PLOS ONE

RESEARCH ARTICLE

Understanding the impacts of missense mutations on structures and functions of human cancer-related genes: A preliminary computational analysis of the COSMIC Cancer Gene Census

Sony Malhotra[1☯¤*], **Ali F. Alsulami**[1☯], **Yang Heiyun**[1], **Bernardo Montano Ochoa**[1], **Harry Jubb**[2], **Simon Forbes**[2], **Tom L. Blundell**[1*]

1 Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom, **2** Wellcome Genome Campus, Hinxton, Cambridgeshire, United Kingdom

☯ These authors contributed equally to this work.
¤ Current address: Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck College, University of London, London, United Kingdom.
* s.malhotra@mail.cryst.bbk.ac.uk (SM); tlb20@cam.ac.uk (TLB)

OPEN ACCESS

Citation: Malhotra S, Alsulami AF, Heiyun Y, Ochoa BM, Jubb H, Forbes S, et al. (2019) Understanding the impacts of missense mutations on structures and functions of human cancer-related genes: A preliminary computational analysis of the COSMIC Cancer Gene Census. PLoS ONE 14(7): e0219935. https://doi.org/10.1371/journal.pone.0219935

Editor: Yang Zhang, University of Michigan, UNITED STATES

Received: March 6, 2019

Accepted: July 3, 2019

Published: July 19, 2019

Data Availability Statement: All relevant gene IDs that were used are within the paper (Table 2) and the mutations used were retrieved from the COSMIC database (https://cancer.sanger.ac.uk/census) using gene names listed in Table 2.

Funding: SM was funded by UK Medical Research Council (X5 06489 DBTMRC Joint Centre Partnership). AFA is funded on a PhD Scholarship by the Kingdom of Saudi Arabia. HJ was supported by a collaboration between Astex Pharmaceuticals

Abstract

Genomics and genome screening are proving central to the study of cancer. However, a good appreciation of the protein structures coded by cancer genes is also invaluable, especially for the understanding of functions, for assessing ligandability of potential targets, and for designing new drugs. To complement the wealth of information on the genetics of cancer in COSMIC, the most comprehensive database for cancer somatic mutations available, structural information obtained experimentally has been brought together recently in COSMIC-3D. Even where structural information is available for a gene in the Cancer Gene Census, a list of genes in COSMIC with substantial evidence supporting their impacts in cancer, this information is quite often for a single domain in a larger protein or for a single protomer in a multiprotein assembly. Here, we show that over 60% of the genes included in the Cancer Gene Census are predicted to possess multiple domains. Many are also multicomponent and membrane-associated molecular assemblies, with mutations recorded in COSMIC affecting such assemblies. However, only 469 of the gene products have a structure represented in the PDB, and of these only 87 structures have 90–100% coverage over the sequence and 69 have less than 10% coverage. As a first step to bridging gaps in our knowledge in the many cases where individual protein structures and domains are lacking, we discuss our attempts of protein structure modelling using our pipeline and investigating the effects of mutations using two of our in-house methods (SDM2 and mCSM) and identifying potential driver mutations. This allows us to begin to understand the effects of mutations not only on protein stability but also on protein-protein, protein-ligand and protein-nucleic acid interactions. In addition, we consider ways to combine the structural information with the wealth of mutation data available in COSMIC. We discuss the impacts of COSMIC

and Wellcome Trust Genome Campus. TLB acknowledges support through his Wellcome Trust Investigator Award 200814/Z/16/Z. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare that they have no competing interests. This does not alter our adherence to PLOS ONE policies on sharing data and materials. Astex Pharmaceuticals has no competing interests. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

missense mutations on protein structure in order to identify and assess the molecular consequences of cancer-driving mutations.

Introduction

Cancer is one of the most common diseases afflicting humanity today and the second leading cause of death globally (WHO Key Facts, Feb 2018). Cancer refers to any genetic disease that leads to an uncontrolled proliferation, causing a tumor. In 2015, there were 90 million cases worldwide and 8.8 million deaths due to cancer[1]. Its toll on the world is expected only to increase in the future.

Drug development is an expensive and time consuming process that can take decades, but the first step for most cancers is to look for a good protein target. Thanks to many breakthroughs in the field of human genome sequencing, we now have a vast amount of information that may improve our understanding of the genetics of cancer. Although we have a good description of mutations that recur in common cancers, defining the structures of the gene products, which is important for predicting the impacts of most mutations, is much more challenging and expensive. This leads to a gap in our understanding of how the sequence data relate to the structure and function of the protein.

In 2003 when the human genome project first sequenced the entire human genome, it cost an estimate of $300 million and a world spanning initiative (https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/). Since then many improvements and breakthroughs have been made in the field of DNA sequencing, drastically decreasing its cost and time consumption. In 2015, the cost of generating a high quality sequence of the whole human genome had fallen to below $1500, and the time required had dropped from 13 years to just 1 or 2 days[2]. These technologies have been collectively termed second generation or next generation sequencing.

The reduced cost in genome sequencing has allowed researchers to find trends in mutations in many tumors taken from patients around the world. Along with the cheaper sequencing technology has come the need for online databases to store sequence data. COSMIC[3], the Catalogue of Somatic Mutations in Cancer, is currently the most comprehensive database of mutations in cancer. Started in 2004, COSMIC provides curated information on somatic mutations. It combines large scale genome screening data from over 32,000 genomes (v86, August 2018), and manual curation of over 25,000 individual publications.

An important focus of the manual curation in COSMIC is the Cancer Gene Census (CGC) [4], a list of genes with substantial literature describing their impacts in cancer development, diseases caused, and indications of the mechanisms involved. There are currently 719 genes in the Cancer Gene Census, which are causally implicated in oncogenesis, and are divided into tier1 (574) and tier2 (145) types. For a gene to be included in the census, there has to be genetic evidence from two or more independent reports showing mutations in the gene in primary patient material, and ideally biological information supporting the oncogenic effects of the mutations. Tier1 genes have a documented activity relevant to cancer, and the mutations in the gene product promote oncogenic transformation and change the activity of the gene product, whereas for the genes in tier2 there is less evidence for their roles in cancer. The census does not include genes that experience only altered levels of expression in cancer cells, or genes that experience epigenetic changes such as methylation of CpG dinucleotides within promoter regions. These are likely the consequences rather than determinants of the oncogenesis.

Protein structures for Cancer Gene Census analysis: COSMIC-3D

To have a better understanding of the structural and functional impacts of the cancer-related mutations, it is important to map these mutations on to the protein structure and analyse their interactions with other cellular macromolecules (such as proteins, nucleic acids, ligands etc.).

COSMIC-3D (http://cancer.sanger.ac.uk/cosmic3d/) provides a new bioinformatics platform for analyzing mutations in some of the 9300 genes in COSMIC including the 390 genes from the Cancer Gene Census onto the experimentally-derived human protein structures[5]. By mapping the mutation data onto the crystal structure of the protein, COSMIC-3D provides a helpful route to understand the structural context of the mutation in terms of its interaction with the other residues in the same protein or with other molecules when the structure of the protein is available in a bound conformation. However, as not all mutations will directly impact on interaction interfaces, further predictive tools are required. The first challenge is to use the experimental data brought together in COSMIC-3D to understand further the impacts of different missense (nonsynonymous) mutations from COSMIC.

Identifying driver mutations

Cancer originates from genetic alteration(s) that affect cellular processes and division. Genes that are highly mutated and lead to cancer progression are known as drivers, which can be characterized as either oncogenes (activating) or tumor suppressors (inactivating)[6]. Candidate driver genes have often been identified based on mutation frequency of that gene compared to the background mutation rate, which is very challenging to estimate due to variability between cancer samples and cancers type[7].

There are three ways that are commonly used to identify background mutations: first frequency-based approaches[8] based on synonymous mutation rates; secondly, feature approaches, such as guanine and cytosine (GC) content, gene density, nucleosome occupancy, distance to telomere and centromere *etc.*[9]; and thirdly function-based methods that consider mutations in the conserved region of the protein that might have functional impacts[10], estimated on the basis of chemical and structure similarity between wildtype and mutant amino acid. The number of samples does not matter in the functional assay unlike frequency estimation methods[8].

Estimation of synonymous mutation rates is problematic where genes have very small numbers of synonymous mutations; here the rates can be estimated by mutations occurring at intron and unrelated regions assuming mutations occur there naturally, which is not always true. Driver genes are difficult to identify either by the background frequency rate or functional-based methods. This is mainly because, when there are several other genes present in the same pathway, mutation of the first gene could give a selective advantage for a tumor to progress, and therefore, other gene mutations will infrequently act as drivers[11]. Although most of the cancer-driver genes are associated with one cancer type, there are a few genes present in more than one cancer type such as TP53[12]. A new PanCancer study has identified a total of 299 driver genes using multiple bioinformatics algorithms[13].

Identification of driver mutations in the patient genomes from a set comprising all occurring mutations is a daunting task and needs functional tests that are usually time consuming and laborious. Hence, the driver mutations are usually identified on the basis of their recurrence at a particular position in all samples. The ones with highest frequency are usually identified as likely driver mutations. Where the mutation frequency is not helpful, possible drivers are often suggested on the basis of 3D proximity to each other or to other frequent mutations in that gene[14]. However, identification of driver mutations from the set of more prevalent

passenger mutations remains an important step in the development of effective and targeted therapies towards different cancer types.

Here, we focus on understanding the effects of mutations in multicomponent molecular assemblies, found in the cytoplasm, nucleus, membranes and vesicles in the cell. This comprises a major challenge as further structural information is required to understand the impacts of mutations. Although ~500 of the 719 proteins in the CGC have an experimental structure in the PDB, less than a fifth of the experimental structures reported have 90–100% coverage of the full sequence. Furthermore, complete structures are available for very few of the multiprotein assemblies that are required for cellular function. We show that mutations in the CGC of COSMIC likely affect protein stability as well as protein-protein, protein-ligand and protein-nucleic acid interactions. The importance of mutations listed in COSMIC that affect such interactions has been emphasised in recent analyses[15–18]. In order to understand the structural and functional impacts of genes from CGC, we have predicted structural information where it is not experimentally available and mapped mutations not only onto the structures of individual domains, but also multidomain and multicomponent systems, using statistical and machine-learning methods to predict their impacts, often through allosteric mechanisms. We illustrate our approach with case studies not only where structures are experimentally defined and therefore provide a reliable basis for the predictive methods, but also where individual domains or subunits are defined, and full-length proteins or multicomponent systems need to be modelled. We have selected examples that include the impacts of mutations on protein-protein (Ras with Son of Sevenless homolog protein, SMAD2 homodimer), protein-ligand (BRAF-inhibitor complex) and protein-nucleic acid (androgen receptor) interactions in important cell regulatory systems. This approach adds to our understanding of cancer target function and helps in distinguishing functionally important mutations.

Results

Mapping sequence domains to the genes in the Cancer Gene Census

In order to gain insights into the functions and impacts of mutations of the proteins with unknown structure, where possible we map sequences of gene products to sequence domains using an HMM[19] search against the PFam[20] database. Of the 719 genes included in the Cancer Gene Census, 205 genes are single domain and 476 genes are predicted to be multidomain, leaving 38 genes with no PFam domain predicted using HMMER3 (Fig 1A). Furthermore, many are either homo-oligomers or contributors to multicomponent assemblies, often varying over space and time. This presents a major challenge to experimental approaches, which many of us are pursuing. However, computational analyses of structures and their interactions will likely be required for many years to come.

Beyond COSMIC-3D

COSMIC-3D is severely limited by the availability of experimentally solved protein structures present in the PDB. Of the 719 genes in the CGC, 469 have a structure representation in the PDB, leaving 250 without PDB structures. This number of available structures is further limited by the fact that most often the structure solved for a gene product does not have a100% coverage of its sequence. In the set of 469 genes with known structures, only 87 structures have 90–100% coverage over the sequence (see Fig 1B) and 69 have less than 10% coverage. Furthermore, many of the 250 genes with no structure representation have a large number of mutations documented in the CGC (Fig 1C). We have also assessed the protein sequences of these 250 genes with no known structure information in terms of their disorder content (Fig 1D) using DISOPRED3[21], and have shown that many have a high percentage of residues in

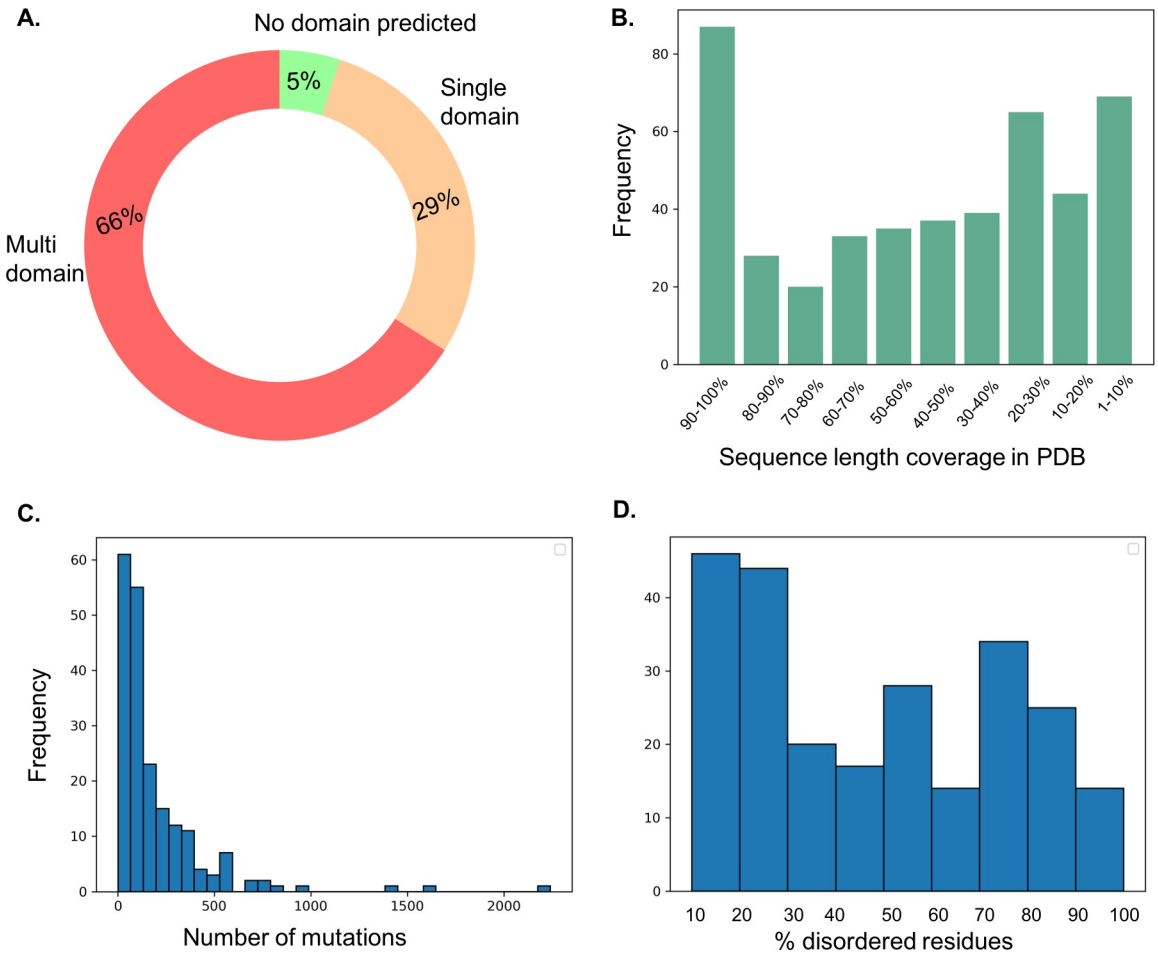


Fig 1. A. Mapping sequence domains to genes in the cancer gene census. 66% of the genes are predicted to have multiple domains and 29% of the genes are single domain. B. The genes in the Cancer Gene Census with their structural coverage in the protein databank. C. The distribution of mutations reported in the genes from Cancer Gene Census that do not have structure representation in the Protein DataBank. D. Histogram showing the distribution of residues in disordered regions for the protein sequences of the genes from Cancer Gene Census that do not have structure representation in the Protein DataBank.

https://doi.org/10.1371/journal.pone.0219935.g001

disordered regions (5% of genes have >90% residues in disordered regions). Clearly, to interpret the effects of these on protein function, one needs to build structural models. We are in the process of organizing these structural models and the predicted effects of mutations in the form of a database (Alsulami AF, P. H. M. Torres and Blundell, TL, under preparation).

Effects of mutations on protein structure and function

Owing to their genetic instability the cancer samples are highly heterogeneous and possess many missense mutations. However, most of these mutations are likely to be neutral or passengers; only a few have deleterious effects and are driver mutations under positive selection pressure. Both oncogenes and tumor suppressor genes are involved in a dense network of interactions with other proteins, nucleic acids and small molecules. Therefore, by combining knowledge of the mutation data with information on molecular interactions, we can identify the molecular mechanisms of carcinogenesis and the likely impacts of mutations in driver genes [22,23].

Recently, there have been attempts to map the mutation data from different cancer-cell types onto protein structures and hence identify clusters of mutations[14,24]. This helps in identifying the new targets, types of interactions disrupted upon mutations and the potential functional effects. However, experiments in structural biology and mutagenesis studies comparing the free energy differences between wildtype and mutant proteins are costly and time consuming. Nevertheless, databases such as ProTherm[25] provide a resource for experimental thermodynamic data on mutant proteins, allowing for larger scale studies of mutation impacts.

This has led to the development of many computational algorithms to study missense mutations and their impacts on protein stability and function. There are several different approaches that have been used to study the impacts of mutations. Most sequence-based approaches consider the local conservation patterns in homologues to predict how damaging mutations at a certain residue would be. SIFT[26] and PolyPhen[27] are very popular sequence-based methods. Structure-based approaches, which make use of the protein 3D structure (either experimentally derived or modelled), typically fall into the categories of potential energy functions or machine-learning methods. The physics-based methods of predicting the effects of mutations rely heavily on the position of side chains in order to define interactions and clashes; therefore, they require accurate positioning of the atoms including bound waters. On the other hand methods that are based on either statistical potentials, such as BeATMuSiC[28], or rely on structural profiles[29] usually require only an approximation of the protein structure. STRUM[30] shows that the predictions of effects of mutations rely more on the accurate prediction of the global fold and have only a marginal dependence on the accuracy of protein structures.

Some early methods such as SDM[31,32] use environment-specific substitution tables, while others such as PoPMuSiC[33,34] use potential energy functions to calculate the change in free energy. More recently, some structure-based approaches, such as mCSM, have used machine-learning methods. These can come in different flavours such as mCSM-PPI[35] (protein-protein interactions), mCSM-lig[36,37] (ligand binding) and mCSM-NA[38] (nucleic acid binding) or neural networks (PoPMuSiC-2) to predict the impact of mutations. There are various ways of feeding structural data into a machine-learning algorithm; one approach has been to turn the structure into a graph-based signature, which is the principle behind mCSM.

There are also molecular dynamics-based methods that can predict the effects of single point mutations at protein-protein interfaces[39]. Kellogg *et al*. have investigated the performance and accuracy of different protocols to predict effects of mutations by extensively searching through alternative conformations[40]. Such analyses of the wealth of mutation data in tumors, using these various mutation-analysis softwares, can be used to evaluate their ability to predict carcinogenic mutations.

Preliminary studies of Cancer Gene Census proteins

Hallmark genes in the CGC are classified following the approach of Hanahan and Weinberg [41], exemplifying the following biological capabilities: proliferative signalling, suppression growth, escaping immune response to cancer, invasion and metastasis, tumor promoter inflammation, and cell replicative immortality. Hallmark genes are marked in the CGC and have manually curated information available on protein function.

In order to illustrate the challenge of understanding the impacts of the mutations we have extended the modeling to examples from the Hallmark dataset in the Cancer Gene census as these have a huge amount of manually curated information on the functional effects of mutations that will help in understanding the structural changes upon mutations. We have chosen

five examples of proteins that have features characteristic of the Cancer Gene Census. These include structures that are experimentally defined and therefore provide a more certain basis for assessing the impacts of the modelling software; they also include modelled structures where predictions of protomer structures are generally reliable, but where multicomponent assemblies are more challenging. They include homo-oligomeric structures, hetero-oligomers/multiprotein assemblies with protein-protein interactions, nucleic acid-binding proteins, protein-ligand interactions and membrane proteins.

Multiprotein assemblies

Hetero-complex of Ras protein with Son of Sevenless protein homolog 1

Ras is a signalling molecule that acts as a switch by shuttling between the active (GTP bound) and inactive (GDP bound) form. Ras forms a complex with Ras-specific nucleotide exchange factor, Son of Sevenless (SOS), which helps in activating receptors that signal through tyrosine kinases. It is known that the unregulated activation of Ras is a hallmark of many cancers[42].

Here, we have mapped the mutations known for HRAS-1 from COSMIC onto the hetero complex of SOS with HRAS-GDP (PDBID: 1XD2[42]). For HRAS-1, 59 residue positions are documented to have mutations in at least one cancer sample (Fig 2A). We have mapped the mutations known for HRAS-1 from COSMIC onto the ternary complex of SOS with HRAS-GDP (PDBID: 1XD2). The most frequently mutated residues are G12, G13 and Q61. Gao *et al.* [24] have recently identified mutational 3D clusters, which assist in identifying the possible driver mutations in cancer targets. Q61 was observed to be part of one such cluster, which has other residues with low mutation frequency (colored in purple) and these spatially

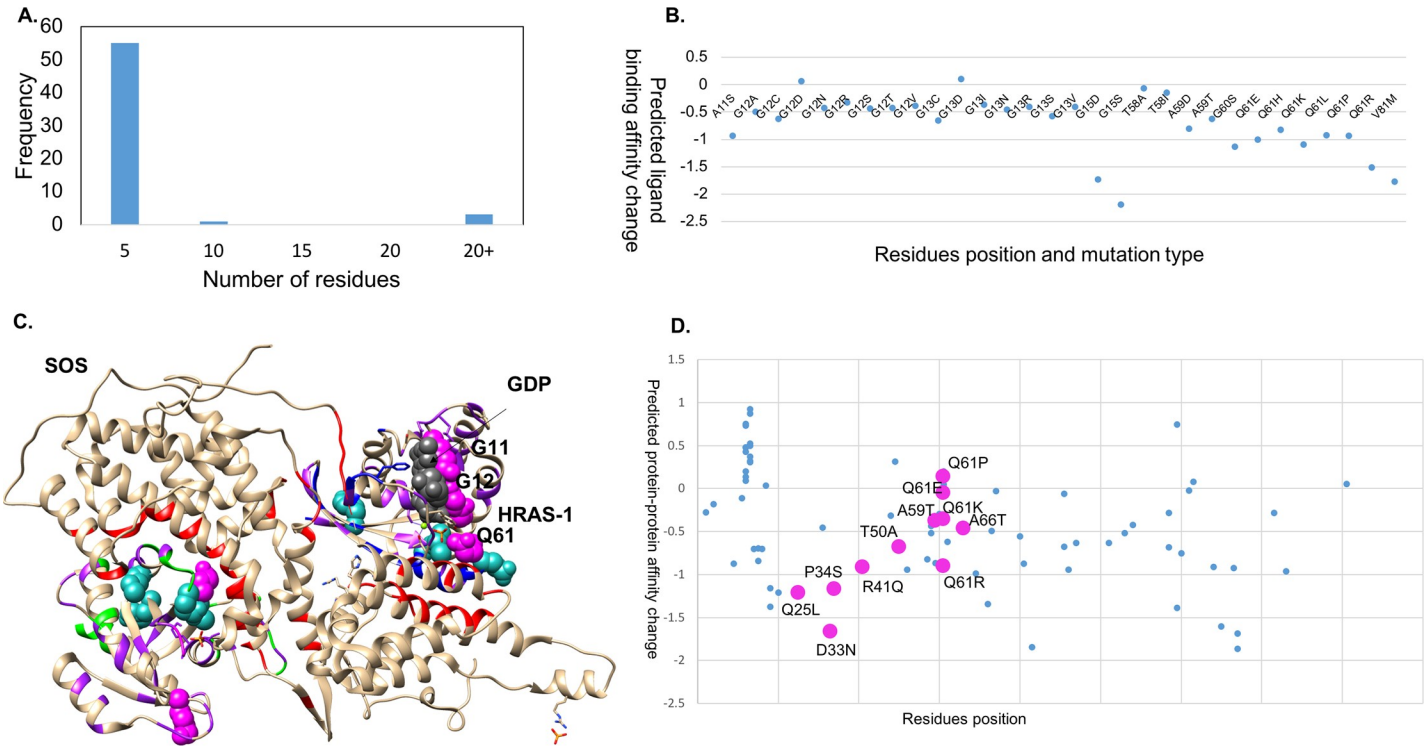


Fig 2. Mutations in HRAS-1. A. Frequency distribution in cancer samples in COSMIC. C. Mutations mapped on to the hetero-complex of HRAS-1 with Son of Sevenless. The driver mutations are colored in magenta (recurrence > = 10) and sea green (recurrence > = 3). B. mCSM-lig values for the mutated sites. D. mCSM-protein-protein values for the interface residues. The residues present within the 5Å of the ligand binding site (GDP), are highlighted as pink circles.

https://doi.org/10.1371/journal.pone.0219935.g002

close residues are within 5Å of the ligand binding site (GDP), highlighting their functional role (Fig 2B and 2C).

We used mCSM-PPI[35], trained on effects of mutations on protein-protein interfaces, to predict the effects for the interface residues (between HRAS-1 and SOS), which are reported to have mutations (eight residues including Q61, shown as magenta in Fig 2D) in COSMIC. The majority of the interface residues were observed to have destabilizing effects for the interface. Q61, present within 5Å of the ligand GDP, is the most recurrent mutation, with a count of 659 and was observed to be mutated to seven other residues, which are all predicted to reduce ligand binding affinity using mCSM-lig (Fig 2B). Three of these (Q61R, Q61K and Q61E) are predicted to have destabilizing effects on protein-protein interactions (Fig 2D). Hence, on the one hand the driver mutations have impacts on both protein-protein and protein-ligand interactions, but on the other, because mutating functionally important residues comes with a fitness cost, the changes are to amino acids with ddG values close to zero.

Homo-dimers SMAD2

Smad2 is a receptor-regulated Smad (R-Smad), a functional class involved in ligand specific, TGF-β-cell-signaling pathways and implicated to function as tumor suppressors[43]. The pathways involving TGF-β are known to regulate cell growth, proliferation, apoptosis, differentiation and developmental pathways[44] and are initiated by cytokine binding to the TGF-β transmembrane receptors, kinase activation, recruitment of a specific R-Smad and phosphorylation of the SSXS motif (pSer motif) at the C-terminal, formation of a hetero-oligomer of R-Smad and Smad4 (ubiquitous, comediator Smad)[43,45]. The hetero-oligomer regulates the expression of genes in the nucleus in response to the specific ligand. Mutations in this pathways are known to cause human cancer and developmental disorders[4,46].

Although the unphosphorylated Smad2 is a monomer, phosphorylated Smad2 is known to form a homotrimer both *in vitro* and *in vivo*[43]. The phosphorylated C-terminus of one protomer contacts the L3/B8 loop-strand pocket of the adjacent protomer, using the two pSer residues as anchors. The residues that mediate interactions in the homotrimer are observed to be conserved in Smad4 and also there have been suggestions that the same surface of Smad2 is used to form a hetero-complex with Smad4[43,45].

We used the structure of the central domain (MH2 domain) (PDBID: 1KHX, 1.8 Å) to map the mutations from COSMIC. There are nine mutations that occur at least three times (Fig 3A), and are seen at the protein-protein interface of the heterotrimer (shown in magenta and green spheres in Fig 3B), other than S276, which is at the core of the MH2 domain of Smad2 and is responsible for the structural stability of each protomer. Most mutations at the residue positions were predicted to have a destabilizing effect at the protein-protein interface (Fig 3C) using mCSM-PPI.

Hence, the mapping of the cancer related mutations onto the surface of the MH2 domain of Smad2 implies that these mutations alter the formation of homo/hetero complex formation and hence might affect the tumor suppressor roles of Smad proteins. A similar spectrum of mutations is discussed for Smad4 and Smad3[46,47].

Small-molecule ligand-binding proteins

BRAF-MAP2K1. RAF kinases are Ser/Thr kinases known to have three isoforms in human: ARAF, BRAF and CRAF. They are involved in the MAPK (Mitogen-Activated Protein Kinase) pathway, which plays role in cell growth, ageing and differentiation. RAF kinases are activated by the GTP-bound RAS molecules, which in turn phosphorylate MAP kinases resulting in downstream cell signaling towards cell cycle progression and transformation. RAF

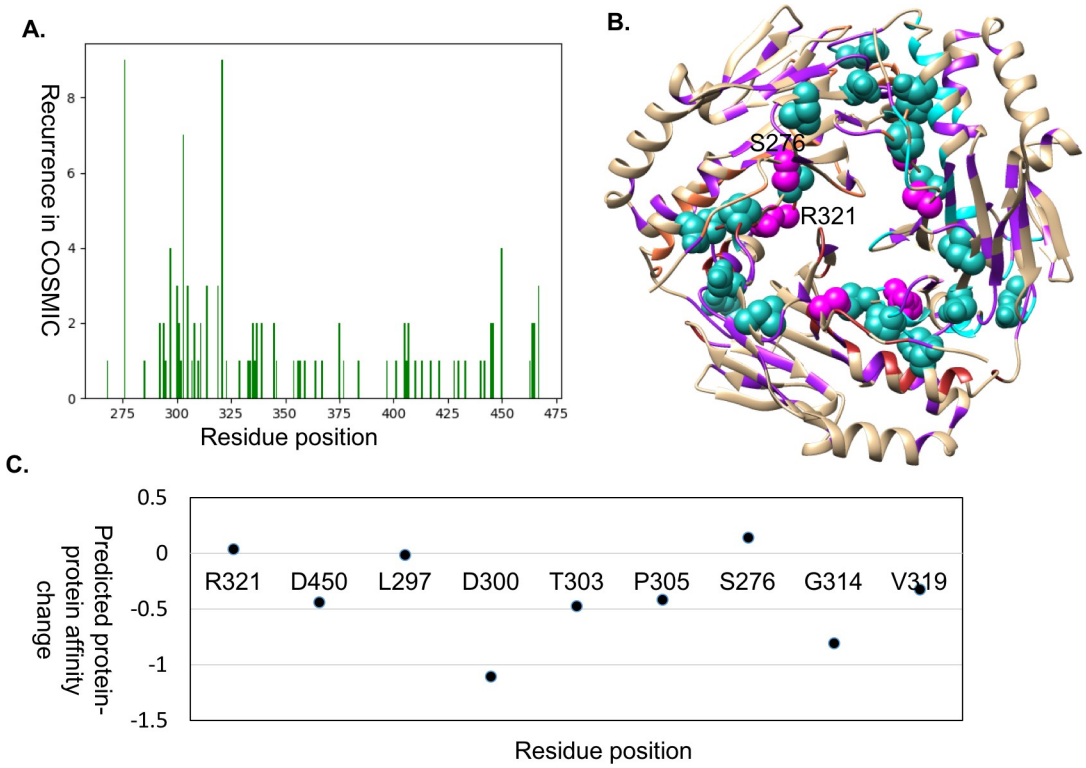


Fig 3. Mutations in Smad2. A. Frequency distribution in cancer samples in COSMIC. B. Mutations mapped on the homotrimer of Smad2. The residues that are documented to have mutations are shown in purple ribbon, and the driver sites are marked in sea green (at least three times) and magenta (more than three times) spheres. C. mCSM-protein-protein values for the most frequently mutated interface residues.

https://doi.org/10.1371/journal.pone.0219935.g003

molecules are known to have mutations that are implicated in human cancers due to their constitutive activation.

The most frequently mutated residue in BRAF (V600) is located in the kinase domain close to the sites of phosphorylation (T599 and S602[48]), which are responsible for the downstream signaling. The mutation V600E is reported to mimic the effect of phosphorylation as it has higher kinase activity than the wild type[49].

The Cancer Gene Census in COSMIC has data from 50,176 samples and 50,742 missense mutations for BRAF (Fig 4A and 4B). We used the structure with PDB IDs: 4MBJ (protein-ligand, structure of BRAF kinase domain with an inhibitor) and 4MNE (protein-protein complex, structure of BRAF kinase domain with MAP2K1) to map the mutations to understand the functional impacts of the mutations and envisage the roles of driver mutations at protein-protein and protein-ligand interactions.

We studied the effects of mutations on the inhibitor binding on the BRAF kinase domain as the recurrent mutations were observed to be present near the inhibitor-binding site (within 5Å of the inhibitor). The majority of these were predicted to decrease the inhibitor-binding affinity and hence might contribute towards resistance (Fig 4C and 4D). The 42 residue positions are reported to have a recurrence rate of at least three (sea green Fig 4D) and 18 residue positions have a recurrence of at least ten (magenta in Fig 4D).

We identified interface residues from the complex structures of BRAF, as the residues in the biological assembly of 4MNE having a Cβ-Cβ distance of 7Å or less of the homodimer interface of BRAF and the BRAF interface with MAP2K1. There are 23 interface residues at

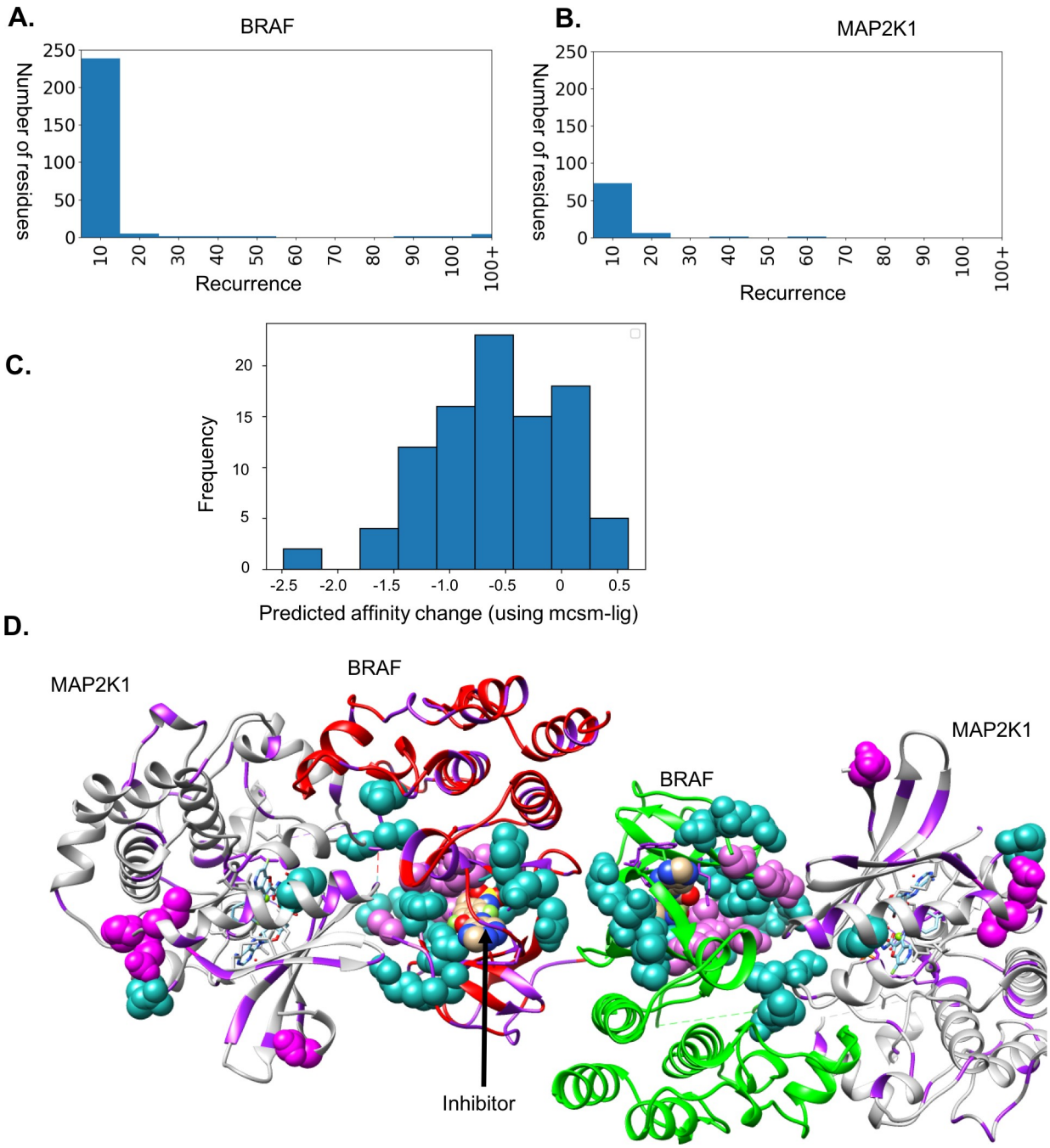


Fig 4. Mutations in BRAF. A. Frequency distribution in cancer samples in COSMIC for BRAF. B. Frequency distribution in cancer samples in COSMIC for MAP2K1. C. mCSM-lig values for the driver residues and protein-ligand interface residues. D. Mutations mapped on the complex of BRAF and MAP2K1. The residues that are documented to have mutations are shown in purple ribbon, and the driver sites are marked in sea green (at least three times) and magenta (more than three times) spheres.

https://doi.org/10.1371/journal.pone.0219935.g004

the homodimer interface and 31 residues are at the interface with MAP2K1. Of these interfacial residues, nine residues have mutations documented in the cancer gene census (Table 1). We predicted the effects of these mutations using mCSM-PPI and most of them were predicted to be destabilizing.

Table 1. Prediction of effects of mutations at the protein-protein interface using mCSM protein-protein in BRAF.

BRAF-BRAF interface			BRAF-MAP2K1 interface		
Residue position	**Recurrence**	**Average ddG (mCSM-PPI)**	**Residue position**	**Recurrence**	**Average ddG (mCSM-PPI)**
L588	3	-0.274	S614	4	0.224
L515	1	-0.628	S616	7	-0.119
D586	9	-0.427	I617	1	-0.211
R509	1	-2.634	L618	8	-0.871
			H539	1	-1.253

https://doi.org/10.1371/journal.pone.0219935.t001

The oncogenic mutations in BRAF are known to act through complicated and diverse mechanisms, for example the mutants possessing the most recurring V600E alteration increase the kinase activity of BRAF, whereas other less common mutations decrease the kinase activity but still promote the downstream phosphorylation and signalling using a CRAF-dependent pathway[50,51].

DNA-binding proteins

Androgen receptor. The androgen receptor (AR), a member of the steroid hormone nuclear receptor family, plays an important role in sexual differentiation and also has many important biological roles such as the development of the cardiovascular and immune systems. AR signaling is also reported to have a role in the development of tumors, and is an important target for prostate cancer[52]. AR has two main domains: ligand-binding domain (LBD, residue range: 668–918)[53], which binds 5-α dihydrotestosterone (DHT) and activates downstream signaling including phosphorylation of the second messenger signaling cascade and DNA-binding domain (DBD residue range: 538–629) [54], which regulates the target gene expression[55]. The N-terminal region of AR (~ 500 amino acids) is intrinsically disordered and has no defined 3D structure (Fig 5A). The structures of both these domains have been solved independently but a linker region (residue range: 630–667) does not have a structure and was modeled using (PDB IDs; 5CJ6 and 2AM9) as template and subsequently we used Modeller to link the DNA binding domain and ligand binding domain together.

There are four types of mutations observed in the AR receptor: missense substitution, insertion or deletion, partial gene deletion, and intronic mutation. We focused on the missense mutations[56], for which there are 482 unique mutations reported in COSMIC for AR, 221 unique mutations from 789 samples were mapped to the AR LBD, DNA-binding domain and linker between the two (Fig 5B, shown as purple). Most of the frequently observed mutations cluster around the ligand-binding pocket (DHT (5-aplha-dihydrotestosterone, shown in grey) binding site (Fig 5B). We predicted the effects of these mutations on the protein stability using mCSM and SDM2 (Fig 5C). The majority of the frequently observed mutations (magenta circles) were predicted to destabilise the protein.

Table 2. Details on the human proteins implicated in cancer, studied here.

Target example	Gene ID (UniProt)	PDB structure	Gene name
Ras Protein and Son of Sevenless hetero complex	P01112 and Q07889	1XD2	HRAS-1
SMAD2	Q15796	1KHX	SMAD2
BRAF and MAP2K1	P15056 and Q02750	4MBJ,4MNE	BRAF and MAP2K1
Androgen receptor	P10275	5CJ6 and 2AM9	AR
Transforming growth factor beta receptor II (TGF-R2)	P15056	1H4I, 3I44, and 1H4J	BRAF
ATP1A1 a sodium/potassium ATPase pump	P05023	2ZXE	ATP1A1

https://doi.org/10.1371/journal.pone.0219935.t002

As the LBD of AR is homodimeric, we measured the effects of the mutations on the protein-protein interactions using mCSM-PPI (Fig 5C). All mutations in the ligand-binding domain within 7Å of any atom of an interface residues between the two chains were predicted to be destabilizing the protein-protein interactions. Using mCSM-lig we estimated the impacts of mutations within 7 Å of the ligand on the ligand (DHT)-binding: they were all predicted to have detabilising effects on ligand binding (Fig 5D). These destabilising mutations cluster mainly around the ligand-binding site and are believed to cause perturbation by increasing the mobility of an adjacent helix [57]. Experimental evidence showed that mutation of F877L, T878, H875Y decrease the sensitivity of AR toward non-steroidal antagonists such as hydroxyflutamide, bicalutamide, and enzalutamide converting it into full agonist[57].

We also estimated the impacts of mutations on DNA-binding using the mCSM-NA software (Fig 5E) and all mutations (with a frequency between 4 and 6 and within 7 Å of DNA) were observed to highly reduce the DNA-binding affinity.

Membrane proteins

Transforming growth factor beta receptor II (TGF-R2). The TGF-R2 protein belongs to the TGF-β cytokine superfamily and regulates multiple cellular activities, playing a key role in development, tissue homeostasis, and immune modulation[58]. TGF-R2 consists of an extracellular domain (residue range: 25–130), a single transmembrane domain (residue range: 157–188) and a kinase domain (residue range: 190–541) in the intracellular region[59,60]. TGF-R2 forms a homo/heterodimer upon binding the ligand, which further triggers the TGF-β signaling pathway.

We modeled the structure the transmembrane domain and the missing regions between the kinase domain and the transmembrane residue range: 191–239, using (PDB IDs: 1H4I, 3I44, and 1H4J) as template. We then mapped the mutation data on to the modeled structure (Fig 6A). The potential driver mutations with frequencies more than eight, shown as magenta circles in Fig 6B and 6C, were predicted to have a destabilizing effect on the protein stability using mCSM (Fig 6B) and SDM2 (Fig 6C). The most frequently mutated residue R528 (67 times, to residues Gly, His, Leu and Phe) is a key residue for maintaining the stability of the kinase domain (highly buried and forms salt bridge with E428, Fig 6A). Mutating R528 to other residue types will affect the protein stability and is also predicted as highly destabilizing with mCSM and SDM2 (marked with oval in Fig 6B). Experimental evidence has shown that mutating R528 leads to conformational changes and hence alters the kinase function[61].

Transmembrane Pump: ATP1A1 a sodium/potassium ATPase pump. Na^+/K^+ ATPase, a sodium-potassium pump, expressed in all animal cells, belongs to the class IIC of P-type ATPases that utilize ATP[62]. Na^+/K^+ ATPase assists in maintaining the pH as well as a low sodium ion intracellular concentration and a high potassium extracellular concentration. Na^+/K^+ ATPase, one of the most important active transporter, works towards maintaining a resting membrane potential and signal transduction[63].

We modeled ATP1A1 (Fig 7A) using the sodium-potassium pump (PDB ID: 2ZXE[64]) structure at 2.4AÅ resolution as a template (89% percent identity and coverage of 96%). There are three alpha-domains present in the intracellular region of ATP1A1 model: A-domain, N-domain, and P-domain. The alpha-N domain and the alpha-A domain are stabilized by a salt bridge interaction[64] between E223 and R551 (Fig 7A, shown in black spheres). ATP binding occurs close to this salt bridge and mutation of E223 and R551 will eliminate the ATP binding. Pharmacologically Na^+/K^+ ATPase can be inhibited by digoxin, which is used to treat heart failure. Na^+/K^+ ATPase has been suggested as a potential chemotherapy target for cancer[65].

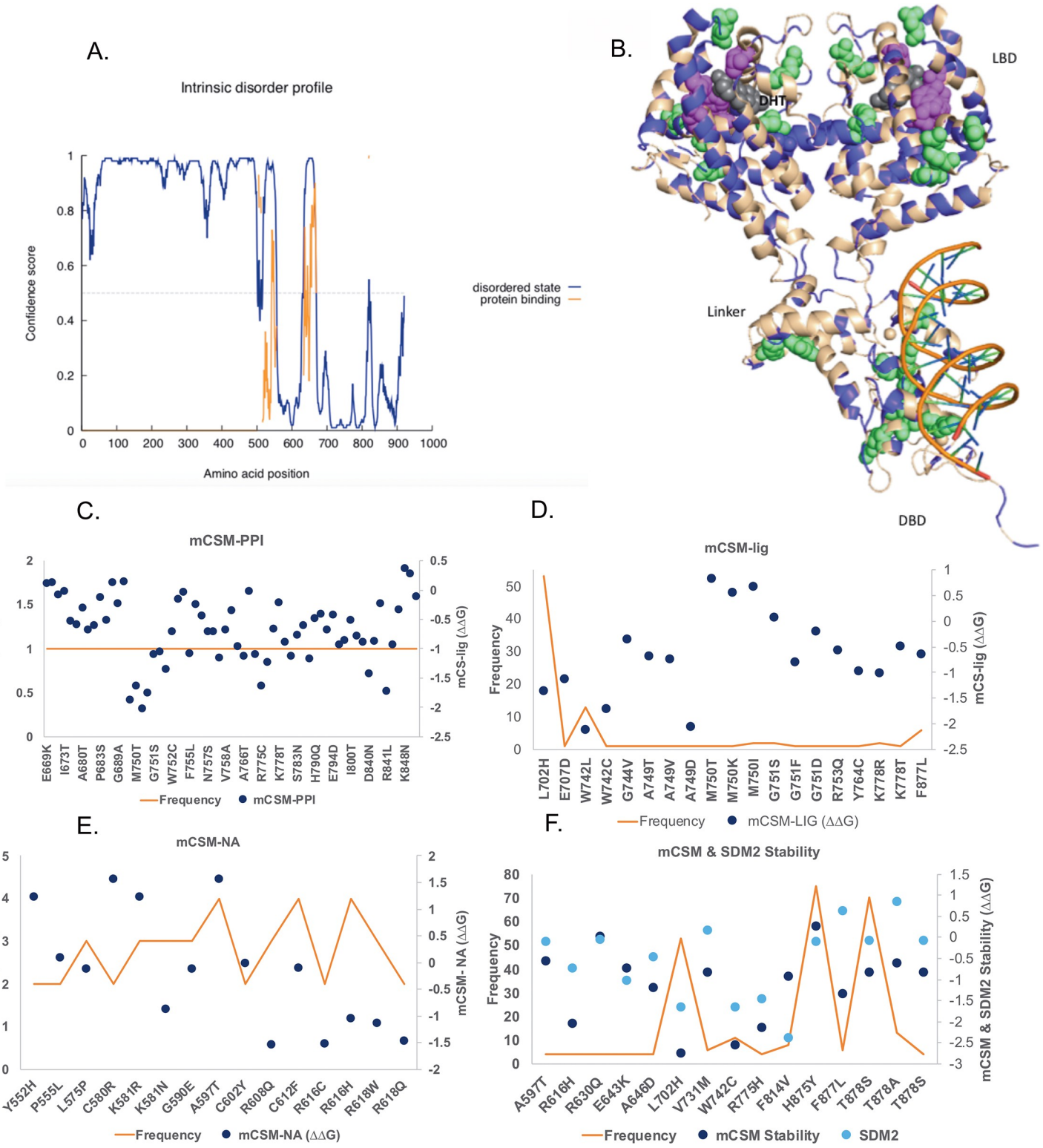


Fig 5. Mutations in androgen receptor. A. The intrinsically disordered region of the androgen receptor (predicted using PSIPRED). B. Missense mutations (shown in purple) from COSMIC mapped on to the modeled structure of androgen receptor homodimer structure. Driver mutation sites (recurrence > = 13), indicated by magenta spheres, are located mainly around the ligand-binding domain and the residues with a mutation frequency between 4 and 6 are shown in light green. C. The changes in protein-protein binding affinity predicted using mCSM-PPI of the residues present in the homodimer interface (from both chains). D. Changes in the ligand binding affinity predicted using mCSM-lig for the residues present within the 7Å of DHT ligand and the most recurring mutations are highlighted in magenta. E.

Changes in the protein-DNA binding affinity predicted using mCSM-NA for mutations that occur within 7Å of the DNA and the most frequent mutations highlighted in light green. F. mCSM, SDM2 stability predictions of mutation reported more than 4 times in COSMIC.

https://doi.org/10.1371/journal.pone.0219935.g005

150 unique mutations in ATP1A1 have been observed in 210 samples recorded in the COSMIC database. Four mutations (V332G, P333L, E786V and D747N, marked as pink circles in Fig 7A), are present within 5 AÅ of the three potassium ions (shown as gold spheres, Fig 7A). All missense mutations from COSMIC were mapped onto the Na/K model (Fig 7A). The most frequent mutation is present in the transmembrane region, L104R (M1 helix, shown in magenta, Fig 7A) has been reported 49 times and is also predicted as highly destabilizing by mCSM (ddG = -1.72) and SDM (ddG = -2.73). L292 in M3 and G99 in M1 are not observed to be frequently mutated as they function as central residues for the movement of M1 to open the gate for ions to enter into the cationic pocket, whereas E334 in M4 is part of the gate that binds to a potassium ion in the occluded stage[66] (Fig 7A). Since the movements of transmembrane domains are essential for ions to be transported in and out of the cells, mutations around the cationic pocket or in the trans-membrane region, which has to move to allow ions to be transported, will disrupt the function of the Na/K transporter. There are multiple studies on L104R mutation indicating that R104 creates a positive charge causing structure alteration around cationic pocket, as a result of which the potassium binding pocket is disrupted and cell depolarization observed[67,68]. Fig 7B, highlights that most frequently mutated residues are potential drivers predicted to have a destabilizing effect on the protein stability using mCSM and SDM (Fig 7C).

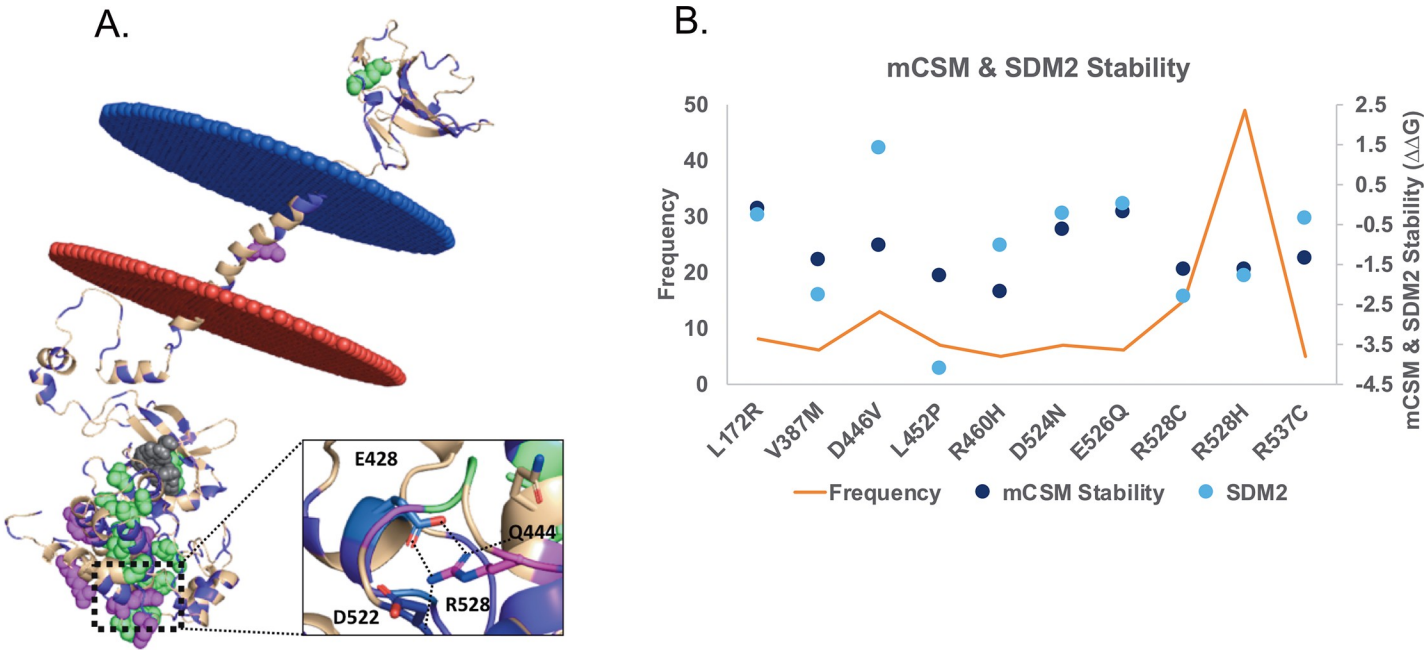
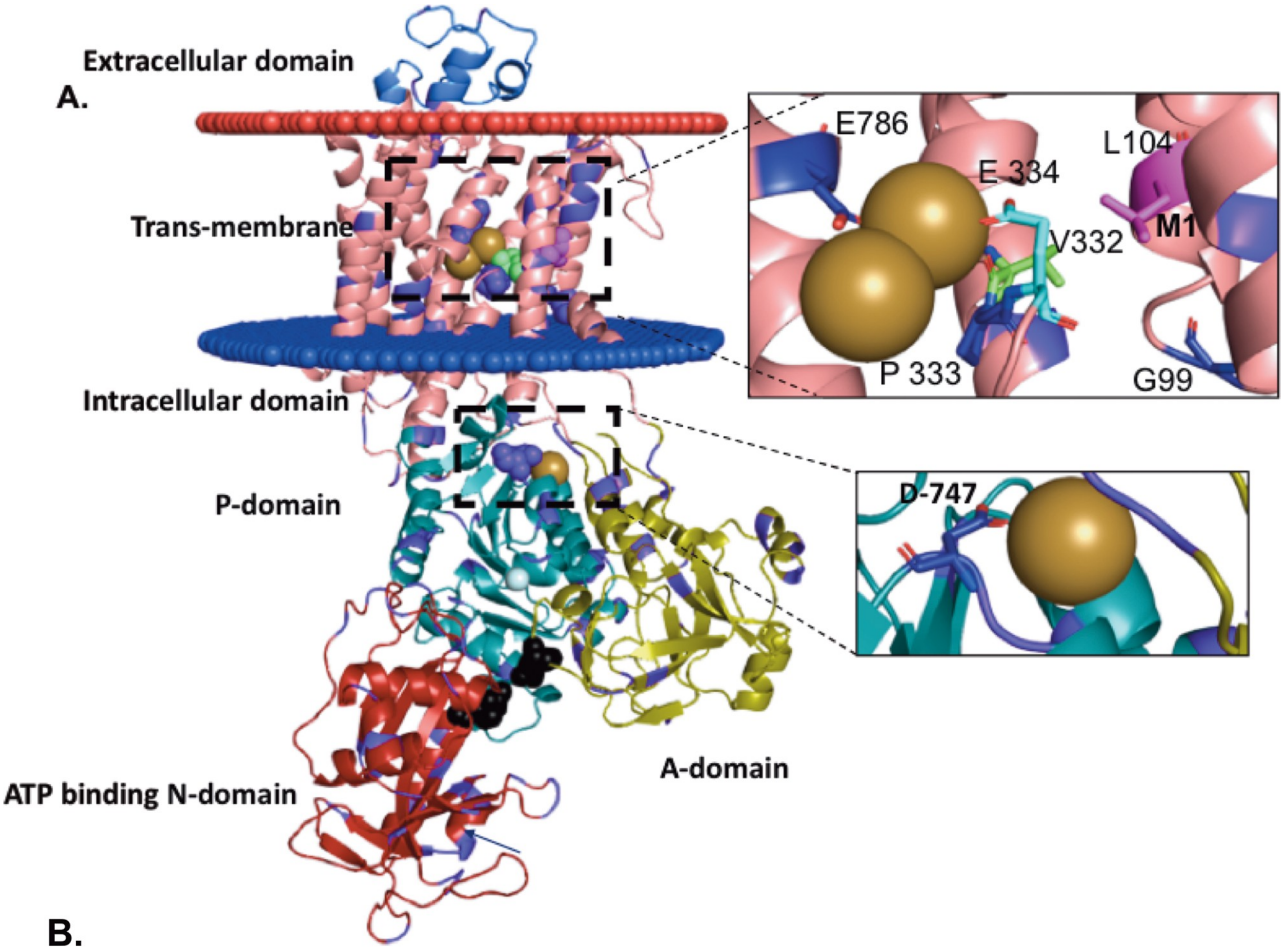


Fig 6. Mutations in TGF. A. Missense mutations (in purple) from COSMIC were mapped on to the modeled structure of *TGF-R2*. Residues with a mutation frequency between 4 and 6 are shown in light green spheres, and the residues with mutation frequency more than 8 are shown in magenta. The interactions of the most frequently mutated residue R528 with its surrounding residues are shown. B. The changes in the energy (ddG) corresponding to the stability of the protein structure predicted using mCSM, SDM2 for all missense mutations and the most frequent mutations are shown in magenta.

https://doi.org/10.1371/journal.pone.0219935.g006


A.
Extracellular domain
Trans-membrane
Intracellular domain
P-domain
A-domain
ATP binding N-domain
E786
L104
E 334
V332
M1
P 333
G99
D-747

B.

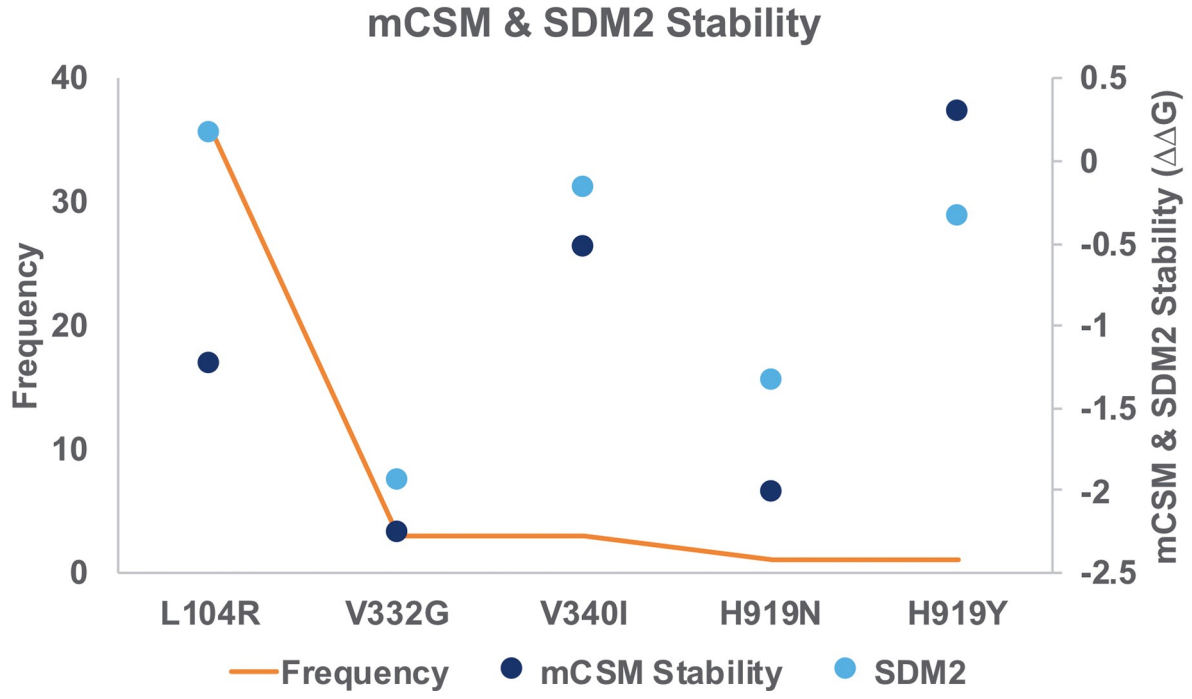

mCSM & SDM2 Stability
Frequency
mCSM & SDM2 Stability (ΔΔG)
40
30
20
10
0
0.5
0
-0.5
-1
-1.5
-2
-2.5
L104R
V332G
V340I
H919N
H919Y
Frequency
mCSM Stability
SDM2

Fig 7. Mutations in sodium potassium pump. A. The modeled structure of sodium potassium pump ATP1A1; each domain labelled in the figure is represented in a different colour. Mutation data from the COSMIC database are mapped on to the structure in purple, the driver mutation, L104, is shown in magenta, the driver sites (recurrence > = 3) are marked in sea green. The potassium ions are shown in gold spheres and the magnesium ions in white spheres. Mutated residues within 5Å of potassium ions are shown in stick. The salt bridge connecting the N-domain to the A-domain is shown in black spheres. B. The changes in the energy (ddG) corresponding to the stability of the protein structure predicted using mCSM and SDM2.

https://doi.org/10.1371/journal.pone.0219935.g007

Discussion and future perspectives

Identifying driver mutations in cancer targets is essential to guide new therapeutics. However, mutations can act at a distance from ligand binding sites not only in well-defined allosteric sites in the same subunit by changing stability but also through disturbing protein interactions with another protein, nucleic acid, metal ion or ligand. Methods to define drivers include determining their impacts on physico-chemical properties as well as understanding the roles of the mutated side chains in protein 3D structure, e.g. solvent accessibility, hydrogen bonding, and surface accessibility. There are several computational approaches to prediction of the impact of mutations on protein stability, for example SDM[31,32], and mCSM[35,36,38]. However, the lack of defined 3D structures in complexes (heterodimer, homodimer, DNA, RNA) makes it difficult to predict the impact of mutations on protein function. Usually mutations in a conserved region are recognized as drivers, whereas mutations in a non-conserved region are classified as passengers. A major challenge is to identify mutations that are outside the conserved region but lead to cancer progression. Destabilizing effects of the glioblastoma missense mutations have been observed in the protein-protein and protein-ligand interfaces [69–71]. With respect to the systems studied here most mutations appear in the interface, binding site, and between domains in ATP1A1, SMAD2, and BRAF-MAP2K1, but others can allosterically affect these interactions. It is essential to model full multicomponent complexes (heterodimer, homodimer, DNA, RNA) in order to explain the impacts on the interface, DNA, and RNA binding. These are particularly important for predictive algorithms that depend on structure such as those encoded in software such as mCSM and SDM.

Here, we have analysed the effects of the most recurrent mutations on protein-protein (hetero-Ras protein with Son of Sevenless protein and homo-SMAD2 homodimer), protein-DNA (androgen receptor with its target DNA) and protein-ligand (BRAF kinase with an inhibitor) interfaces. In the protein-ligand cases, many of the most recurrent mutations were clustered around the ligand-binding site and were predicted to decrease the inhibitor-binding affinity. Similarly, all mutations with high frequency and within the 7 Å of DNA were observed to highly reduce the DNA-binding activity. In the protein-protein complex of BRAF-kinase dimer with MAP2K1, both homo and hetero interfaces are tightly packed and comprise of 23 and 31 interface residues respectively. The mutations in these interface residues were predicted to be destabilizing and hence affect cell signalling and function.

3D hotspot clustering is one of the methods used to study driver mutations. Recently 3D structural information has been used to identify driver mutations in cancer and other diseases. Using the Fragment-Hotspot program[72] to identify druggable sites in conjunction with Hotspt3D[73], HotMAPS[74], and Mutation3D[67], which use 3D structure to identify mutation clusters in cancer should give valuable information on identifying driver mutations. The Pan-Cancer analysis has shown that structure-based approaches are more reliable but less sensitive than sequence approaches in identifying driver mutations than other methods for the dataset used[13]. In our example of the Ras dimer with SOS protein, a residue with a large number of mutations, Q61, was a part of a cluster of spatially close residues, which has other residues with low mutation frequency. The residues with low frequency are within 5Å of the ligand binding site (GDP), highlighting their functional role. However, this method is limited

to a good 3D structure, defined experimentally or from homology, and cannot be applied to mutations that occur in intrinsic-disordered regions of the protein, which occur very often in proteins from the Cancer Gene Census.

We have described an approach that will help in predicting those mutations that are damaging and functionally important. This will help in identifying potential driver mutations and prioritize mutations for experimental testing which will ultimately help in guiding drug design.

Materials and methods

Protein structure prediction

Where 3D structures are not available for genes, they can be constructed using a variety of software available to search for homologues of known structure and to use appropriate structures as templates for comparative modelling. We have used our in-house modeling pipeline VIVACE[75], which is built in Python using the Ruffus module[76], and combines template searching, single or multiple template alignment, modelling, model quality assessment (NDOPE, GA341, SOAP), optional disordered-region predictions into a single automated program that can be easily parallelized in multiprocessor systems.

In order to identify the homologues, a sequence-structure homology recognition program, FUGUE[77], uses environment-specific substitution tables, which take into account both amino acid sequence information and the local structural environment (secondary structure, solvent accessibility and sidechain interactions) to identify sequences that are compatible with a known protein fold. The search is facilitated by the TOCCATA database (http://mordred.bioc.cam.ac.uk/toccata/), which includes profiles of aligned structures of homologues from the PDB and is organized for use with FUGUE. Originally TOCCATA profiles were for domains assigned from SCOP[78] and CATH superfamilies[79]. The recent VIVACE update now includes all PDBs grouped by CD-HIT[80]. A PSI-BLAST[81] search is run concurrently with FUGUE, thus preventing VIVACE from missing templates that have been submitted to PDB since the most recent CATH and SCOP updates. The total number of PDB domain structures in TOCATTA has increased from 228,000 to 475,000 (the figures refer to the number of structure domains as represented in SCOP[78] and CATH[79]), with many not associated with superfamilies but ensuring access to recent structures. Profiles that share the same CATH–SCOP consensus are also linked together during the template selection phase, where FUGUE is used to consider all the templates in a profile to find the best matches. This is to mitigate the problem of trapping the best template in a mediocre profile.

Following the template-selection phase, up to five of the best templates are picked for alignment using BATON, a streamlined version of the program COMPARER[82]. The resultant alignment is finally used to create the model using MODELLER[83].

An average of ~four models of 202 proteins without crystal structures were produced using the VIVACE pipeline, with ~60% built from more from more than one homologue and with an average FUGUE z-score of 13.97. The average percentage identity of templates, calculated for the final alignment made by BATON between the model and the template structures, was 29.6%, while the average PID of the closest homologue for each gene was 54.2%. Taking only the longest model for each gene, the average coverage is 54.0% and average length 305 residues.

Mapping mutations on the protein structures

Chimera[84] and PyMol (https://pymol.org/2/) were used to view the 3D-structure of the protein (Table 2) and mutation positions were obtained from the CGC page of the COSMIC

database (https://cancer.sanger.ac.uk/census), and the search for the gene name was performed in the search tool. The mutations retrieved from the CGC were then mapped onto the structure.

Predicting effects of mutations

Upon modeling we mapped the mutations from the COSMIC database onto the sequences and 3D-structures to study their effects on protein structure and function using our statistical and machine-learning based methods, namely SDM[31,32] and mCSM[35,36,38] respectively, to measure the effects of mutations on protein stability and protein-protein, protein nucleic acid or protein-ligand interactions.

Author Contributions

Conceptualization: Sony Malhotra, Tom L. Blundell.

Data curation: Sony Malhotra, Ali F. Alsulami.

Formal analysis: Sony Malhotra, Ali F. Alsulami, Yang Heiyun.

Investigation: Sony Malhotra, Ali F. Alsulami.

Methodology: Sony Malhotra.

Software: Bernardo Montano Ochoa.

Supervision: Sony Malhotra, Tom L. Blundell.

Writing – original draft: Sony Malhotra, Ali F. Alsulami, Tom L. Blundell.

Writing – review & editing: Sony Malhotra, Harry Jubb, Simon Forbes, Tom L. Blundell.

References

1. GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet Lond Engl. 2016; 388: 1459–1544. https://doi.org/10.1016/S0140-6736(16)31012-1
2. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016; 17: 333–351. https://doi.org/10.1038/nrg.2016.49 PMID: 27184599
3. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017; 45: D777–D783. https://doi.org/10.1093/nar/gkw1121 PMID: 27899578
4. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004; 4: 177–183. https://doi.org/10.1038/nrc1299 PMID: 14993899
5. Jubb HC, Saini HK, Verdonk ML, Forbes SA. COSMIC-3D provides structural perspectives on cancer genetics for drug discovery. Nat Genet. 2018; 50: 1200–1202. https://doi.org/10.1038/s41588-018-0214-9 PMID: 30158682
6. Zhang J, Liu J, Sun J, Chen C, Foltz G, Lin B. Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing. Brief Bioinform. 2014; 15: 244–255. https://doi.org/10.1093/bib/bbt042 PMID: 23818492
7. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499: 214–218. https://doi.org/10.1038/nature12213 PMID: 23770567
8. Evans P, Avey S, Kong Y, Krauthammer M. Adjusting for background mutation frequency biases improves the identification of cancer driver genes. IEEE Trans Nanobioscience. 2013; 12: 150–157. https://doi.org/10.1109/TNB.2013.2263391 PMID: 23694700
9. Makova KD, Hardison RC. The effects of chromatin organization on variation in mutation rates in the genome. Nat Rev Genet. 2015; 16: 213–223. https://doi.org/10.1038/nrg3890 PMID: 25732611

10. Korthauer KD, Kendziorski C. MADGiC: a model-based approach for identifying driver genes in cancer. Bioinformatics. 2015; 31: 1526–1535. https://doi.org/10.1093/bioinformatics/btu858 PMID: 25573922
11. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 2012; 40: e169. https://doi.org/10.1093/nar/gks743 PMID: 22904074
12. Stacey SN, Sulem P, Jonasdottir A, Masson G, Gudmundsson J, Gudbjartsson DF, et al. A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. Nat Genet. 2011; 43: 1098–1103. https://doi.org/10.1038/ng.926 PMID: 21946351
13. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018; 173: 371–385.e18. https://doi.org/10.1016/j.cell.2018.02.060 PMID: 29625053
14. Gao J, Chang MT, Johnsen HC, Gao SP, Sylvester BE, Sumer SO, et al. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med. 2017;9. https://doi.org/10.1186/s13073-017-0399-z
15. Nishi H, Tyagi M, Teng S, Shoemaker BA, Hashimoto K, Alexov E, et al. Cancer missense mutations alter binding properties of proteins and their interaction networks. PloS One. 2013; 8: e66273. https://doi.org/10.1371/journal.pone.0066273 PMID: 23799087
16. Jubb HC, Pandurangan AP, Turner MA, Ochoa-Montaño B, Blundell TL, Ascher DB. Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol. 2017; 128: 3–13. https://doi.org/10.1016/j.pbiomolbio.2016.10.002 PMID: 27913149
17. David A, Sternberg MJE. The Contribution of Missense Mutations in Core and Rim Residues of Protein–Protein Interfaces to Human Disease. J Mol Biol. 2015; 427: 2886–2898. https://doi.org/10.1016/j.jmb.2015.07.004 PMID: 26173036
18. Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, et al. Unexplored therapeutic opportunities in the human genome. Nat Rev Drug Discov. 2018; 17: 317–332. https://doi.org/10.1038/nrd.2018.14 PMID: 29472638
19. Eddy SR. Accelerated Profile HMM Searches. PLOS Comput Biol. 2011; 7: e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361
20. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016; 44: D279–D285. https://doi.org/10.1093/nar/gkv1344 PMID: 26673716
21. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinforma Oxf Engl. 2015; 31: 857–863. https://doi.org/10.1093/bioinformatics/btu744
22. Stehr H, Jang S-HJ, Duarte JM, Wierling C, Lehrach H, Lappe M, et al. The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. Mol Cancer. 2011; 10: 54. https://doi.org/10.1186/1476-4598-10-54 PMID: 21575214
23. Ryslik GA, Cheng Y, Cheung K-H, Bjornson RD, Zelterman D, Modis Y, et al. A spatial simulation approach to account for protein structure when identifying non-random somatic mutations. BMC Bioinformatics. 2014; 15: 231. https://doi.org/10.1186/1471-2105-15-231 PMID: 24990767
24. Porta-Pardo E, Garcia-Alonso L, Hrabe T, Dopazo J, Godzik A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. PLOS Comput Biol. 2015; 11: e1004518. https://doi.org/10.1371/journal.pcbi.1004518 PMID: 26485003
25. Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 2006; 34: D204–206. https://doi.org/10.1093/nar/gkj103 PMID: 16381846
26. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 2006; 7: 61–80. https://doi.org/10.1146/annurev.genom.7.080505.115630 PMID: 16824020
27. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7: 248–249. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512
28. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D. BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations. Nucleic Acids Res. 2013; 41: W333–339. https://doi.org/10.1093/nar/gkt450 PMID: 23723246
29. Brender JR, Zhang Y. Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. PLoS Comput Biol. 2015; 11: e1004494. https://doi.org/10.1371/journal.pcbi.1004494 PMID: 26506533
30. Quan L, Lv Q, Zhang Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinforma Oxf Engl. 2016; 32: 2936–2946. https://doi.org/10.1093/bioinformatics/btw361

31. Pandurangan AP, Ochoa-Montaño B, Ascher DB, Blundell TL. SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res. 2017; https://doi.org/10.1093/nar/gkx439

32. Worth CL, Preissner R, Blundell TL. SDM—a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res. 2011; 39: W215–222. https://doi.org/10.1093/nar/gkr363 PMID: 21593128

33. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinforma Oxf Engl. 2009; 25: 2537–2543. https://doi.org/10.1093/bioinformatics/btp445

34. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics. 2011; 12: 151. https://doi.org/10.1186/1471-2105-12-151 PMID: 21569468

35. Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014; 30: 335–342. https://doi.org/10.1093/bioinformatics/btt691 PMID: 24281696

36. Pires DEV, 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; 6: 29575. https://doi.org/10.1038/srep29575 PMID: 27384129

37. Pires DEV, Blundell TL, Ascher DB. Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res. 2015; 43: D387–391. https://doi.org/10.1093/nar/gku966 PMID: 25324307

38. Pires DEV, Ascher DB. mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res. 2017; https://doi.org/10.1093/nar/gkx236

39. Dourado DFAR Flores SC. A multiscale approach to predicting affinity changes in protein-protein interfaces. Proteins. 2014; 82: 2681–2690. https://doi.org/10.1002/prot.24634 PMID: 24975440

40. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins. 2011; 79: 830–838. https://doi.org/10.1002/prot.22921 PMID: 21287615

41. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000; 100: 57–70. https://doi.org/10.1016/s0092-8674(00)81683-9 PMID: 10647931

42. Sondermann H, Soisson SM, Boykevisch S, Yang S-S, Bar-Sagi D, Kuriyan J. Structural analysis of autoinhibition in the Ras activator Son of sevenless. Cell. 2004; 119: 393–405. https://doi.org/10.1016/j.cell.2004.10.005 PMID: 15507210

43. Wu JW, Hu M, Chai J, Seoane J, Huse M, Li C, et al. Crystal structure of a phosphorylated Smad2. Recognition of phosphoserine by the MH2 domain and insights on Smad function in TGF-beta signaling. Mol Cell. 2001; 8: 1277–1289. PMID: 11779503

44. Heldin CH, Miyazono K, ten Dijke P. TGF-beta signalling from cell membrane to nucleus through SMAD proteins. Nature. 1997; 390: 465–471. https://doi.org/10.1038/37284 PMID: 9393997

45. Souchelnytskyi S, Tamaki K, Engström U, Wernstedt C, ten Dijke P, Heldin CH. Phosphorylation of Ser465 and Ser467 in the C terminus of Smad2 mediates interaction with Smad4 and is required for transforming growth factor-beta signaling. J Biol Chem. 1997; 272: 28107–28115. https://doi.org/10.1074/jbc.272.44.28107 PMID: 9346966

46. Fleming NI, Jorissen RN, Mouradov D, Christie M, Sakthianandeswaren A, Palmieri M, et al. SMAD2, SMAD3 and SMAD4 Mutations in Colorectal Cancer. Cancer Res. 2013; 73: 725–735. https://doi.org/10.1158/0008-5472.CAN-12-2706 PMID: 23139211

47. Shi Y, Hata A, Lo RS, Massagué J, Pavletich NP. A structural basis for mutational inactivation of the tumour suppressor Smad4. Nature. 1997; 388: 87–93. https://doi.org/10.1038/40431 PMID: 9214508

48. Zhang BH, Guan KL. Activation of B-Raf kinase requires phosphorylation of the conserved residues Thr598 and Ser601. EMBO J. 2000; 19: 5429–5439. https://doi.org/10.1093/emboj/19.20.5429 PMID: 11032810

49. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, et al. Mutations of the BRAF gene in human cancer. Nature. 2002; 417: 949–954. https://doi.org/10.1038/nature00766 PMID: 12068308

50. Haarberg HE, Smalley KSM. Resistance to Raf inhibition in cancer. Drug Discov Today Technol. 2014; 11: 1–116. https://doi.org/10.1016/j.ddtec.2014.03.013

51. Holderfield M, Deuker MM, McCormick F, McMahon M. Targeting RAF kinases for cancer therapy: BRAF mutated melanoma and beyond. Nat Rev Cancer. 2014; 14: 455–467. https://doi.org/10.1038/nrc3760 PMID: 24957944

52. Davey RA, Grossmann M. Androgen Receptor Structure, Function and Biology: From Bench to Bedside. Clin Biochem Rev. 2016; 37: 3–15. PMID: 27057074

53. Nadal M, Prekovic S, Gallastegui N, Helsen C, Abella M, Zielinska K, et al. Structure of the homodimeric androgen receptor ligand-binding domain. Nat Commun. 2017; 8: 14388. https://doi.org/10.1038/ncomms14388 PMID: 28165461

54. Shaffer PL, Jivan A, Dollins DE, Claessens F, Gewirth DT. Structural basis of androgen receptor binding to selective androgen response elements. Proc Natl Acad Sci U S A. 2004; 101: 4758–4763. https://doi.org/10.1073/pnas.0401123101 PMID: 15037741

55. Gao W, Bohl CE, Dalton JT. Chemistry and structural biology of androgen receptor. Chem Rev. 2005; 105: 3352–3370. https://doi.org/10.1021/cr020456u PMID: 16159155

56. Brinkmann AO, Jenster G, Ris-Stalpers C, van der Korput JA, Brüggenwirth HT, Boehmer AL, et al. Androgen receptor mutations. J Steroid Biochem Mol Biol. 1995; 53: 443–448. PMID: 7626493

57. Lallous N, Volik SV, Awrey S, Leblanc E, Tse R, Murillo J, et al. Functional analysis of androgen receptor mutations that confer anti-androgen resistance identified in circulating cell-free DNA from prostate cancer patients. Genome Biol. 2016; 17: 10. https://doi.org/10.1186/s13059-015-0864-1 PMID: 26813233

58. Neuzillet C, Tijeras-Raballand A, Cohen R, Cros J, Faivre S, Raymond E, et al. Targeting the TGFβ pathway for cancer therapy. Pharmacol Ther. 2015; 147: 22–31. https://doi.org/10.1016/j.pharmthera.2014.11.001 PMID: 25444759

59. Tebben AJ, Ruzanov M, Gao M, Xie D, Kiefer SE, Yan C, et al. Crystal structures of apo and inhibitor-bound TGFβR2 kinase domain: insights into TGFβR isoform selectivity. Acta Crystallogr Sect Struct Biol. 2016; 72: 658–674. https://doi.org/10.1107/S2059798316003624

60. Deep S, Walker KP, Shu Z, Hinck AP. Solution structure and backbone dynamics of the TGFbeta type II receptor extracellular domain. Biochemistry. 2003; 42: 10126–10139. https://doi.org/10.1021/bi034366a PMID: 12939140

61. Horbelt D, Guo G, Robinson PN, Knaus P. Quantitative analysis of TGFBR2 mutations in Marfan-syndrome-related disorders suggests a correlation between phenotypic severity and Smad signaling activity. J Cell Sci. 2010; 123: 4340–4350. https://doi.org/10.1242/jcs.074773 PMID: 21098638

62. Katz AI. Renal Na-K-ATPase: its role in tubular sodium and potassium transport. Am J Physiol. 1982; 242: F207–219. https://doi.org/10.1152/ajprenal.1982.242.3.F207 PMID: 6278949

63. Clausen MV, Hilbers F, Poulsen H. The Structure and Function of the Na,K-ATPase Isoforms in Health and Disease. Front Physiol. 2017; 8: 371. https://doi.org/10.3389/fphys.2017.00371 PMID: 28634454

64. Shinoda T, Ogawa H, Cornelius F, Toyoshima C. Crystal structure of the sodium-potassium pump at 2.4 A resolution. Nature. 2009; 459: 446–450. https://doi.org/10.1038/nature07939 PMID: 19458722

65. Mijatovic T, Dufrasne F, Kiss R. Na+/K+-ATPase and cancer. Pharm Pat Anal. 2012; 1: 91–106. https://doi.org/10.4155/ppa.12.3 PMID: 24236716

66. Einholm AP, Toustrup-Jensen M, Andersen JP, Vilsen B. Mutation of Gly-94 in transmembrane segment M1 of Na+,K+-ATPase interferes with Na+ and K+ binding in E2P conformation. Proc Natl Acad Sci U S A. 2005; 102: 11254–11259. https://doi.org/10.1073/pnas.0501201102 PMID: 16049100

67. Meyer DJ, Gatto C, Artigas P. On the effect of hyperaldosteronism-inducing mutations in Na/K pumps. J Gen Physiol. 2017; 149: 1009–1028. https://doi.org/10.1085/jgp.201711827 PMID: 29030398

68. Beuschlein F, Boulkroun S, Osswald A, Wieland T, Nielsen HN, Lichtenauer UD, et al. Somatic mutations in ATP1A1 and ATP2B3 lead to aldosterone-producing adenomas and secondary hypertension. Nat Genet. 2013; 45: 440–444, 444e1-2. https://doi.org/10.1038/ng.2550 PMID: 23416519

69. Goncearenco A, Li M, Simonetti FL, Shoemaker BA, Panchenko AR. Exploring Protein-Protein Interactions as Drug Targets for Anti-cancer Therapy with In Silico Workflows. Methods Mol Biol Clifton NJ. 2017; 1647: 221–236. https://doi.org/10.1007/978-1-4939-7201-2_15

70. Li M, Goncearenco A, Panchenko AR. Annotating Mutational Effects on Proteins and Protein Interactions: Designing Novel and Revisiting Existing Protocols. Methods Mol Biol Clifton NJ. 2017; 1550: 235–260. https://doi.org/10.1007/978-1-4939-6747-6_17

71. Nishi H, Fong JH, Chang C, Teichmann SA, Panchenko AR. Regulation of protein–protein binding by coupling between phosphorylation and intrinsic disorder: analysis of human protein complexes. Mol Biosyst. 2013; https://doi.org/10.1039/C3MB25514J

72. Radoux CJ, Olsson TSG, Pitt WR, Groom CR, Blundell TL. Identifying Interactions that Determine Fragment Binding at Protein Hotspots. J Med Chem. 2016; https://doi.org/10.1021/acs.jmedchem.5b01980

73. Niu B, Scott AD, Sengupta S, Bailey MH, Batra P, Ning J, et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat Genet. 2016; 48: 827–837. https://doi.org/10.1038/ng.3586 PMID: 27294619

74. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res. 2016; 76: 3719–3731. https://doi.org/10.1158/0008-5472.CAN-15-3190

75. Ochoa-Montaño B, Mohan N, Blundell TL. CHOPIN: a web resource for the structural and functional proteome of Mycobacterium tuberculosis. Database. 2015;2015: bav026. https://doi.org/10.1093/database/bav026

76. Goodstadt L. Ruffus: a lightweight Python library for computational pipelines. 2010; 26: 2778–2779. https://doi.org/10.1093/bioinformatics/btq524 PMID: 20847218

77. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol. 2001; 310: 243–257. https://doi.org/10.1006/jmbi.2001.4762 PMID: 11419950

78. Hubbard TJP, Ailey B, Brenner SE, Murzin AG, Chothia C. SCOP: A structural classification of proteins database. Nucleic Acids Research. 1999. pp. 254–256. https://doi.org/10.1093/nar/27.1.254 PMID: 9847194

79. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, et al. CATH: Comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 2015; 43: D376–D381. https://doi.org/10.1093/nar/gku947 PMID: 25348408

80. Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22: 1658–1659. https://doi.org/10.1093/bioinformatics/btl158 PMID: 16731699

81. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25: 3389–3402. https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694

82. Salit A, Blundellt TL. Definition of General Topological Equivalence in Protein Structures A Procedure Involving Comparison of Properties and Relationships through Simulated Annealing and Dynamic Programming. 1990; 403–428.

83. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993; 234: 779–815. https://doi.org/10.1006/jmbi.1993.1626 PMID: 8254673

84. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem. 2004; 25: 1605–1612. https://doi.org/10.1002/jcc.20084 PMID: 15264254