutations, such as COSMIC (FORBES et al., 2009) and ExAC (KARCZEWSKI et al., 2017) which require extra steps toward the curation of their data.

Lastly, more sophisticated computational approaches, such as deep learning, can

[1] <https://git-scm.com/>
[2] <https://bitbucket.com>

be introduced as an attempt to analyze broader perspectives of mutations in kinases. However, more data is needed to apply such techniques and, as discussed in section 3.1.1, comparative modeling methods for predicting the 3D structure could greatly benefit future implementations of this study by providing a broader range of data.

Appendix

APPENDIX A – Publications and Events

EXPERT OPINION ON DRUG DISCOVERY, 2017
VOL. 12, NO. 6, 553–563
https://doi.org/10.1080/17460441.2017.1322579

Taylor & Francis
Taylor & Francis Group

REVIEW

Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design

Amanda T.S. Albanaz[a,b*], Carlos H.M. Rodrigues[a,b*], Douglas E.V. Pires[a*] and David B. Ascher[a,c,d]

[a]Centro de Pesquisas René Rachou, FIOCRUZ, Belo Horizonte, MG, Brazil; [b]Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; [c]Department of Biochemistry, University of Cambridge, Cambridge, Cambridgeshire, UK; [d]Department of Biochemistry and Molecular Biology, University of Melbourne, Melbourne, Victoria, Australia

ABSTRACT

Introduction: Mutations introduce diversity into genomes, leading to selective changes and driving evolution. These changes have contributed to the emergence of many of the current major health concerns of the 21st century, from the development of genetic diseases and cancers to the rise and spread of drug resistance. The experimental systematic testing of all mutations in a system of interest is impractical and not cost-effective, which has created interest in the development of computational tools to understand the molecular consequences of mutations to aid and guide rational experimentation.

Areas covered: Here, the authors discuss the recent development of computational methods to understand the effects of coding mutations to protein function and interactions, particularly in the context of the 3D structure of the protein.

Expert opinion: While significant progress has been made in terms of innovative tools to understand and quantify the different range of effects in which a mutation or a set of mutations can give rise to a phenotype, a great gap still exists when integrating these predictions and drawing causality conclusions linking variants. This often requires a detailed understanding of the system being perturbed. However, as part of the drug development process it can be used preemptively in a similar fashion to pharmacokinetics predictions, to guide development of therapeutics to help guide the design and analysis of clinical trials, patient treatment and public health policy strategies.

ARTICLE HISTORY
Received 27 January 2017
Accepted 20 April 2017

KEYWORDS
Mutational analysis; genetic diseases; drug resistance; cancer; drug design; molecular mechanism; genotype-phenotype association

1. Introduction

Changes at the genetic level can result in drastic changes in cellular phenotypes and behavior. These changes can lead to disease, or provide selective advantages that promote the development of drug resistance. In particular, non-synonymous single-nucleotide polymorphisms (nsSNPs) within the protein coding regions of the genome have been strongly associated with occurrence and predisposition of human disease and drug resistance, sparking great interest from the research community.

The rapid developments in high-throughput sequencing, including dramatic drops in the cost, have created vast opportunities to understand the link between our genomes and phenotypes. This has opened up the promises of personalized medicines, targeted therapies, and targeted public health policies. In order to fully realize the potential of these developments, however, we still need to improve our understanding of what are the molecular consequences of a given mutation, and how do these lead to a given phenotype.

While considerable resources have been invested in the experimental evaluation of genomic mutations, characterizing mutation effects is a challenging task and impractical to systematically experimentally evaluate all possible mutations for a given protein of interest, even more considering the range of different mechanisms in which mutations can affect protein function and interactions. Traditional experimental approaches are also not efficient enough or do not achieve scalability required to provide real time guidance into patient treatment and public health policy. This has led to significant interest in the development of computational approaches to rapidly and accurately evaluate the effects of mutations. Figure 1 summarizes how *in silico* mutation analysis can be helpful in deconvoluting genotype-phenotype associations obtained from the wealth of genomic variation generated from sequencing efforts, including shedding light into disease predisposition and its mechanisms in a molecular level. Such methods can also be used to mutation prioritization for further experimental investigation, identification, and anticipation of resistant variants and resistance hotspots, knowledge that can be applied in the design of drugs less prone to resistance as well as to drive the development of public health policies and aid in establishing more appropriate and personalized treatments.

2. Analyzing the effects of mutations

The two most commonly used methods by clinical geneticists to look at the effects of coding nsSNP mutations in the human genome are SIFT [1] and Polyphen [2]. Other approaches

CONTACT David B. Ascher david.ascher@unimelb.edu.au

*These authors contributed equally for this work.

Article highlights

- Scalable and reliable structural based computational approaches are providing detailed insight into the molecular consequences of coding mutations.
- These have been used to guide patient treatment strategies for renal cell carcinoma and genetic diseases.
- Using these methods, drug resistance mutations can be identified and predicted.
- Used in a preemptive fashion, these can help guide drug development in the search for new therapeutics less likely to develop resistance.
- Mutations can give rise to a phenotype through different molecular mechanisms which can be assessed via integration of computational methods.

This box summarizes key points contained in the article.

include CADD [3] and MutationTaster [4]. These approaches use the protein sequence to evaluate whether a given mutation is likely to be pathogenic or not. However, they have been limited by the lack of mechanistic information they provide and their overestimation of mutations likely to be pathogenic [5]. Structural approaches can complement these analyses by providing detailed mechanistic information, but historically have involved a trade-off between scalability and molecular level mechanistic information, with molecular dynamics approaches providing greater atomic detail, but proving impractical for comprehensive analysis of a large number of different mutations.

In the 1990s, efforts to utilize the expanding structural information available for many proteins led to the development of SDM [6], the first method for predicting the effects

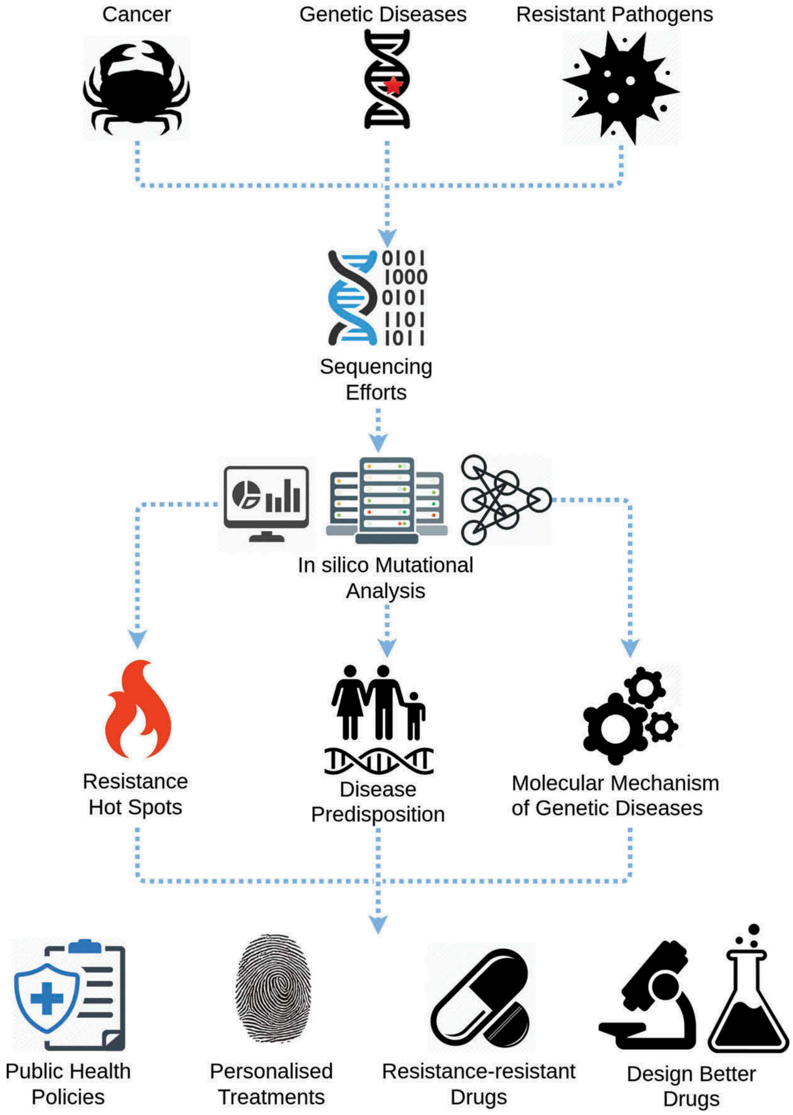


Figure 1. The use of in silico mutational analysis to tackle drug resistance and genetic diseases. Sequencing efforts generate a wealth of genomic variation. Computational mutation analysis can help deconvolute genotype-phenotype associations aiding in understanding the molecular mechanism of diseases and disease predisposition as well as in mutation prioritization for experimental validation, identification of resistant variants and resistance hot-spots, which can then fed into drug design pipelines as well drive the development of public health policies and choice of more appropriate and personalized treatments.

Table 1. Recent structure-based computational methods for analyzing the effects of coding mutations.

Method	Web server[a]	Publication year	Reference[b]
Effects of Mutations on Protein Stability and Folding			
SDM	http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php	2011	[23,24]
	http://structure.bioc.cam.ac.uk/sdm2	2017	
PoPMuSiC 2.1	http://babylone.ulb.ac.be/popmusic	2011	[25]
mCSM-Stability	http://structure.bioc.cam.ac.uk/mcsm/stability	2014	[13]
DUET	http://structure.bioc.cam.ac.uk/duet	2014	[12]
ENCoM	http://bcb.med.usherbrooke.ca/encom.php	2015	[26]
MAESTROweb	https://biwww.che.sbg.ac.at/maestro/web	2016	[27]
STRUM	http://zhanglab.ccmb.med.umich.edu/STRUM/	2016	[28]
ELASPIC	http://elaspic.kimlab.org	2016	[29]
Effects of Mutations on Protein-Protein Binding Affinity			
BeAtMuSiC	http://babylone.ulb.ac.be/beatmusic/	2013	[30]
mCSM-PPI	http://structure.bioc.cam.ac.uk/mcsm/protein_protein	2014	[13]
mCSM-AB	http://structure.bioc.cam.ac.uk/mcsm_ab	2016	[9]
MutaBind	https://www.ncbi.nlm.nih.gov/projects/mutabind	2016	[31]
Effects of Mutations on Protein-Nucleic Acid Interactions			
mCSM-NA	http://structure.bioc.cam.ac.uk/mcsm/protein_dna	2014	[13,11]
	http://structure.bioc.cam.ac.uk/mcsm_na	2017	
Effect of Mutations on Protein-Small Molecule Interactions			
mCSM-Lig	http://structure.bioc.cam.ac.uk/mcsm_lig	2016	[14]
CSM-Lig	http://structure.bioc.cam.ac.uk/csm_lig	2016	[10]

[a] The URLs link to the webserver to run the method. Links last accessed in April 2017.

[b] The primary reference describing the method, and which should be cited if used.

of mutations on protein folding and stability. Subsequent efforts by other groups led to a range of methods to predict the same effects, improving upon the accuracy but not considering the other potential structural effects mutations might lead to.

This was first addressed through the systematic application of cut-off scanning matrices [7,8] to quantitatively and scalably predict the effects of mutations on the binding affinities to other ligands, including other proteins, nucleic acids, small molecules, and metal ions [9–14]. Table 1 presents a summary of the main structure-based methods proposed over the past years to analyze the different effects of mutations on coding regions. While this started to allow the deconvolution of the individual molecular changes that might be occurring, the big question limiting their application, especially in a clinical setting, was how do these individual effects combine to lead to a phenotype? Recent efforts have started to integrate these structural effects in order to better understand phenotypes, and have been used to look at a number of different human health problems driven by mutations in protein coding regions [14–22].

3. Using mutation analysis to guide treatment: toward personalized treatments

3.1. Cancers

By analyzing the molecular effects of mutations in common renal cell carcinoma genes, including *p15* and *SDHA*, these have been correlated to a patient's risk of developing renal carcinoma. This was best demonstrated by recent studies looking at mutations in the von Hippel–Lindau protein (VHL) associated with the development of clear cell renal cell carcinoma (ccRCC) [15,16,32,33]. By assessing whether a mutation affected the stability of the protein, or disrupted interactions to Elongin or HIF-1α, a patient could be classified into high-, medium-, and low-risk groups that could help guide screening strategies and provide more focused genetic counseling. The available clinical data from over 100 patients was integrated with a saturation mutagenesis analysis of all possible mutations on VHL producing Symphony, a relational database mapping experimental and predicted risks of mutations to its molecular mechanism, aiding the characterization of newly discovered variants.

Understanding cancer genetics has been important for the diagnosis and treatment of a range of other cancers [34,35], with increasing interest in how the structural impacts of mutations can be used to interpret sequence information. This has led to recent efforts to map the COSMIC database onto protein structures.

3.2. Mendelian genetic diseases

Alkaptonuria (AKU), also known as ochronosis or black bone disease, is a rare recessive inherited genetic disease and first metabolic disorder firstly described over 100 years ago. AKU is caused by coding mutations that disrupt structure and function of the enzyme homogentisate 1,2-dioxygenase (HGD), related to phenylalanine and tyrosine metabolism. HGD gene product folds to form a homo-hexamer disposed as two stacked trimers, quaternary structure which is necessary for enzyme function.

Two comprehensive analysis on AKU causing mutations were carried out in an attempt to characterize the potential molecular mechanisms on which mutations could disruption enzyme activity [17,18].

Mutation effects on protein monomer stability as well as protein-protein and protein-ligand affinity were predicted with the DUET, mCSM-PPI and mCSM-Lig web servers respectively. Three mutation clusters emerged from this analysis, regarding the molecular mechanism for structure and function disruption: (a) mutations that greatly affected monomer stability, therefore preventing oligomer formation; (b) mutations greatly reducing protein-protein affinity between the hexamer components, also preventing proper oligomer formation and (c) mutations with mild effects on both monomer stability and protein-protein affinity, which together caused functional impairment. The structural analysis of mutations in other Mendelian diseases, for example ornithine transcarbamylase deficiency [36], have identified that disease causing mutations lead to altered protein stability and interactions. Mutations with these molecular consequences occurred in roughly similar proportions to those observed in AKU.

These observations have been validated experimentally and expanded to examine all known disease causing mutations for inclusion in the HGD mutation database [37], which could hopefully guide the development of new, more effective and personalized drugs to treat this condition. For example, subsequent efforts have identified molecular stabilizers that reverse the effects of the destabilizing mutations, analogous to the recent successes on p53. They have also been used to classify patients in the SONIA2 clinical trial, as we know that the molecular mechanism of a mutation can alter how patients may respond to therapeutics [38].

Structural mutation analysis techniques have started to play important roles in the diagnosis of rare Mendelian genetic diseases. For example, establishing the genetic basis of epilepsy is a fundamental step for disease prognosis and choice of patient treatments [38]. Recently, these methods were used to not only identify the genetic cause of a previously undiagnosed or characterized human cohesinopathy but also characterize the molecular mechanism, subsequently experimentally validated [39]. The potential for the structural characterization of mutations to impact upon clinical practice will only continue to grow with the increasing availability of structural information, and routine use of exome sequencing in patient care.

3.3. *Screening for drug resistance in tuberculosis*

The reduction of sequencing costs, and improvements in accuracy and sensitivity, have led to interest in using high-throughput sequencing to diagnose patients, and identify drug resistance mutations. For infectious diseases such as tuberculosis (TB), where the drug susceptibility screening is time consuming and costly, genomic sequencing opens up the possibility of being able to more rapidly identify the correct treatment strategies for a patient, but also to guide public health policy by following the spread of resistance. Experimental innovations have allowed researchers to sequence the TB genome based on a sample of the patient's sputum, and Public Health England is now sequencing all new TB cases in the UK.

Many resistance mutations in TB have been well characterized, but one of the limitations of these approaches is how to interpret novel mutations identified within the genome. Due to the lack of horizontal gene transfer, TB is an ideal pathogen to apply structural based mutational analysis approaches. Looking at mutations in rpoB and katG, which leads to rifampicin and isoniazid resistance, respectively, clear structural features were identified that correlated strongly with the resulting effectiveness of the drugs (MIC) [40]. A number of resistance mutations have also been observed across protein-protein interfaces, which raises the interesting hypothesis that similar to Mendelian disease mutations, those at interfaces might be prone to lead to disease and resistance because they have a lower fitness cost associated to them than those in the active site that completely disrupt activity [36,41,42].

While previous experimental and clinical knowledge about the effect of a given mutation in a given strain on drug susceptibility will always provide the gold standard for predicting and identifying drug resistance, structural based approaches complement this limited available information by providing the power to look at novel mutations.

4. Targeting resistance mutations: toward resistance-resistant therapies

4.1. *HIV protease 1 inhibitors*

HIV protease catalyzes the cleavage of the polypeptide precursors into mature enzymes and structural proteins, an essential step in the HIV-1 replication cycle. Inhibitors targeting the HIV protease have been in clinical use since 1995 and include darunavir, amprenavir, atazanavir, nelfinavir, indinavir, saquinavir, and lopinavir [43,44].

Due to the HIV's error prone replication, resistance mutations against these inhibitors have evolved rapidly and been widely observed clinically, limiting the effectiveness of these therapies. These include mutations in the active site (V32I, L33F, I54M, and I84V) that through changes in hydrogen bonding and Van der Waals interactions between the inhibitors and the catalytic site amino acids, can reduce their binding affinities [45,46].

A better understanding of the effects of mutations on inhibitor binding and their molecular mechanism giving rise to resistance are crucial for designing novel drugs, more effectively and less prone to failure. Computational structure-based methods play an important role in tackling this challenge. The mCSM suite was successfully used to predict the effect of the aforementioned mutations upon the binding affinities. Molecular dynamics simulations have also been used to elucidate the effects of the protease inhibitor resistance mutations D30N, I50V, I54M, and V82A, providing interesting mechanistic information on how these mutations alter binding affinities, including changes in the binding conformation (I50V), conformational changes (I54M) and large enthalpic changes reducing binding affinity (V82A) [47]. While genomic methods have proven unreliable for phenotypic characterization of HIV [48], this potentially offers a means to better leverage this information and suggests ways to guide new designs that avoid these common hotspots.

The last HIV protease inhibitor approved, darunavir, was designed with this in mind and is capable of inhibiting the replication of both wild-type and multidrug-resistant strains of HIV-1. While earlier inhibitors interacted with the side-chains of Asp-28 and Asp-30, darunavir contained a *bis*-tetrahydrofuranylurethane functional group that made close, tight interactions with the main chain of these residues, making only

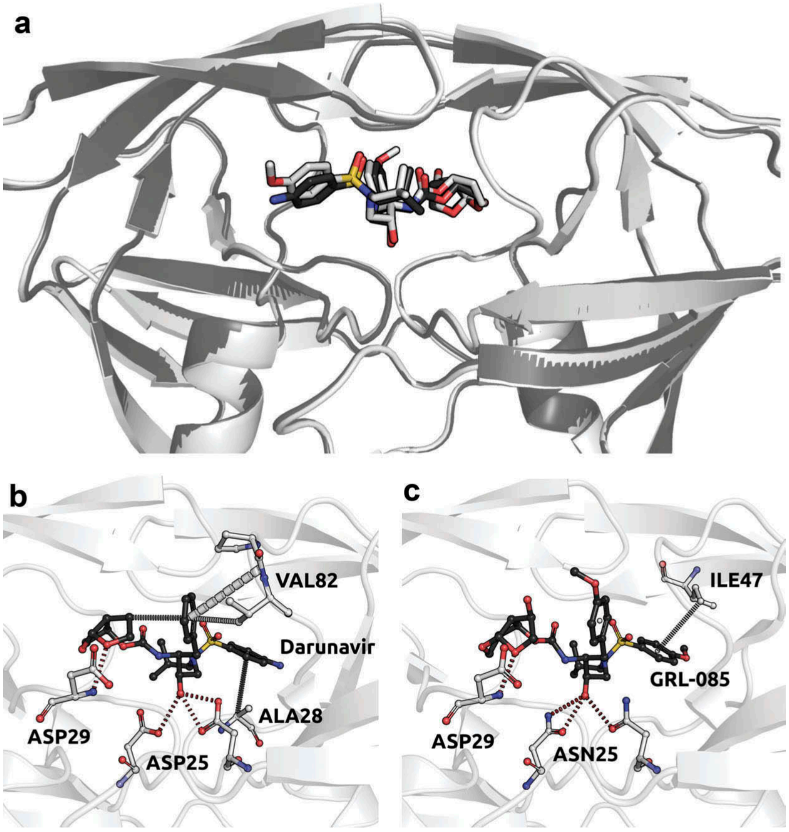


Figure 2. HIV-1 protease in complex with the non-peptidic inhibitor GRL-085 and darunavir (PDB: 5COO and 4HLA, respectively). (a) Shows the two aligned structures of HIV-1 protease in complex with GRL-085 (light gray) and darunavir (dark gray). (b) Depicts the main interactions between the key residues of the binding site of HIV-1 protease and darunavir. (c) Shows the interactions between GRL-085 and the wild-type protease, calculated by Arpeggio [50].

minimal interactions with the side chains [49]. This made darunavir less sensitive to substitutions in either of these positions. Figure 2(a) depicts an alignment between darunavir and a non-peptidic inhibitor GRL-085 and the interactions made by the inhibitors (Figure 2(b,c), respectively).

Many resistant strains against darunavir, however, have emerged. These mutations often lead to a change in the conformation of the active site residues, reducing affinity for darunavir, but also leading to a significant fitness cost [51]. In the effort to avoid these resistance mutations, current medicinal chemistry efforts have identified potent inhibitors that differ from the currently approved protease inhibitors by the number and proximity of contacts to the main chains of these catalytic amino acids [49]. These compounds will be hopefully even more effective therapeutics that are significantly less prone to develop resistance.

4.2. Influenza neuraminidase inhibitors

Influenza neuraminidase inhibitors (NAIs) are the major specific anti-influenza drugs used clinically, despite the emergence of resistance [52]. Currently, the NAIs oseltamivir, zanamivir, peramivir, and laninamivir (currently approved only in Japan) have been approved to prevent and treat influenza A and B [52–55]. Many governments have stockpiled resources of these drugs in the event of an Influenza outbreak. During the recent H1N1 and H7N9 influenza outbreaks, significant resources were focused on identifying and monitoring potential resistance mutations, primarily through genetic screening, with sporadic oseltamivir-resistant 2009 H1N1 virus infections identified. Thus, understanding the mechanisms of influenza NA drug resistance is crucial to develop drugs that can get around mutations and be more successful to fight the epidemics and pandemics [52].

A strong correlation has been observed between mutations that affect the slow binding and dissociation of these NAIs, and the association with resistance [56]. Resistance mutations that have been observed to residues E119 and I222 of Influenza A lead to high and slight resistance to oseltamivir and zanamivir, respectively [57]. Figure 3(a,b) highlight these resistance hotspots on the solved complex of the neuraminidase with oseltamivir and the interactions established on the wild-type protein. Mutations on E119, include substitutions to Gly, Asp, Ala, Ile, and Val, lead to the loss of a salt bridge to the inhibitors [58], with zanamivir showing less susceptibility due to the presence of the 4-guanidino group that maintains typical interactions [52].

Mutations at I222 alter the hydrophobic drug-binding pocket. While I222R leads to a reduction in oseltamivir, peramivir, and zanamivir effectiveness [53,59,60], the I222L mutation, which is also found in Influenza B, has been reported to not lead to significant drug resistance [52]. The other common mutation in N2 is R292K, which leads to resistance against

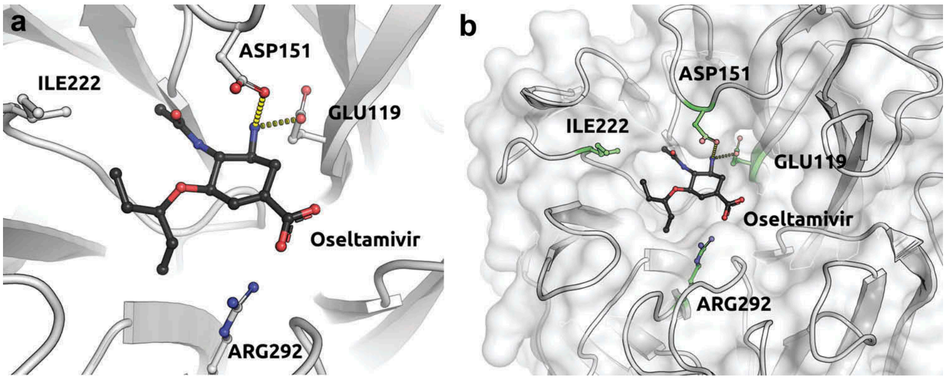


Figure 3. Neuraminidase subtype 2 of Influenza A in complex with Oseltamivir (PDB: 4GZP). (a) Shows the main resistance hot-spot residues Glu119, Asp151 and Ile222 shown as sticks. The two negatively charged residues interact with Oseltamivir via ionic interactions shown as dashes, as calculated by Arpeggio [50]. Arg292, another important binding residue is also shown. (b) Shows the four aforementioned residues and the oseltamivir molecule in a surface perspective.

oseltamivir and peramivir and a slight reduction of zanamivir and laninamivir effectiveness [53].

Following treatment with oseltamivir, the N1 subtype-specific substitution H274Y has also been observed, leading to resistance to this drug and also peramivir, but not to zanamivir and laninamivir [61,62]. The change in volume of the side chains upon this mutation causes the carbonyl group of E276 to be shifted into the binding site of the enzyme, disturbing the hydrophobic pocket that would accommodate the pentyloxy group of oseltamivir [62].

Therefore in efforts to overcome some of these resistance problems, the guanidino group of zanamivir and the hydrophobic pentyloxy group of oseltamivir were merged [61]. The guanidino group was capable of inhibiting the spread of Influenza A with the hydrogen bond interactions between the guanidino group and neuraminidase binding site crucial for the inhibition of the enzyme and virus replication [62,63]. However, the inhibition profile of MS-257 and zanamivir was comparable against the E119V and I222L mutant strains [52].

The sequence database compiled by the WHO containing lists of amino acid substitutions in the neuraminidase has been widely used to identify key mutations and regions, guiding genomic analysis of resistance and proving invaluable for testing new compounds targeting inhibition of neuraminidase [64,65]. It has also facilitated the use of next-generation sequencing to detect resistance markers in the NA gene and predict the effect of drug treatment [66], which have been complemented by the use of structural-based approaches to identify likely resistance mutations.

4.3. Kinase drug development

4.3.1. Kinase inhibition

Abnormal regulation of kinases through occurrence of mutations is responsible for many human diseases, including metabolic disorders and certain types of cancer [67]. The development of small-molecule kinase inhibitors has therefore been seen as an attractive treatment option [68]. Unlike conventional chemotherapy (cytotoxic), molecular targeted therapies using kinase inhibitors are designed to act at specific biological points that are essential for development of tumor cells [69].

The design of kinase inhibitors has great impact on their efficacy and sensitivity to resistance. The first kinase inhibitors developed targeted the ATP-binding site via competitive binding. As resistance to these inhibitors was identified, other strategies including allosteric and covalently bound inhibitors were used to avoid these common resistance mutations [68].

4.3.2. ATP-competitive inhibitors –first generation

ATP-competitive kinase inhibitors inhibit ATP binding in the catalytic site of the target kinase, or bind at alternative sites to induce conformational molecular changes that inhibit the activity of the enzyme [69]. Imatinib was the first kinase small-molecule inhibitor clinically approved by the US Food and Drug Administration (FDA) for treatment of chronic myeloid leukemia [70]. Imatinib binds to the active site of the target enzyme preventing other substrates from phosphorylation and consequently inhibiting kinase activity. Figure 4(a) shows the Abelson tyrosine-protein kinase 2 (ABL2) in complex with imatinib. The inhibitor only binds to the enzyme when it is in inactive conformation. Another example of an inhibitor with a mechanism similar to imatinib is gefitinib which is used for treatment of non-small-cell lung cancer through inhibition of the epidermal growth factor receptor (EGFR).

Despite the success of imatinib, studies have shown that patients can develop resistance and relapse after initial response to therapy. The effect of mutations linked to imatinib resistance were analyzed by mCSM-Lig [14], which could correctly identify resistance mutations located even quite distal from the active site. mCSM-Lig quantitatively predicts the effect of mutations on small molecule affinity. Resistance mutations of competitive inhibitor, however, can exist by shifting the preference of the protein toward the natural ligand (ATP), not necessarily by dramatically reducing the affinity of the protein to the drug. Interestingly, using a fold-ratio between the predicted affinity effect on the natural ligand and the drug, mCSM-Lig was successful in identifying the majority of the imatinib resistance mutations.

Several mechanisms of resistance have been observed, including mutations in the BCR-ABL kinase domain, with the most common resistant observed the gatekeeper mutant T315I [71]. This amino acid substitution eliminates a critical

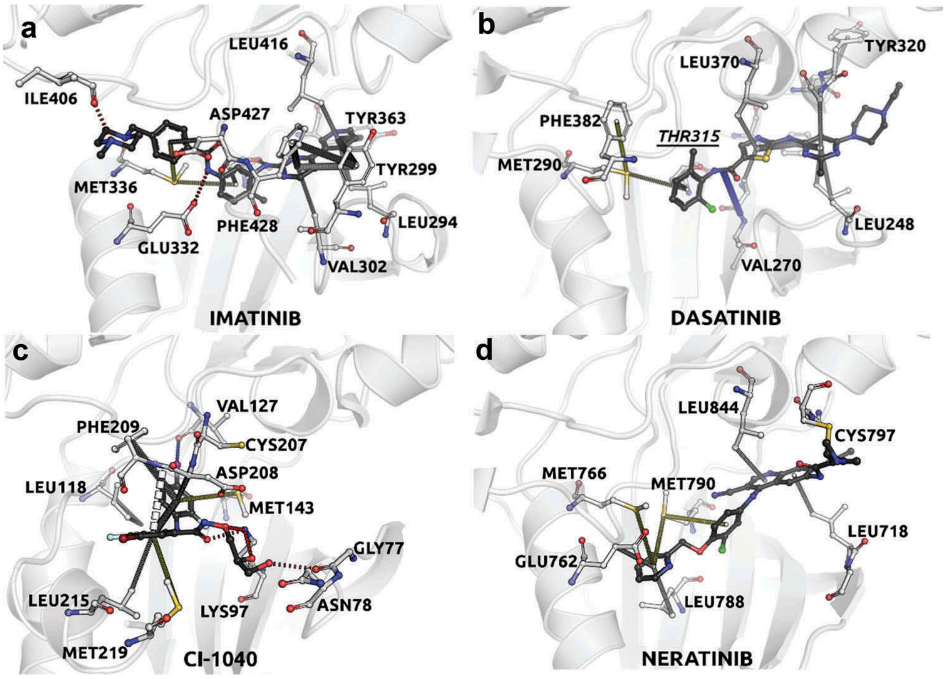


Figure 4. Four generations of kinase inhibitors. (a) Shows ABL2 in complex with first generation kinase inhibitor Imatinib (PDB: 3GVU). Imatinib binds to the active site of the enzyme preventing other substrates from phosphorylation only when the ABL2 is in inactive conformation. (b) Shows ABL1 in complex with second generation inhibitor Dasatinib (PDB: 2GQG). Dasatinib is a multitargeted tyrosine kinase inhibitor more potent than Imatinib due to its capability of binding to the enzyme in inactive imatinib-bound conformation, also effective against several imatinib-resistant mutations, except for T315I gatekeeper mutation as a result of a crucial hydrogen bond with T315 (underlined) for the stabilization of the complex. (c) Shows MEK1 in complex with CI-1040 allosteric kinase inhibitor adjacent to the ATP binding site of the enzyme (PDB: 1S9J). The third generation of kinase inhibitors can bind either to the kinase domain or to other sites giving them clear advantage over ATP-competitive in first and second generation. (d) Shows EGFR mutant T790M/L858R in complex with fourth generation kinase inhibitor Neratinib (PDB: 3W2Q). Unlike first and second generation inhibitors, this fourth generation inhibitor binds covalently to the kinase active site, blocking ATP binding.

oxygen molecule needed for hydrogen bonding between imatinib and the ABL kinase, and also introduces a steric clash preventing drug binding. The gatekeeper residue determines the relative accessibility of a hydrophobic pocket located adjacent to the ATP-binding site, which is important for imatinib binding given that hydrophobic interactions are crucial for inhibitor binding affinity [68,72,73]. In fact, mutations in gatekeeper residues have also been studied for other kinases in different types of cancer, such as the Threonine 790 of EGFR in Lung cancer that mutates to a methionine (T790M) increasing the affinity for ATP and making it difficult for the gefitinib to compete for the binding site [74–76]. Such mechanisms of resistance have contributed to the development of more sophisticated generations of inhibitors with mechanisms to overcome resistances conferred by these gatekeeper mutations.

4.3.3. ATP-competitive inhibitors –second generation

The second generation of small-molecule kinase inhibitors preferentially binds to regions outside the ATP-binding site, for example, to the inactive conformation, also known as DFG-out, of the protein kinase. The transition from the active conformation to DFG-out conformation exposes additional hydrophobic pockets adjacent to the ATP site that can be used by the inhibitors to stabilize the kinase in its inactive conformation [77], preventing ATP binding.

Dasatinib is a multitargeted tyrosine kinase inhibitor that targets oncogenic pathways and is a more potent inhibitor than imatinib that binds only when the ABL enzyme is in its inactive conformation. Dasatinib is also effective against several imatinib-resistant ABL mutations that occur in regions that are in contact with imatinib or mutations involved in stabilization of specific inactive imatinib-bound conformation of the enzyme. However, the T315I gatekeeper mutation is also resistant to dasatinib due crucial hydrogen bond with the T315 side chain [78]. Figure 4(b) shows ABL1 in complex with dasatinib. The main residues involved in the binding of the drug are highlighted, including T315.

4.3.4. Allosteric inhibitors – third generation

These inhibitors regulate the kinase activity in an allosteric manner, exhibiting a higher degree of selectivity due the exploitation of binding sites and regulatory mechanisms that are specific to a particular kinase [68]. Figure 4(c) shows the allosteric inhibitor CI-1040 binding MEK1 immediately adjacent to the ATP binding site.

This class of inhibitors can bind either to the kinase domain (or close to the ATP binding site) or to sites outside the kinase domain. These range of options for inhibiting the catalytic activity of kinases represent clear advantages over the ATP-competitive inhibitors [79,80]. However, the lack of methods to identify such inactive conformations or binding modes in kinases to drive the development of this type of inhibitor still remains a challenge [81]. Inhibitors that disrupt formation of the higher order oligomers, which play an important role in achieving high signal-to-noise throughout the signal transduction process, have also proven to be effective kinase inhibitors that avoid the common ATP resistance mutations [82–84].

ABL001, also known as Asciminib, is a potent and selective third generation kinase inhibitor with activity against chronic myeloid leukemia and Philadelphia chromosome-positive (Ph +) acute lymphoblastic leukemia. ABL001 binds to the myristoyl pocket of ABL1 kinase leading to a formation of an inactive kinase conformation [85]. Recent studies have shown that treatment with ABL001 combined with ATP-competitive inhibitors can help prevent resistance in chronic myeloid leukemia [86,87].

4.3.5. Covalent inhibitors – fourth generation

Recent studies [88,89] described a fourth class of kinase inhibitors that are capable of forming covalent bonds to the kinase active site, most frequently by reacting with a nucleophilic cysteine residue. Unlike first- and second-generation inhibitors, the fourth generation blocks the binding of ATP irreversibly preventing the kinase from being activated. Figure 4(d) shows the fourth-generation inhibitor Neratinib (HKI-272) in complex with EGFR kinase T790M mutant, making a covalent bond to Cysteine 797.

4.3.6. Tackling kinase inhibitor resistance

Much of the effort to target and avoid resistance against common kinase inhibitors has focused on the development of inhibitors with different modes of action. This has in part been driven by the lack of selectivity of the early inhibitors that targeted the ATP-binding site – which is highly conserved among many proteins. Structural methods such as mCSM-lig and molecular dynamics approaches have been able to correctly identify and predict likely resistance mutations, which could also potentially facilitate the design of new inhibitors avoiding these resistance hotspots, similar to the efforts in antiviral inhibitor design. However, more practically, as sequencing of cancers is becoming more routine, these methods offer the opportunity to help guide the selection of the most effective therapeutics- facilitating the widespread implementation of personalized medicine.

The advent of fast and precise computational methods to predict effect of mutations can be leveraged to assist and guide the development of new drugs. Since resistance can emerge from different molecular mechanisms, current predictors can be integrated in novel drug resistance identification methods that can then be used in large-scale screening to identify better protein targets, identify and avoid potential resistance hotspots as well as optimize ligand affinity and selectivity, driving the experimental design of better, more potent and efficacious drugs.

5. Expert opinion

While significant progress has been made in terms of innovative tools to understand and quantify the different range of effects in which a mutation or a set of mutations can give rise to a phenotype, a great gap still exists when integrating these predictions and drawing causality conclusions linking variants, compounded by the need for detailed information regarding the system/protein. The availability of scalable, effective computational methods to assess mutation effects creates new opportunities of development of such integrated approaches and decipher complex genomic background patterns, shedding light into their role in the emergence of a given phenotype and molecular mechanisms of action. This capability can then be used to systematically study, for instance, how drug resistance emerges on specific drug targets, aiding the drug development process. Initial efforts on that matter have focused on preparing predictors and databases for specific diseases and proteins; however, greater effort needs to be invested in making these predictors user friendly, integrated, and accessible to geneticists. This is particularly important considering that most structural information is a snapshot of a protein conformation, but how mutations affect the equilibrium between different states can play a very important role in disease and drug resistance [90]. A complementary and important effort refers to the collection and curation of experimental data regarding mutation effects linked to phenotype in comprehensive databases. This information forms the evidence set necessary for the proposal of novel computational methods as well as the improvement of current approaches. Initiatives like the Platinum database [91], the first curated online database linking effects of mutations on protein-small-molecule affinity for complexes with known structures, are fundamental.

Despite this limitation, these methodologies have already provided invaluable insights into many diseases. Current genomic analyses are dependent upon preexisting information; either extensive genomic or biochemical analyses. This limits the insight and information that can be drawn regarding novel mutations. As these structural methods become more widely used, they will complement traditional analyses methods to provide much greater power from genomic analysis.

In the shorter term, the ability of these methods to predict likely resistance mutations before they arise offers enormous potential throughout the drug development process. Peter Coleman first suggested that the design of inhibitors that resemble transition state analogs should be more resilient to the development of resistance. Out of this, Zanamivir was developed, the first successful structure guided drug development, but as we have seen over the intervening years resistance against Relenza has been widely reported, although it has been less prone to resistance than Oseltamivir.

During the development of a recent class of *Mycobacterium tuberculosis* IMPDH inhibitors, structural-guided mutational prediction was used to identify likely resistance mutations, defined in this case as point mutations that disrupted inhibitor binding, but did not affect NAD binding, protein solubility or formation of the active tetramer. One mutation in particular, Y487C, was highlighted, and subsequently confirmed to be one of the few mutations to arise during resistance screening [92]. Subsequent drug development attempts avoided this resistance hotspot and were active against the Y487C mutant [93]. This also enables the analysis of multiple mutations, some of which have been characterized to facilitate the development of resistance. In many cases, these seem to increase protein stability or natural ligand binding, which can be decreased due to the primary resistance mutation.

While current medicinal chemistry efforts are currently normally retroactive – we observe which mutations arise in the lab or clinic and then design new generations of inhibitors to target or avoid them – the power of computational mutational analysis enables us to preemptively identify likely resistance hotspots, and to take this information under consideration when optimizing candidate molecules. In a similar fashion to how experimental structures [94–98] and pharmacokinetic predictors are now widely used to guide medicinal chemistry efforts [99], playing a role in dramatically reducing failure rates of clinical trials due to these problems. The use of *in silico* mutational analysis in the development of new therapeutics will hopefully avoid likely resistance mutations. While the evolutionary forces and the constant selective battle makes the development of resistance somewhat inevitable, this will hopefully aid in the development of the next generation of therapeutics that are more resistant to the development of resistance.

Funding

This work was funded by the Jack Brockhoff Foundation (JBF 4186, 2016) and a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (MR/M026302/1). This research was supported by the Victorian Life Sciences Computation Initiative (VLSCI), an initiative of the Victorian Government, Australia, on its Facility hosted at the University of Melbourne (UOM0017). DEV Pires receives support from the René Rachou Research Center (CPqRR/FIOCRUZ Minas), Brazil. DB Ascher is supported by a C. J. Martin Research Fellowship from the National Health and Medical Research Council of Australia (APP1072476), and the Department of Biochemistry, University of Melbourne.

Declaration of interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

ORCID

David B. Ascher http://orcid.org/0000-0003-2948-2413

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–1081.
2. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7 (4):248–249.
3. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar;46(3):310–315.
4. Schwarz JM, Rodelsperger C, Schuelke M, et al. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010 Aug;7(8):575–576.
5. Rethink the links between genes and disease. Nature. 2016 Oct 13;538(7624):140.
6. Topham CM, Srinivasan N, Blundell TL. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Eng. 1997 Jan;10(1):7–21.
7. Pires DE, De Melo-Minardi RC, Dos Santos MA, et al. Cutoff scanning matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. BMC Genomics. 2011 Dec 22;12(Suppl 4):S12.
8. Pires DE, De Melo-Minardi RC, Da Silveira CH, et al. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. Bioinformatics. 2013 Apr 01;29(7):855–861.
9. Pires DE, Ascher DB. mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. Nucleic Acids Res. 2016 Jul 08; 44(W1):W469–73.
   - • **A high-throughput and accurate method to predict the effects of mutations on antibody-antigen binding affinity. Used for antibody maturation and to predict likely antibody escape mutations.**
10. Pires DE, Ascher DB. CSM-lig: a web server for assessing and comparing protein-small molecule affinities. Nucleic Acids Res. 2016 Jul 08;44(W1):W557–61.
11. Pires DE, Ascher DB. mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res. DOI: 10.1093/nar/gkx236.
   - • **Optimized method to predict the effect of mutations on protein-nucleic acid binding.**
12. Pires DE, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W314–9.
   - • **An integrated structural method to predict effects of mutations on protein stability.**
13. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014 Feb 01; 30(3):335–342.
   - •• **Comprehensive platform for analysis of the effects of mutations on protein structure and function, including the first published methods to assess the affects of mutations on protein-protein and protein-nucleic acid binding affinity.**
14. Pires DE, Blundell TL, Ascher DB. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep. 2016 Jul 07;6:29575.
   - •• **The first scalable and accurate method to predict the effects of single-point mutations on protein-small-molecule interactions. This was capable of identifying and anticipating drug resistance mutations.**
15. Andrews KA, Vialard L, Ascher DB, et al. Tumour risks and genotype–phenotype–proteotype analysis of patients with germline mutations in the succinate dehydrogenase subunit genes SDHB, SDHC, and SDHD. Lancet. 2016 Feb 25;387:S19.
16. Gossage L, Pires DE, Olivera-Nappa A, et al. An integrated computational approach can classify VHL missense mutations according to risk of clear cell renal carcinoma. Hum Mol Genet. 2014 Nov 15;23(22):5976–5988.
17. Nemethova M, Radvanszky J, Kadasi L, et al. Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. Eur J Hum Genet. 2016 Jan;24(1):66–72.
18. Usher JL, Ascher DB, Pires DE, et al. Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: identification of novel mutations. JIMD Rep. 2015 Feb;15(24):3–11.
19. Kano FS, Souza-Silva FA, Torres LM, et al. The presence, persistence and functional properties of plasmodium vivax duffy binding protein II antibodies are Influenced by HLA class II allelic variants. Plos Negl Trop Dis. 2016 Dec;10(12):e0005177.
20. Silvino AC, Costa GL, Araujo FC, et al. Variation in human cytochrome P-450 drug-metabolism genes: a gateway to the understanding of plasmodium vivax relapses. Plos One. 2016;11(7):e0160172.
21. White RR, Ponsford AH, Weekes MP, et al. Ubiquitin-dependent modification of skeletal muscle by the parasitic nematode, trichinella spiralis. Plos Pathog. 2016 Nov;12(11):e1005977.

22. Pires DE, Chen J, Blundell TL, et al. In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep. 2016 Jan 22;6:19848.

•• **An integrated pipeline using changes in protein structure and function to elucidate the relationship between genotype and phenotype.**

23. Worth CL, Preissner R, Blundell TL. SDM–a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W215–22.

24. Pandurangan AP, Ochoa-Montaño B, Ascher DB, et al. SDM: a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res. Fourth coming.

25. Dehouck Y, Kwasigroch JM, Gilis D, et al. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics. 2011 May 13;12:151.

26. Frappier V, Chartier M, Najmanovich RJ. ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. Nucleic Acids Res. 2015 Jul 01;43(W1): W395–400.

27. Laimer J, Hiebl-Flach J, Lengauer D, et al. MAESTROweb: a web server for structure-based protein stability prediction. Bioinformatics. 2016 May 01;32(9):1414–1416.

28. Quan L, Lv Q, Zhang Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics. 2016 Oct 01;32(19):2936–2946.

29. Witvliet DK, Strokach A, Giraldo-Forero AF, et al. ELASPIC webserver: proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. Bioinformatics. 2016 May 15;32(10):1589–1591.

30. Dehouck Y, Kwasigroch JM, Rooman M, et al. BeAtMuSiC: prediction of changes in protein-protein binding affinity on mutations. Nucleic Acids Res. 2013 Jul;41(Web Server issue): W333–9.

31. Li M, Simonetti FL, Goncearenco A, et al. MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions. Nucleic Acids Res. 2016 Jul 08;44(W1):W494–501.

32. Jafri M, Wake NC, Ascher DB, et al. Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. Cancer Discov. 2015 Jul;5(7):723–729.

33. Casey RT, Ascher DB, Rattenberry E, et al. SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. Mol Genet Genomic Med. DOI: 10.1002/mgg3.279.

34. Paez JG, Janne PA, Lee JC, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004 Jun 04;304(5676):1497–1500.

35. Lievre A, Bachet JB, Le Corre D, et al. KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. Cancer Res. 2006 Apr 15;66(8):3992–3995.

36. Jubb HC, Pandurangan AP, Turner MA, et al. Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol. DOI: 10.1016/j.pbiomolbio.2016.10.002.

37. Zatkova A, Sedlackova T, Radvansky J, et al. Identification of 11 novel homogentisate 1,2 dioxygenase variants in alkaptonuria patients and establishment of a novel LOVD-based HGD mutation database. JIMD Rep. 2012;4:55–65.

38. Poduri A, Sheidley BR, Shostak S, et al. Genetic testing in the epilepsies-developments and dilemmas. Nat Rev Neurol. 2014 May;10(5):293–299.

39. Soardi FC, Machado-Silva A, Linhares ND, et al. Familial STAG2 germline mutation defines a new human cohesinopathy. Npj Genomic Medicine. 2017;2(1):7.

40. Phelan J, Coll F, McNerney R, et al. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med. 2016 Mar 23;14(1):31.

• **Structural characterization of *Mycobacterium Tuberculosis* drug resistance mutations, highlighting the strong correlation between MIC and structural effects.**

41. Ascher DB, Jubb HC, Pires DE, et al. Protein-protein interactions: structures and druggability. In: Scapin G, Patel D, Arnold E eds. Multifaceted roles of crystallography in modern drug discovery. Netherlands: Springer; 2015. p. 141–163.

42. Pandurangan AP, Ascher DB, Thomas SE, et al. Genomes, structural biology, and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochem Soc Trans. 2017;45(2):303–311. DOI: 10.1042/BST20160422

43. Park JH, Sayer JM, Aniana A, et al. Binding of clinical inhibitors to a model precursor of a rationally selected multidrug resistant HIV-1 protease is significantly weaker than that to the released mature enzyme. Biochemistry. 2016 Apr 26;55(16):2390–2400.

44. Zhang H, Wang YF, Shen CH, et al. Novel P2 tris-tetrahydrofuran group in antiviral compound 1 (GRL-0519) fills the S2 binding pocket of selected mutants of HIV-1 protease. J Med Chem. 2013 Feb 14;56(3):1074–1083.

45. Koh Y, Amano M, Towata T, et al. In vitro selection of highly darunavir-resistant and replication-competent HIV-1 variants by using a mixture of clinical HIV-1 isolates resistant to multiple conventional protease inhibitors. J Virol. 2010 Nov;84(22):11961–11969.

46. Hosseini A, Alibes A, Noguera-Julian M, et al. Computational prediction of HIV-1 resistance to protease inhibitors. J Chem Inf Model. 2016 May 23;56(5):915–923.

47. Hu G, Ma A, Dou X, et al. Computational studies of a mechanism for binding and drug resistance in the wild type and four mutations of HIV-1 protease with a GRL-0519 inhibitor. Int J Mol Sci. 2016 May 27;17(6).

48. Hanna GJ, D'Aquila RT. Clinical use of genotypic and phenotypic drug resistance testing to monitor antiretroviral chemotherapy. Clin Infect Dis. 2001 Mar 01;32(5):774–782.

49. Koh Y, Nakata H, Maeda K, et al. Novel bis-tetrahydrofuranylurethane-containing nonpeptidic protease inhibitor (PI) UIC-94017 (TMC114) with potent activity against multi-PI-resistant human immunodeficiency virus in vitro. Antimicrob Agents Chemother. 2003 Oct;47(10):3123–3129.

50. Jubb HC, Higueruelo AP, Ochoa-Montano B, et al. Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. J Mol Biol. 2017 Feb 03;429(3):365–371.

•• **Novel method for calculating and displaying all possible intra- and inter-molecular interactions.**

51. Yoshimura K, Kato R, Kavlick MF, et al. A potent human immunodeficiency virus type 1 protease inhibitor, UIC-94003 (TMC-126), and selection of a novel (A28S) mutation in the protease active site. J Virol. 2002 Feb;76(3):1349–1358.

52. Wu Y, Gao F, Qi J, et al. Resistance to mutant group 2 influenza virus neuraminidases of an oseltamivir-zanamivir hybrid inhibitor. J Virol. 2016 Dec 01;90(23):10693–10700.

53. Yen HL. Current and novel antiviral strategies for influenza infection. Curr Opin Virol. 2016 Jun;18:126–134.

54. Hata A, Akashi-Ueda R, Takamatsu K, et al. Safety and efficacy of peramivir for influenza treatment. Drug Des Devel Ther. 2014;8:2017–2038.

55. Moscona A. Neuraminidase inhibitors for influenza. N Engl J Med. 2005 Sep 29;353(13):1363–1373.

56. McKimm-Breschkin JL, Barrett S. Neuraminidase mutations conferring resistance to laninamivir lead to faster drug binding and dissociation. Antiviral Res. 2015 Feb;114:62–66.

57. Richard M, Ferraris O, Erny A, et al. Combinatorial effect of two framework mutations (E119V and I222L) in the neuraminidase active site of H3N2 influenza virus on resistance to oseltamivir. Antimicrob Agents Chemother. 2011 Jun;55(6):2942–2952.

58. Okomo-Adhiambo M, Demmler-Harrison GJ, Deyde VM, et al. Detection of E119V and E119I mutations in influenza A (H3N2) viruses isolated from an immunocompromised patient: challenges in diagnosis of oseltamivir resistance. Antimicrob Agents Chemother. 2010 May;54(5):1834–1841.

59. Van Der Vries E, Stelma FF, Boucher CA. Emergence of a multidrug-resistant pandemic influenza A (H1N1) virus. N Engl J Med. 2010 Sep 30;363(14):1381–1382.
60. McKimm-Breschkin JL. Influenza neuraminidase inhibitors: antiviral action and mechanisms of resistance. Influenza Other Respir Viruses. 2013 Jan;7(Suppl 1):25–36.
61. Kerry PS, Mohan S, Russell RJ, et al. Structural basis for a class of nanomolar influenza A neuraminidase inhibitors. Sci Rep. 2013 Oct;16(3):2871.
62. Collins PJ, Haire LF, Lin YP, et al. Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants. Nature. 2008 Jun 26;453(7199):1258–1261.
63. Niikura M, Bance N, Mohan S, et al. Replication inhibition activity of carbocycles related to oseltamivir on influenza A virus in vitro. Antiviral Res. 2011 Jun;90(3):160–163.
64. Meijer A, Rebelo-de-Andrade H, Correia V, et al. Global update on the susceptibility of human influenza viruses to neuraminidase inhibitors, 2012-2013. Antiviral Res. 2014;110:31–41.
65. Hurt AC, Besselaar TG, Daniels RS, et al. Global update on the susceptibility of human influenza viruses to neuraminidase inhibitors, 2014-2015. Antiviral Res. 2016;132:178–185.
66. Parker J, Chen J. Application of next generation sequencing for the detection of human viral pathogens in clinical specimens. J Clin Virol. 2017 Jan;86:20–26.
67. Lahiry P, Torkamani A, Schork NJ, et al. Kinase mutations in human disease: interpreting genotype-phenotype relationships. Nat Rev Genet. 2010 Jan;11(1):60–74.
68. Zhang J, Yang PL, Gray NS. Targeting cancer with small molecule kinase inhibitors. Nat Rev Cancer. 2009 Jan;9(1):28–39.
69. Gharwan H, Groninger H. Kinase inhibitors and monoclonal antibodies in oncology: clinical implications. Nat Rev Clin Oncol. 2016 Apr;13(4):209–227.
70. Agafonov RV, Wilson C, Kern D. Evolution and intelligent design in drug development. Front Mol Biosci. 2015;2:27.
71. Weisberg E, Manley P, Mestan J, et al. AMN107 (nilotinib): a novel and selective inhibitor of BCR-ABL. Br J Cancer. 2006 Jun 19;94(12):1765–1769.
72. Azam M, Seeliger MA, Gray NS, et al. Activation of tyrosine kinases by mutation of the gatekeeper threonine. Nat Struct Mol Biol. 2008 Oct;15(10):1109–1118.
73. Engelman JA, Janne PA. Mechanisms of acquired resistance to epidermal growth factor receptor tyrosine kinase inhibitors in non-small cell lung cancer. Clin Cancer Res. 2008 May 15;14(10):2895–2899.
74. Tetsu O, Hangauer MJ, Phuchareon J, et al. Drug resistance to EGFR inhibitors in lung cancer. Chemotherapy. 2016;61(5):223–235.
75. Klebl BM, Muller G. Second-generation kinase inhibitors. Expert Opin Ther Targets. 2005 Oct;9(5):975–993.
76. Yun CH, Mengwasser KE, Toms AV, et al. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. Proc Natl Acad Sci U S A. 2008 Feb 12;105(6):2070–2075.
77. Cowan-Jacob SW, Jahnke W, Knapp S. Novel approaches for targeting kinases: allosteric inhibition, allosteric activation and pseudokinases. Future Med Chem. 2014 Apr;6(5):541–561.
78. Tokarski JS, Newitt JA, Chang CY, et al. The structure of Dasatinib (BMS-354825) bound to activated ABL kinase domain elucidates its inhibitory activity against imatinib-resistant ABL mutants. Cancer Res. 2006 Jun 01;66(11):5790–5797.
79. Fasano M, Della Corte CM, Califano R, et al. Type III or allosteric kinase inhibitors for the treatment of non-small cell lung cancer. Expert Opin Investig Drugs. 2014 Jun;23(6):809–821.
80. Wu P, Clausen MH, Nielsen TE. Allosteric small-molecule kinase inhibitors. Pharmacol Ther. 2015 Dec;156:59–68.
81. Muller S, Chaikuad A, Gray NS, et al. The ins and outs of selective kinase inhibitor development. Nat Chem Biol. 2015 Nov;11(11):818–821.
82. Blaszczyk M, Harmer NJ, Chirgadze DY, et al. Achieving high signal-to-noise in cell regulatory systems: spatial organization of multiprotein transmembrane assemblies of FGFR and MET receptors. Prog Biophys Mol Biol. 2015 Sep;118(3):103–111.
83. Liang S, Esswein SR, Ochi T, et al. Achieving selectivity in space and time with DNA double-strand-break response and repair: molecular stages and scaffolds come with strings attached. Struct Chem. 2016;28(1):161–171.
84. Sibanda BL, Chirgadze DY, Ascher DB, et al. DNA-PKcs structure suggests an allosteric mechanism modulating DNA double-strand break repair. Science. 2017 Feb 03;355(6324):520–524.
85. Wylie AA, Schoepfer J, Jahnke W, et al. The allosteric inhibitor ABL001 enables dual targeting of BCR-ABL1. Nature. 2017 Mar 30;543(7647):733–737.
86. Wylie A, Schoepfer J, Berellini G, et al. ABL001, a potent allosteric inhibitor of BCR-ABL, prevents emergence of resistant disease when administered in combination with nilotinib in an in vivo murine model of chronic myeloid leukemia. Blood. 2014;124(21):398–398.
87. Eide CA, Savage SL, Heinrich MC, et al. Combining the allosteric ABL1 tyrosine kinase inhibitor ABL001 with ATP-competitive inhibitors to suppress resistance in chronic myeloid leukemia. Blood. 2016;128(22):2747–2747.
88. Tan L, Wang J, Tanizaki J, et al. Development of covalent inhibitors that can overcome resistance to first-generation FGFR kinase inhibitors. Proc Natl Acad Sci U S A. 2014 Nov 11;111(45):E4869–77.
89. Zou Y, Xiao J, Tu Z, et al. Structure-based discovery of novel 4,5,6-trisubstituted pyrimidines as potent covalent Bruton's tyrosine kinase inhibitors. Bioorg Med Chem Lett. 2016 Jul 01;26(13):3052–3059.
90. Ascher DB, Wielens J, Nero TL, et al. Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. Sci Rep. 2014 Apr 23;4:4765.
91. Pires DE, Blundell TL, Ascher DB. Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res. 2015 Jan;43 (Database issue):D387–91.
92. Singh V, Donini S, Pacitto A, et al. The inosine monophosphate dehydrogenase, GuaB2, is a vulnerable new bactericidal drug target for tuberculosis. ACS Infect Dis. 2017 Jan 13;3(1):5–17.
93. Park Y, Pacitto A, Bayliss T, et al. Essential but not vulnerable: indazole sulfonamides targeting inosine monophosphate dehydrogenase as potential leads against mycobacterium tuberculosis. ACS Infect Dis. 2017 Jan 13;3(1):18–33.
94. Ascher DB, Cromer BA, Morton CJ, et al. Regulation of insulin-regulated membrane aminopeptidase activity by its C-terminal domain. Biochemistry. 2011 Apr 05;50(13):2611–2622.
95. Hermans SJ, Ascher DB, Hancock NC, et al. Crystal structure of human insulin-regulated aminopeptidase with specificity for cyclic peptides. Protein Sci. 2015 Feb;24(2):190–199.
96. Chai SY, Yeatman HR, Parker MW, et al. Development of cognitive enhancers based on inhibition of insulin-regulated aminopeptidase. BMC Neurosci. 2008 Dec 03;9(Suppl 2):S14.
97. Albiston AL, Morton CJ, Ng HL, et al. Identification and characterization of a new cognitive enhancer based on inhibition of insulin-regulated aminopeptidase. Faseb J. 2008 Dec;22(12):4209–4217.
98. Sigurdardottir AG, Winter A, Sobkowicz A, et al. Exploring the chemical space of the lysine-binding pocket of the first kringle domain of hepatocyte growth factor/scatter factor (HGF/SF) yields a new class of inhibitors of HGF/SF-MET binding. Chem Sci. 2015;6(11):6147–6157.
99. Pires DE, Blundell TL, Ascher DB. pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem. 2015 May 14;58(9):4066–4072.

FIOCRUZ MINAS
Centro de Pesquisas René Rachou

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

CPqRR - Fiocruz Minas

kamp: a structure-based computational approach for predicting activating mutations in kinases

Carlos H. M. Rodrigues[1,2], David B. Ascher[1,3,4], Douglas E. V. Pires[1]

[1]René Rachou Research Center,- CPqRR, Fiocruz -- Minas/Brazil - [3]Department of Biochemistry,- University of Cambridge/United Kingdom
[2]Department of Biochemistry and Immunology, -Universidade Federal de Minas Gerais/Brazil - [4]Department of Biochemistry,- University of Melbourne/Australia
Email: chmrodrigues@ufmg.br, dascher@svi.edu.au, douglas.pires@cpqrr.fiocruz.br

Introduction

- ❖ **Kinases** phosphorylate more than 30% of all cellular proteins, modulating their activities and interactions (Cohen, 2001).
- ❖ **Disregulation** of catalytic activity of kinases, through the introduction of **dominant activating mutations**:
  - ➢ Metastasis of many cancers, which has driven the widespread design and use of **kinase inhibitors;**
- ❖ We present **kamp**, a novel machine learning method for **predicting missense activating mutations** in kinases from a **structural perspective**;

Goals

- ❖ Computational analysis of structural data on mutations in kinases:
  - ➢ Understanding the **role of these mutations in diseases** and guide the development of improved and more personalized treatment strategies;
- ❖ **New *in silico* method to predict activating mutations in kinases:**
  - ➢ No robust computational methods for identifying activating mutations have been proposed yet.

Methodology

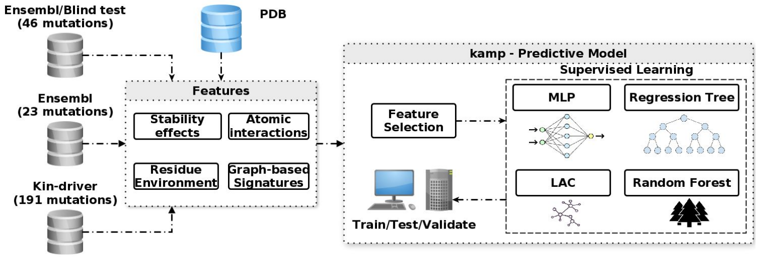


Figure 1. kamp workflow. Experimental data on kinase missense mutations from databases available in the literature were collected, and those were mapped to structures in the PDB. These were used to calculate a set of structural features, including changes in stability, interactions established by wildtype residue and residue environment attributes. To choose the most discriminative attributes a filtering step is performed via feature selection using Principal Component Analysis and Information Gain. The final set of attributes were used for supervised learning to select the best performing model, capable of identifying activating mutations in kinases..

Results

Table 1. Prediction performance for different classification algorithms per mutation class.

Classifier	Precision	Recall	AUC	F-measure	Class
LAC	0,760	0,987	0,904	0,859	Activating
LAC	0,857	0,200	0,904	0,324	Nonactivating
MLP	0,867	0,929	0,839	0,897	Activating
MLP	0,776	0,633	0,839	0,697	Nonactivating
Regression (M5P)	**0,910**	**0,981**	**0,922**	**0,944**	**Activating**
Regression (M5P)	**0,938**	**0,750**	**0,922**	**0,833**	**Nonactivating**
Random Forest	0,893	0,981	0,942	0,935	Activating
Random Forest	0,933	0,700	0,942	0,8	Nonactivating

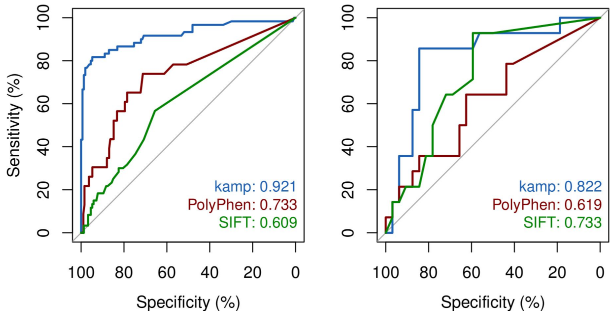
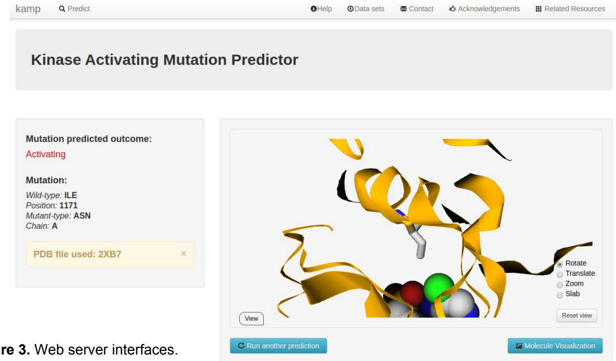


Figure 2. ROC curves showing the performance of kamp compared against the well established methods SIFT (Ng, P.C. and Henikoff,S. 2003) and PolyPhen (Adzhubei, I.A. et al., 2010), for the complete training set (on the left hand side) and the blind test (on the right hand side).



Kinase Activating Mutation Predictor

Mutation predicted outcome:
Activating

Mutation:
Wild-type: **ILE**
Position: **1171**
Mutant-type: **ASN**
Chain: **A**

PDB file used: 2XB7 ×

Figure 3. Web server interfaces.

Conclusions

- ❖ **kamp**, for the first successfully predicts activating mutations on kinases.
- ❖ Very effective in identifying activating mutations (91% precision) as well as non-activating mutations (93%).
- ❖ Outperforms well established sequence-based predictors.
- ❖ It is implemented as a user-friendly web server availble online at http://biosign.cpqrr.fiocruz.br/kamp.

References

1.Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods, 7(4), 248–249.
2.Cohen, P. (1982). The role of protein phosphorylation in neural and hormonal control of cellular activity. Nature, 296, 613–620.
3.Cohen, P. (2001). The role of protein phosphorylation in human health and disease. Euro. J. Biochem., 268(19), 5001–5010.
4.Ng, P. C. and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res., 31(13), 3812–3814.

Newton Fund
FAPEMIG
Fiocruz Minas
Centro de Pesquisas René Rachou
CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

Computational Study and Inference of Mutations Affecting Viral Fitness and Escape from Immune System in the HIV-1 Envelope Glycoprotein

CPqRR / FIOCRUZ

Amanda T. S. Albanaz[1,2], Carlos H. M. Rodrigues[1,2], David B. Ascher[1,4], Douglas E. V. Pires[1]

[1]Biosystems Informatics Group - CPqRR, Fiocruz - Minas Gerais/Brazil; [2]Institute of Biological; Sciences - Centro Universitário Una/Brazil; [3]Department of Biochemistry and Immunology - Universidade Federal de Minas Gerais/Brazil; [4]Department of Biochemistry - University of Cambridge/United Kingdom

Email: amanda.albanaz@cpqrr.fiocruz.br, chmrodrigues@ufmg.br, dascher@svi.edu.au, douglas.pires@cpqrr.fiocruz.br

Introduction

➢Over 70 million people have been infected with HIV-1 in the last 50 years.

➢The virus can infect the human immune system by binding of glycoprotein120 (gp120) with their natural ligand CD4 on surface of host cells.

➢The human immune response is based on the recognition of gp120 by broadly neutralizing antibodies (NAbs).

➢Treatments have been developed based on antibody therapy against gp 120.

➢Rapid virus evolution has limited these treatments:

❖ evasion mechanisms include mutations that decrease dramatically the binding affinity and neutralization sensitivity of NAbs;

❖ these mutations also effect the viral fitness.

Challenge → Predict and anticipate escape mutations → Identify hot spots for antibody design

➢ Over the years *in silico* methods have been helping decipher the effects of mutations in diseases, and elucidating their molecular mechanism.

➢ These can be used to assemble a platform for studying and predicting escape mutations.

Goals

➢**Elucidate the molecular mechanism of escape mutations and their effects on antigen/antibody complex:**

❖ Computationally analyzing mutations in the gp120 structure in complex with VRC01/CD4/b12[3] antibodies.

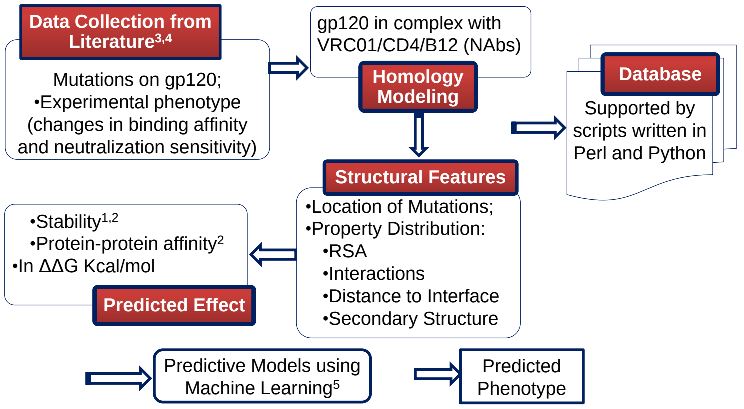
➢**New *in silico* methods to explain:**

❖ How mutations lead to evasion of the immune system?

➢ **And predict:**

❖ Binding affinity, Neutralization sensitivity.

Methodology



Data Collection from Literature[3,4]

Mutations on gp120;
•Experimental phenotype
(changes in binding affinity and neutralization sensitivity)

gp120 in complex with VRC01/CD4/B12 (NAbs)

Homology Modeling

Database

Supported by scripts written in Perl and Python

Structural Features

•Location of Mutations;
•Property Distribution:
•RSA
•Interactions
•Distance to Interface
•Secondary Structure

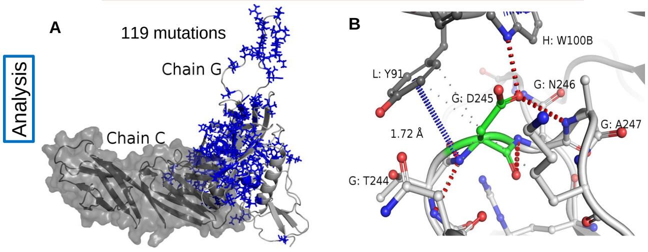
•Stability[1,2]
•Protein-protein affinity[2]
•In ΔΔG Kcal/mol

Predicted Effect

Predictive Models using Machine Learning[5]

Predicted Phenotype

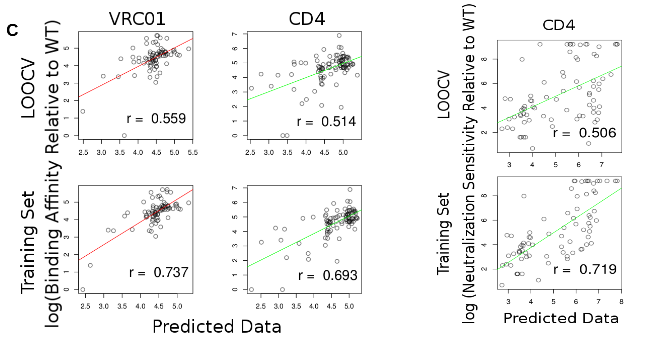
Results



Predicting CD4 escape mutations in gp120 via binary classification

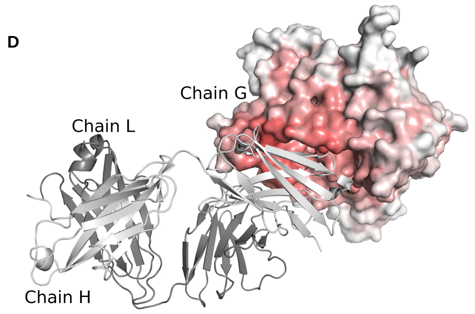
Accuracy: 0.805 AUC: 0.811 Precision: 0.895 Recall: 0.819

Predicting effects of mutations in gp120 via regression tree



A - gp120 (chain G) in complex with CD4 (chain G), and residues shown in blue. **B** – Contact interface presenting interactions between Asp 245 (chain G – gp120), Tyr 91(light chain (L)) and Trp 100B (heavy chain (H)) of VRC01. **C** – Plots presenting the best predictive models found for complexes gp120-VRC01/CD4 and experimental data. Data in logarithmic scale.

Hot spots of escape mutations



D – Heatmap of escape mutations calculated with trained model for gp120 (chain G) in complex with VRC01 (chain L and H),. Hot spots are shown in red.

Conclusions and Future Directions

➢Integration of the effects of mutations on antigen stability, antigen-antibody affinity and mutations structural features, were used to develop a predictive model for escape mutations.

❖ These models achieved correlations of up to r = 0.74 and accuracy of 80.5%

➢ Extrapolation of the trained predictive model allowed identification of hot spots for escape mutations which can be used in further analysis

➢Future efforts will involve considering the effect of these mutations on viral fitness, as well as taking into account other structural effects for model refinement.

➢These can then be used in the design of novel antibodies therapies

References

1.PIRES, D. E. V . ; ASCHER, D. B. ; BLUNDELL, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. **Nucleic Acids Research,** p. W314-W319, 2014.

2.PIRES, D. E. V. ; ASCHER, D. B. ; BLUNDELL, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. **Bioinformatics**, v. 30, p. 335-342, 2013.

3.LI, Yuxing et al. Mechanism of Neutralizing by the Broadly Neutralizing HIV-1 Monoclonal Antibody VRC01. **Journal of Virology**, v.85, n.17, p.8954-8967, 2011.

4.LYNCH, Rebecca M. HIV-1 Fitness Cost Associated with Escape from the VRC01 Class of CD4 Binding Site Neutralizing Antibodies. **Journal of Virology,** v.89, n.8, p.4201-4213, 2015.

5.Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. The WEKA Data Mining Software: An Update; **SIGKDD Explorations**, v. 11, n.1, 2009.

Newton Fund   FAPEMIG   CPqRR - Fiocruz Minas   UNIVERSITY OF CAMBRIDGE

APPENDIX B – Web Components of Kinact

Table 10 – Components used for building Kinact web server.

Component	Version	External link
Bootstrap	3.3.7	<http://getbootstrap.com>
Font Awesome	4.6.3	<http://fontawesome.io/>
Datatables	1.9.4	<http://www.datatables.net>
Selectize	0.12.4	<https://github.com/selectize/selectize.js>
Wallop	2.4.1	<http://pedroduarte.me/wallop>
SweetAlert	4.1.9	<https://limonte.github.io/sweetalert2/>
Perfect Scrollbar	0.6.16	<https://github.com/noraesae/perfect-scrollbar>
3Dmol.js	1.0.1	<http://3dmol.csb.pitt.edu>
MSAViewer	1.0	<http://msa.biojs.net/>
BaguetteBox	1.8.0	<https://github.com/feimosi/baguetteBox.js>
Flask	0.11.1	<http://flask.pocoo.org/>
Redis	2.10.5	<https://redis.io/>
PostgreSQL	9.3	<https://www.postgresql.org/>

Bibliography

ABDI, H.; WILLIAMS, L. J. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, v. 2, n. 4, p. 433–459, jul 2010. ISSN 19395108. Disponível em: <http://doi.wiley.com/10.1002/wics.101>. Cited on page 45.

ADZHUBEI, I. A. et al. A method and server for predicting damaging missense mutations. *Nature methods*, Nature Publishing Group, v. 7, n. 4, p. 248–249, 2010. Cited 4 times on pages 24, 25, 40, and 42.

ALBANAZ, A. T. et al. Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. *Expert Opinion on Drug Discovery*, v. 12, n. 6, p. 553–563, jun 2017. ISSN 1746-0441. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/17460441.2017.1322579>. Cited on page 24.

ALBERTS, B. et al. *Molecular Biology of the Cell*. 6. ed. New York: Garland Science, 2014. ISBN 9780815344322. Cited on page 17.

BAIROCH, A. et al. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, v. 33, n. DATABASE ISS., 2005. ISSN 03051048. Cited on page 35.

BARTLETT, G. J. et al. Analysis of Catalytic Residues in Enzyme Active Sites. *Journal of Molecular Biology*, v. 324, n. 1, p. 105–121, nov 2002. ISSN 00222836. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0022283602010367>. Cited on page 43.

BAXEVANIS, A. D.; BATEMAN, A. The Importance of Biological Databases in Biological Discovery. In: *Current Protocols in Bioinformatics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2015. p. 1.1.1–1.1.8. Disponível em: <http://doi.wiley.com/10.1002/0471250953.bi0101s50>. Cited on page 24.

BERMAN, H. M. et al. The protein data bank. *Nucleic acids research*, v. 28, n. 1, p. 235–242, 2000. ISSN 0305-1048. Cited on page 25.

BOSE, R. et al. Activating HER2 Mutations in HER2 Gene Amplification Negative Breast Cancer. *Cancer Discovery*, v. 3, n. 2, p. 224–237, feb 2013. ISSN 2159-8274. Disponível em: <http://cancerdiscovery.aacrjournals.org/lookup/doi/10.1158/2159-8290.CD-12-0349>. Cited 2 times on pages 17 and 53.

BOUDEAU, J. et al. Emerging roles of pseudokinases. *Trends in Cell Biology*, v. 16, n. 9, p. 443–452, 2006. ISSN 09628924. Cited on page 18.

BOYER, P. D. Energy, Life, and ATP. *Bioscience Reports*, v. 18, n. 3, p. 97–117, 1998. ISSN 01448463. Cited on page 18.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. ISSN 08856125. Cited on page 47.

CHAMARY, J. V.; HURST, L. D. The Price of Silent Mutations. *Scientific American*, v. 300, n. 6, p. 46–53, jun 2009. ISSN 0036-8733. Disponível em:

<http://www.nature.com/doifinder/10.1038/scientificamerican0609-46>. Cited on page 21.

CHAPMAN, B.; CHANG, J. Biopython. *ACM SIGBIO Newsletter*, v. 20, n. 2, p. 15–19, aug 2000. ISSN 01635697. Cited 2 times on pages 42 and 43.

CHASMAN, D.; ADAMS, R. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation11Edited by F. Cohen. *Journal of Molecular Biology*, v. 307, n. 2, p. 683–706, mar 2001. ISSN 00222836. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0022283601945103>. Cited on page 40.

COHEN, P. The origins of protein phosphorylation. *Nature cell biology*, v. 4, n. 5, p. E127–E130, 2002. ISSN 14657392. Cited on page 17.

COHEN, P. T. W. The role of protein phosphorylation in neural and hormonal control of cellular activity. *Nature*, v. 296, n. 5858, p. 613–620, 1982. ISSN 0028-0836. Cited on page 17.

CORPAS, M. et al. BioJS: an open source standard for biological visualisation – its status in 2014. *F1000Research*, feb 2014. ISSN 2046-1402. Disponível em: <http://f1000research.com/articles/3-55/v1>. Cited on page 73.

DHOMEN, N.; MARAIS, R. New insight into BRAF mutations in cancer. *Current Opinion in Genetics & Development*, v. 17, n. 1, p. 31–39, feb 2007. ISSN 0959437X. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0959437X06002449>. Cited on page 39.

DOSS, C. G. P.; SETHUMADHAVAN, R. Investigation on the role of nsSNPs in HNPCC genes–a bioinformatics approach. *Journal of biomedical science*, v. 16, p. 42, 2009. ISSN 1423-0127. Cited on page 25.

EFRON, B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, v. 78, n. 382, p. 316–331, jun 1983. ISSN 0162-1459. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1983.10477973>. Cited on page 48.

FISER, A. Protein structure modeling in the proteomics era. *Expert Review of Proteomics*, v. 1, n. 1, p. 97–110, jun 2004. ISSN 1478-9450. Disponível em: <http://www.tandfonline.com/doi/full/10.1586/14789450.1.1.97>. Cited on page 38.

FLANAGAN, S. E.; PATCH, A.-M.; ELLARD, S. Using SIFT and PolyPhen to Predict Loss-of-Function and Gain-of-Function Mutations. *Genetic Testing and Molecular Biomarkers*, v. 14, n. 4, p. 533–537, aug 2010. ISSN 1945-0265. Disponível em: <http://www.liebertonline.com/doi/abs/10.1089/gtmb.2010.0036>. Cited on page 25.

FORBES, S. A. et al. COSMIC (the Catalogue of Somatic Mutations In Cancer): A resource to investigate acquired mutations in human cancer. *Nucleic Acids Research*, v. 38, n. SUPPL.1, p. 652–657, 2009. ISSN 03051048. Cited 3 times on pages 33, 35, and 75.

FRANK, E. et al. Using model trees for classification. *Machine Learning*, v. 32, n. 1, p. 63–76, 1998. ISSN 08856125. Cited on page 46.

FRAPPIER, V.; CHARTIER, M.; NAJMANOVICH, R. J. ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. *Nucleic Acids Research*, Oxford Univ Press, v. 43, n. W1, p. W395–W400, jul 2015. ISSN 0305-1048. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv343>. Cited 2 times on pages 42 and 43.

GHARWAN, H.; GRONINGER, H. Kinase inhibitors and monoclonal antibodies in oncology: clinical implications. *Nature Reviews Clinical Oncology*, v. 13, n. 4, p. 209–227, dec 2015. ISSN 1759-4774. Disponível em: <http://www.nature.com/doifinder/10.1038/nrclinonc.2015.213>. Cited on page 23.

GIBBS, C. S.; ZOLLER, M. J. Rational scanning mutagenesis of a protein kinase identifies functional regions involved in catalysis and substrate interactions. *Journal of Biological Chemistry*, v. 266, n. 14, p. 8923–8931, 1991. ISSN 00219258. Cited on page 19.

GNAD, F. et al. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC genomics*, v. 14 Suppl 3, p. S7, 2013. ISSN 1471-2164. Cited on page 25.

GOLDBERG, J. M. et al. Kinannote, a computer program to identify and classify members of the eukaryotic protein kinase superfamily. *Bioinformatics*, v. 29, n. 19, p. 2387–2394, 2013. ISSN 14602059. Cited on page 34.

GRABINER, B. C. et al. A Diverse Array of Cancer-Associated MTOR Mutations Are Hyperactivating and Can Predict Rapamycin Sensitivity. *Cancer Discovery*, v. 4, n. 5, p. 554–563, may 2014. ISSN 2159-8274. Disponível em: <http://cancerdiscovery.aacrjournals.org/cgi/doi/10.1158/2159-8290.CD-13-0929>. Cited on page 17.

GRINBERG, M. *Flask web development: developing web applications with python.* [S.l.]: " O'Reilly Media, Inc.", 2014. Cited on page 65.

GUEROIS, R.; NIELSEN, J. E.; SERRANO, L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *Journal of Molecular Biology*, v. 320, n. 2, p. 369–387, jul 2002. ISSN 00222836. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0022283602004424>. Cited on page 40.

HALL, M. et al. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, v. 11, n. 1, p. 10, nov 2009. ISSN 19310145. Cited 2 times on pages 45 and 47.

HANKS, S. K.; HUNTER, T. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, v. 9, n. 8, p. 576–96, may 1995. ISSN 0892-6638. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/7768349>. Cited on page 19.

HANKS, S. K.; QUINN, A. M.; HUNTER, T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science (New York, N.Y.)*, v. 241, n. 4861, p. 42–52, jul 1988. ISSN 0036-8075. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/3291115>. Cited on page 19.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, v. 143, n. 1, p. 29–36, apr 1982. ISSN 0033-8419. Disponível em: <http://pubs.rsna.org/doi/10.1148/radiology.143.1.7063747>. Cited on page 63.

HUBBARD, T. The Ensembl genome database project. *Nucleic Acids Research*, v. 30, n. 1, p. 38–41, jan 2002. ISSN 13624962. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/30.1.38>. Cited on page 35.

HUMPHREY, S. J.; JAMES, D. E.; MANN, M. Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends in Endocrinology and Metabolism*, Elsevier Ltd, v. 26, n. 12, p. 676–687, 2015. ISSN 18793061. Disponível em: <http://dx.doi.org/10.1016/j.tem.2015.09.013>. Cited on page 17.

HUSE, M.; KURIYAN, J. The Conformational Plasticity of Protein Kinases. *Cell*, v. 109, n. 3, p. 275–282, may 2002. ISSN 00928674. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0092867402007419>. Cited on page 19.

JAIN, R. *Art of Computer Systems Performance Analysis Techniques For Experimental Design Measurements Simulation And Modeling*. [S.l.]: Wiley Computer Publishing, 2015. 685 p. ISBN 0471503363. Cited on page 51.

JOHNSON, L. N. et al. The structural basis for substrate recognition and control by protein kinases 1. *FEBS Letters*, v. 430, n. 1-2, p. 1–11, jun 1998. ISSN 00145793. Cited on page 21.

JOHNSON, L. N.; NOBLE, M. E.; OWEN, D. J. Active and Inactive Protein Kinases: Structural Basis for Regulation. *Cell*, v. 85, n. 2, p. 149–158, apr 1996. ISSN 00928674. Cited on page 19.

JOLLIFFE, I. *Principal Component Analysis*. [S.l.]: Springer, 2002. (Springer Series in Statistics). ISBN 9780387954424. Cited on page 45.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, v. 349, n. 6245, p. 255–260, jul 2015. ISSN 0036-8075. Cited on page 46.

JUBB, H. C. et al. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of Molecular Biology*, Elsevier Ltd, v. 429, n. 3, p. 365–371, feb 2017. ISSN 00222836. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0022283616305332>. Cited 2 times on pages 42 and 44.

KARCZEWSKI, K. J. et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Research*, v. 45, n. D1, p. D840–D845, jan 2017. ISSN 0305-1048. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw971>. Cited on page 75.

KORNEV, A. P. et al. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proceedings of the National Academy of Sciences*, v. 103, n. 47, p. 17783–17788, nov 2006. ISSN 0027-8424. Cited on page 19.

KOTSIANTIS, S. et al. Using Data Mining Techniques for Estimating Minimum , Maximum and Average Daily Temperature Values. *Maximum and Average Daily Temperature Values", World Academy of Science, Engineering and Technology*, v. 1, n. 1, p. 16–20, 2007. Cited 2 times on pages 46 and 47.

KOTSIANTIS, S. B.; ZAHARAKIS, I. D.; PINTELAS, P. E. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, v. 26, n. 3, p. 159–190, nov 2006. ISSN 0269-2821. Disponível em: <http://link.springer.com/10.1007/s10462-007-9052-3>. Cited on page 46.

LAHIRY, P. et al. Kinase mutations in human disease: interpreting genotype–phenotype relationships. *Nature Reviews Genetics*, Nature Publishing Group, v. 11, n. 1, p. 60–74, jan 2010. ISSN 1471-0056. Disponível em: <http://www.nature.com/doifinder/10.1038/nrg2707>. Cited on page 17.

LANDRUM, M. J. et al. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, v. 42, n. D1, p. 980–985, 2014. ISSN 03051048. Cited on page 35.

MACKAY, D. J. C. Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, Springer Verlag, v. 168, p. 133–166, 1998. Cited 2 times on pages 46 and 47.

MANNING, G. et al. The Protein Kinase Complement of the Human Genome. *Science*, v. 298, n. 5600, p. 1912–1934, 2002. ISSN 00368075. Cited 3 times on pages 18, 29, and 39.

MARSHALL, C. J. MAP kinase kinase kinase, MAP kinase kinase and MAP kinase. *Current Opinion in Genetics & Development*, v. 4, n. 1, p. 82–89, feb 1994. ISSN 0959437X. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/0959437X94900957>. Cited on page 18.

MARTELLI, A.; RAVENSCROFT, A.; ASCHER, D. *Python cookbook*. [S.l.]: " O'Reilly Media, Inc.", 2005. Cited on page 65.

MARTIN, A. C. R. Mapping PDB chains to UniProtKB entries. *Bioinformatics*, v. 21, n. 23, p. 4297–4301, dec 2005. ISSN 1367-4803. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti694>. Cited on page 35.

MISURA, K. M.; BAKER, D. Progress and challenges in high-resolution refinement of protein structure models. *Proteins: Structure, Function, and Bioinformatics*, v. 59, n. 1, p. 15–29, feb 2005. ISSN 08873585. Disponível em: <http://doi.wiley.com/10.1002/prot.20376>. Cited on page 38.

NAKANISHI, Y. et al. Activating Mutations in PIK3CB Confer Resistance to PI3K Inhibition and Define a Novel Oncogenic Role for p110. *Cancer Research*, v. 76, n. 5, p. 1193–1203, mar 2016. ISSN 0008-5472. Disponível em: <http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-15-2201>. Cited on page 24.

NG, P. C.; HENIKOFF, S. Predicting Deleterious Amino Acid Substitutions. *Genome Research*, v. 11, n. 5, p. 863–874, may 2001. ISSN 1088-9051. Disponível em: <http://genome.cshlp.org/cgi/doi/10.1101/gr.176601>. Cited 2 times on pages 24 and 42.

NG, P. C.; HENIKOFF, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, v. 31, n. 13, p. 3812–3814, jul 2003. ISSN 1362-4962. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg509>. Cited 2 times on pages 25 and 40.

OPENSTAX, B. *OpenStax CNX*. 2015. Disponível em: <http://cnx.org/contents/185cbf87-c72e-48f5-b51e-f14f21b5eabd@9.87>. Cited on page 23.

PANDURANGAN, A. P. et al. SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Research*, v. 45, n. W1, p. W229–W235, jul 2017. ISSN 0305-1048. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx439>. Cited 2 times on pages 26 and 42.

PIRES, D. E.; ASCHER, D. B. mCSM-AB: a web server for predicting antibody–antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Research*, v. 44, n. W1, p. W469–W473, jul 2016. ISSN 0305-1048. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw458>. Cited on page 26.

PIRES, D. E.; ASCHER, D. B. mCSM–NA: predicting the effects of mutations on protein–nucleic acids interactions. *Nucleic Acids Research*, apr 2017. ISSN 0305-1048. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx236>. Cited on page 26.

PIRES, D. E. V.; ASCHER, D. B.; BLUNDELL, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Research*, v. 42, n. W1, p. W314–W319, jul 2014. ISSN 0305-1048. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku411>. Cited 2 times on pages 26 and 42.

PIRES, D. E. V.; ASCHER, D. B.; BLUNDELL, T. L. MCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, v. 30, n. 3, p. 335–342, 2014. ISSN 13674803. Cited 3 times on pages 26, 42, and 44.

PIRES, D. E. V.; BLUNDELL, T. L.; ASCHER, D. B. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Scientific Reports*, v. 6, n. 1, p. 29575, sep 2016. ISSN 2045-2322. Disponível em: <http://www.nature.com/articles/srep29575>. Cited on page 26.

PIRES, D. E. V. et al. ACSM: Noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, v. 29, n. 7, p. 855–861, 2013. ISSN 13674803. Cited on page 26.

PIRES, D. E. V. et al. Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, v. 12, n. Suppl 4, p. S12, 2011. ISSN 1471-2164. Disponível em: <http://www.biomedcentral.com/1471-2164/12/S4/S12>. Cited on page 26.

PORTER, C. T.; BARTLETT, G. J.; THORNTON, J. M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic acids research*, v. 32, n. Database issue, p. D129–33, 2004. ISSN 1362-4962. Cited 2 times on pages 42 and 43.

POWERS, D. M. W. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, v. 2, n. 1, p. 37–63, 2011. Cited on page 49.

PRAVDA, L. et al. Anatomy of enzyme channels. *BMC Bioinformatics*, v. 15, n. 1, p. 379, dec 2014. ISSN 1471-2105. Cited on page 43.

RASMUSSEN, C. E.; WILLIAMS, C. K. *Gaussian processes for machine learning*. [S.l.]: MIT press Cambridge, 2006. Cited on page 47.

REGO, N.; KOES, D. 3Dmol.js: molecular visualization with WebGL. *Bioinformatics*, v. 31, n. 8, p. 1322–1324, apr 2015. ISSN 1367-4803. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu829>. Cited on page 69.

RIEDMILLER, M. Advanced supervised learning in multi-layer perceptrons - from backpropagation to adaptive learning algorithms. *Computer Standards and Interfaces*, v. 16, n. 3, p. 265–278, 1994. ISSN 09205489. Cited on page 46.

SHERRY, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, v. 29, n. 1, p. 308–311, jan 2001. ISSN 13624962. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/29.1.308>. Cited on page 35.

SHEVADE, S. et al. Improvements to the SMO algorithm for SVM regression. *IEEE Transactions on Neural Networks*, v. 11, n. 5, p. 1188–1193, 2000. ISSN 10459227. Disponível em: <http://ieeexplore.ieee.org/document/870050/>. Cited on page 26.

SIMONETTI, F. L. et al. Kin-Driver: a database of driver mutations in protein kinases. *Database*, v. 2014, p. bau104–bau104, nov 2014. ISSN 1758-0463. Disponível em: <https://academic.oup.com/database/article-lookup/doi/10.1093/database/bau104>. Cited on page 33.

STAJICH, J. E. The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, v. 12, n. 10, p. 1611–1618, oct 2002. ISSN 10889051. Disponível em: <http://www.genome.org/cgi/doi/10.1101/gr.361602>. Cited on page 43.

SUN, Y.; WONG, A. K. C.; KAMEL, M. S. CLASSIFICATION OF IMBALANCED DATA: A REVIEW. *International Journal of Pattern Recognition and Artificial Intelligence*, v. 23, n. 04, p. 687–719, jun 2009. ISSN 0218-0014. Disponível em: <http://www.worldscientific.com/doi/abs/10.1142/S0218001409007326>. Cited on page 38.

SUTTO, L.; GERVASIO, F. L. Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase. *Proceedings of the National Academy of Sciences*, v. 110, n. 26, p. 10616–10621, jun 2013. ISSN 0027-8424. Disponível em: <http://www.pnas.org/cgi/doi/10.1073/pnas.1221953110>. Cited 2 times on pages 51 and 53.

TIACCI, E. et al. Genomics of Hairy Cell Leukemia. *Journal of Clinical Oncology*, v. 35, n. 9, p. 1002–1010, mar 2017. ISSN 0732-183X. Disponível em: <http://ascopubs.org/doi/10.1200/JCO.2016.71.1556>. Cited 2 times on pages 17 and 23.

TOPHAM, C. M.; SRINIVASAN, N.; BLUNDELL, T. L. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein engineering*, v. 10, n. 1, p. 7–21, jan 1997. ISSN 0269-2139. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/9051729>. Cited 2 times on pages 26 and 42.

VALDMANIS, P. N.; VERLAAN, D. J.; ROULEAU, G. A. The proportion of mutations predicted to have a deleterious effect differs between gain and loss of function genes in neurodegenerative disease. *Human Mutation*, v. 30, n. 3, p. 481–489, 2009. ISSN 10597794. Cited on page 25.

WAN, P. T. et al. Mechanism of Activation of the RAF-ERK Signaling Pathway by Oncogenic Mutations of B-RAF. *Cell*, v. 116, n. 6, p. 855–867, mar 2004. ISSN 00928674. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0092867404002156>. Cited 2 times on pages 51 and 53.

WEBB, B.; SALI, A. Comparative Protein Structure Modeling Using MODELLER. In: *Current Protocols in Protein Science*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2016. p. 2.9.1–2.9.37. Disponível em: <http://doi.wiley.com/10.1002/cpps.20>. Cited on page 38.

WLODAWER, A. et al. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS Journal*, Blackwell Publishing Ltd, v. 275, n. 1, p. 1–21, 2008. ISSN 1742-4658. Disponível em: <http://dx.doi.org/10.1111/j.1742-4658.2007.06178.x>. Cited on page 51.

WORTH, C. L.; PREISSNER, R.; BLUNDELL, T. L. SDM–a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Research*, v. 39, n. suppl, p. W215–W222, jul 2011. ISSN 0305-1048. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr363>. Cited 2 times on pages 26 and 42.

YACHDAV, G. et al. MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, p. btw474, jul 2016. ISSN 1367-4803. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw474>. Cited on page 73.

YU, L. et al. Role of Mg 2+ ions in protein kinase phosphorylation: insights from molecular dynamics simulations of ATP-kinase complexes. *Molecular Simulation*, v. 37, n. 14, p. 1143–1150, dec 2011. ISSN 0892-7022. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/08927022.2011.561430>. Cited on page 18.

ZAKI, M. J.; MEIRA, M. J. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. New York: Cambridge University Press, 2014. ISBN 9780521766333. Disponível em: <http://www.dataminingbook.info/pmwiki.php/Main/BookDownload>. Cited 3 times on pages 45, 48, and 49.