Review

Protein domain-based approaches for the identification and prioritization of therapeutically actionable cancer variants

Elisabetta Grillo a,*, Cosetta Ravelli a, Michela Corsini a, Luca Zammataro b, Stefania Mitola a,*

a Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy
b Division of Artificial Intelligence Systems for Immunoinformatics, Kiromic BioPharma, Inc., Houston, USA

ARTICLE INFO

Keywords:
Protein domain
Somatic mutation
Cancer variant
Low frequency mutation

ABSTRACT

The tremendous number of cancer variants that can be detected by NGS analyses has required the development of computational approaches to prioritize mutations on the basis of their biological and clinical significance. Standard strategies take a gene-centric approach to the problem, allowing exclusively the identification of highly frequent variants. On the contrary, protein domain (PD)-based approaches allow to identify functionally relevant low frequency variants by searching for mutations that recur on analogous residues across homologous proteins (i.e. containing the same PD). Such approaches enable to transfer information about the effects and druggability from one known mutation to unknown ones. Here we describe how PD-based strategies work, and discuss how they could be exploited for mutation prioritization.

The principle that mutations clustered on specific residues of PDs have the same functional consequences and are therapeutically actionable in a similar manner could help the choice of patient-specific targeted drugs, eventually improving the management of cancer patients.

1. Introduction

Precision oncology refers to the possibility of using therapeutics that are specifically tailored to the genetic changes in each person’s cancer. There is widespread enthusiasm for precision oncology, also fueled by the increasing number of cancer-associated genetic changes that can be detected in large cohorts of cancer specimens [e.g., The Cancer Genome Atlas (TCGA) [1], the Cancer Genome Project (CGP) [2], the International Cancer Genome Consortium [3] and the Catalogue of Somatic Mutations in Cancer (COSMIC) [4] through next-generation sequencing (NGS) technologies and by the increasing number of targeted agents becoming available. Understanding the complete landscape of mutations that drive cancer remains an open challenge for precision oncology and it could help in revealing novel tumor-specific therapeutic vulnerabilities [5]. Many research efforts have been spent on accurately characterizing all these molecular alterations. The development of novel computational algorithms to search, analyze and prioritize cancer-associated mutations have led to extraordinary advances [6–12]. These efforts have focused on highly frequent variants, disregarding instead rare and low-frequency mutations. Thus, clinical benefits have been obtained only for a restricted group of cancers, including the frequent BCR-ABL1-fusion-positive chronic myeloid leukemia and BRAF mutated melanoma [13,14], while most tumors have profited much less from precision oncology. This manifests the urgent need for the identification of clinically relevant mutations raising the question of how to identify and characterize clinically relevant rare or low-frequency driver mutations [15]. An accurate characterization of patient-specific molecular alterations could accelerate the process of drug selection, bringing significant clinical benefits. However, the study of each of these variants is time-consuming and extremely expensive, not affordable and applicable in routine clinical activities in economically developed countries and unthinkable in most of other countries.

In this article, we discuss current approaches useful to overcome these limitations that we believe could maximize the benefits of personalized medicine. We will focus on the analysis of cancer mutations at the level of protein domains (PDs) to identify molecular predictors for the stratification of patients and drug selection.

2. Gene mutations in cancer

Cancer develops following the acquisition and accumulation of gene mutations due to the exposure to chemicals or radiations, hormones,
aggregating or other environmental factors. Over time, a number of mutations may accumulate in a single cell, promoting abnormal proliferation and eventually tumorigenesis [16]. In addition to somatic mutations that are acquired during life, inherited germline mutations represent one of the major risk factors in cancer [17]. Here we will focus on somatic cancer mutations. However, the domain-based approaches we discuss in this review may also be applied to germline mutations. Many public databases provide still very little information concerning these variants. A recent update of eBioPortal improved visualization of germline mutations, which anyway remain the minority of all mutations found in this database.

2.1. Driver mutations

Those mutations that confer a growth advantage to cancer cells and are able to initiate or trigger the progression of cancer are called ‘driver’ mutations, while all the others are termed ‘passenger’ variants [18]. Distinguishing between the two classes remains a significant challenge. Traditionally, driver mutations were identified based on their frequency in cancer patients. However, various examples of allosteric driver mutations or rare driver mutations have been provided and could be exploited for precision medicine [15,19].

The introduction of NGS technologies have allowed a fast, cheap and accurate detection of whole-genome pan-cancer mutations and have increased the power of discriminating between driver and passenger mutations, allowing to pinpoint the causative and/or therapeutically actionable variants. For example, MH Bailey and colleagues performed a pan-cancer and pan-software analysis and identified in silico more than 3400 missense driver mutations. More than 60% of these mutations were experimentally validated as drivers or putative clinically actionable mutations [20].

2.2. Frequent or rare mutations

The most characterized variants occur at high-frequency in cancer patients and include, among others, the hotspot mutation V600E of B-Raf in melanoma [21] or APC mutations in colon carcinoma [22]. More recently, large collections of cancer genomes and pan-cancer analyses have led to the identification of similar tumorigenic mutations in different cancers including mutations of TP53 (found in 42% of samples in a pan cancer analysis) PIK3CA, ML2/3/4, KRAS, NRAS and EGFR [23].

On the other hand, rare mutations, defined by several authors as those occurring in less than 1% of cancer patients [15,24], dominate the landscape of cancer-associated alterations. These can drive tumorigenesis and can be therapeutically targeted [15,19]. Remarkably, tumors are heterogeneous and dynamic organs in which a mutation occurs only in a sub-clone in a specific moment of tumor progression. Rare and sub-clonal mutations remained unexplored until the introduction of high sensitivity NGS technologies [25]. Although these rare mutations may be responsible for drug response, tumor relapse or metastasis, they were disregarded for their expensive translational potential. However, understanding their effects on tumor behavior and on the therapeutic response is of great importance for precision oncology. Various bioinformatics and bio-statistical approaches have been proposed for the identification, validation and the analysis of the druggability of rare variants [19,26–29].

3. Protein domains: functional regions

Proteins generally contain one or more functional regions in their sequence, which are commonly termed ‘domains’. In 1973, DB Wet-laufer defined for the first time PDs as stable units of a protein structure that could fold autonomously [30]. PDs retain their cognate biochemical function and 3D structure even when isolated from the protein context. PDs are evolutionary conserved in terms of amino acid sequence, 3D structure and function. The presence of different combinations of domains in different proteins gives rise to the diverse functional protein repertoire found in nature [31,32]. Identifying which domains are present in a given protein provides insight into the function of that protein.

Various databases have been created for the classification, analysis and identification of PDs. These allow to transfer information from characterized sequences to uncharacterized ones. The Structural Classification of Proteins (SCOP) database, created by manual inspection, classifies the PDs of known proteins in tree structures considering evolutionary relationships (family and superfamily) and those that arise from physics and chemistry of proteins (class). The peculiarity of SCOP is to exploit the evolutionary relationships among PDs which imply functional similarity better than simple physical relationships [33,34]. SCOP was later updated to SCOP2 in which the structural and evolutionary relationships are separated, allowing the classification of the homologous proteins into different folds and structural classes while keeping them in the same evolutionary family and superfamily [35].

More recently, the InterPro consortium ([http://www.ebi.ac.uk/interpro/about-consortium/]) collected 13 different databases for protein analysis for PD identification and classification including CATH, CDD, HAMAP, PANTHER, Prosite, Pfam and others [36]. Most of their algorithms are based on the Hidden Markov Model (HMM). The profile HMM is trained on a small representative set of aligned sequences that are known to belong to the family. This model is then used to search all homologous sequences against a large sequence database as UniProtKB to build those that we term ‘protein families’, consisting in groups of proteins that include in their sequence a given domain. On these bases the same protein containing more than one PD will belong to the corresponding distinct protein families. Remarkably, some databases, such as Pfam [37], classify PDs based on their linear amino acid sequences, while others, including CATH [38], exploit three-dimensional structures downloaded from the protein data bank (PDB) to catalogue PDs. In the following chapters, PDs will be named according to their Pfam classification.

4. Domain-based approaches for the identification of mutation hotspots

Over the last few years, numerous ‘gene-based’ methods for the identification of putative oncogenic variants have been developed. By analyzing one single gene at a time they take into account a variety of features, including the evolutionary story of the mutated position, the recurrence of the mutation across different cancer types or its expected effects on protein structure, stability and function [39–42]. However, gene-based approaches have the potential to mischaracterize mutations [43]. Also, different mutations in the same gene can trigger different functional consequences. L. Shen and colleagues classified 33 mutated genes that can act both as oncogenes and tumor suppressors [44]. This may be due to different mutations within the same PD or mutations in different PDs of the same protein. To overcome these limitations alternative methods, including tools to analyze cancer mutations at PD level, have been proposed (Fig. 1).

PD-based approaches search for mutations that cluster on specific amino acid positions (called ‘mutation hotspots’) within a given PD enabling the transfer of functional and clinical information from known variants to uncharacterized mutations. This defines what MG Kann called ‘domain landscape’ [45]. Domain-based methods for mutation analysis rely on the hypothesis that variants clustered on analogous residues within a PD across different proteins have a high probability to elicit similar functional consequences and be therapeutically actionable in a similar manner.

PD-based strategies aggregate mutational data from different proteins containing a given domain and generate a unique mutational profile. The general workflow of PD-based mutational approaches (summarized in Fig. 2) entails 4 steps:
A. PD alignment: the first step is the generation of a multiple sequence alignment of the PDs of all proteins of interest. This is usually achieved through alignment tools such as ClustalO.

B. Creation of a consensus sequence: a sequence including the most represented amino acid found at every position is created starting from the multiple sequence alignment. Some tools also calculate a per-position conservation score, which highlights the functionally important residues within a given PD.

C. Collection and mapping of mutations: somatic mutations of all target proteins are retrieved from public or custom databases and mapped per-position along the consensus sequence. This step allows to cluster all analogous mutations on a single position of the aligned domain.

D. Statistics: statistical models are used to calculate the statistical significance of clustered mutations considering: i) the conservation score of each position; ii) a significance threshold of background mutation frequency.

This general workflow can be integrated or complemented with tools to predict the functional impact of the identified mutations [46], resources to map the mutations on the 3D structure of a given protein [47-51], software that align the 3D structures instead of linear amino acid sequences [52], methods to visualize the position of the mutations in single proteins or the analysis of co-occurrence and mutual exclusivity of mutations [53]. Altogether this workflow defines a ‘domain-based mutation prioritization’. It enables navigating cancer mutations and prioritizing them on the bases of the putative functional impact, inferred from their position within a given domain or from the correspondence with other known oncogenic variants. Also, the aggregation of mutations from several genes increases the statistical power of rare and functional variants (Box 1).

5. PD-based mutation analysis: from genes to drug targets

In 2010, MG Kann’s team released the first bioinformatic resource for PD-based mutational analysis, called domain mapping of disease mutations (DMDM) database [54]. DMDM enables an aggregated view of all human coding disease-related mutations across homologous proteins (i.e. containing the same PD) across cancer types. Such approaches increase the power of the analyses and enable to transfer functional information from one known mutation to unknown ones, thus inferring the functional consequences of a given mutation.
Box 1

Domain-based analyses of cancer mutations enable:

1) a rational mutation prioritization. Mutation hotspots have a high probability to elicit functional consequences, thus helping to select functionally significant mutations among negligible alterations
2) to pinpoint rare variants that would remain disregarded using gene-based strategies
3) to infer the possible mechanism of action of an unknown mutation on the basis of its position within the domain
4) to find correspondence between novel unknown variants and known oncogenic or druggable mutations, thus inferring the possible functional consequences of a given mutation

mutations in colon adenocarcinoma tumor and breast invasive carcinoma, the authors extracted the most significantly mutated PDs and identified various mutation hotspots. Among those, some hotspots aggregate mutations from single genes, while others, including the ones occurring in the CENP-B_N domain, aggregate mutations from numerous genes. By doing so, they demonstrated that through PD-based approaches also low frequency variants reach the statistical significance threshold [45,55]. Interestingly, DMDM compares the cancer-related variants with variants occurring in other human diseases. This peculiar feature of DMDM increases the possibilities to transfer functional information from one known variant to unknown ones [54,56]. In the same period, the group of G. Cavet released a similar interactive web application for browsing mutation clusters among germline and somatic mutations found in cancer and other diseases [57].

In 2015, F Yang and colleagues demonstrated that in the same cancer sample a given hotspot mutation was not present in more than one protein belonging to the same family. Thus, the mutations aggregated in the same hotspot are mutually exclusive [58]. This strongly indicates that analogous mutations act in the same pathway via a parallel mechanism.

More recently, the MutationAligner web resource was developed (http://mutationaligner.org) [59,60]. This user-friendly interactive interface enables visualizing, navigating and analyzing somatic mutation hotspots identified in PDs from TCGA variant data processed by cbioPortal up to spring 2015 across genes and 22 tumor types. Despite this, MutationAligner does not allow to select subgroups of target proteins or select the type of cancer to refine the analysis. Using MutationAligner, ML Miller et al. identified 14 PDs significantly enriched for missense mutations, including the tyrosine kinase (PK_Tyr_Ser-Thr), P53, Ras and Cadherin domains [59]. This is not surprising, as all the identified domains exert biological roles in cell proliferation, differentiation and metastasization. Miller’s analyses also highlighted the correspondence between some well-known cancer driver mutations such as B-RafV600E and KITD816V/Y with uncharacterized low frequency mutations [59,61]. In a similar way, RA Toledo and colleagues showed that seven mutations of the vascular endothelial growth factor receptor 2 (VEGFR2) lay in mutation hotspots recurring in the PK_Tyr_Ser-Thr domain. Functional analysis supported the tumorigenic effects of substitutions L840F and R1032Q of VEGFR2 [62].

The bioinformatic tool LowMACA (Low frequency Mutations Analysis via Consensus Alignment) was designed for the identification of low frequency variants in pan-cancer analysis [53]. LowMACA is a versatile platform that allows to analyze Pfams, subgroups of proteins within a given Pfam or a custom protein list in all or specific cancer types. LowMACA retrieves cancer somatic mutations by querying cbioportal or other freely available databases. Also, LowMACA allows importing mutations from homemade datasets. Using LowMACA G Melloni et al. analyzed the RAS superfamily and identified new putative driver mutations in Rho, Rab and Rheb subfamilies that would have remained disregarded if single genes were considered [53]. By analyzing a subgroup of tyrosine kinase receptors belonging to the PK_Tyr_Ser-Thr protein family via LowMACA we recently found and characterized the previously unknown R1051Q and D1052N mutations of VEGFR2 as activating mutations [61,63]. Combining LowMACA with MutationAligner we revealed the correspondence between R1051Q and D1052N substitutions of VEGFR2 with the well known oncogenic mutations T599I and V600E of B-Raf [61]. On these bases, all mutations occurring at these positions in any tyrosine/Ser/Thr kinase domain-containing protein have a high probability to elicit pro-oncogenic effects.

Sporadic studies have employed PD-based strategies for the identification of novel driver cancer mutations. For example, it was shown that Pkinase, PK_Tyr_Ser-Thr, Y-Phosphatase and Src-homology 2 domains carry stage-specific mutations in invasive ductal breast cancer [64]. Also, two novel hotspot mutations in the immunoglobulin (Ig)-like domain of FGFR2 were found in colon cancer [58] by using PD-based analyses.

In recent years, to benefit from the accuracy of 3D structure-based PDs, domain-based approaches have been implemented to also include structural data from the PDB database. For example, ZR Moghadam and colleagues took advantage of CATH database to extract structure-based PDs for the identification of cancer-type specific somatic mutations recurring across PDs [52]. By using a CATH-based approach M Ashford et al. calculated the 3D proximity of mutation clusters to catalytic residues, protein-protein interfaces and ligand binding regions. Such analysis filtered out mutations that do not affect protein function (probably passengers) enabling a more accurate detection of functionally relevant mutations [65]. Although these methods are limited to the available 3D structures (only about 75% of proteins have a PDB structure) they have the great advantage to allow the visualization of the 3D spatial information about a given mutation hotspot and to infer from this information its putative functional role.

Beside discovering unknown variants and inferring their role in tumorigenesis, PD-based strategies have the potential of highlighting new therapeutically actionable cancer variants. One can speculate that, for example, all mutations analogous to the substitution V600E of B-Raf oncogene, may exhibit strong sensitivity to Vemurafenib (PLX4032) kinase inhibitor, similarly to mutated B-Raf [66,67]. This hypothesis would definitely increase the number of patients candidate for a given drug with tremendous clinical benefits. The analysis of 3D spatial information may further increase the number of putative actionable variants and target patients by pinpointing those additional mutations that occur in residues close to the ones that confer drug sensitivity [49] (Fig. 3).

6. Conclusions

The possibility to identify and attribute a functional role to mutations that may be causative of cancer is not only an important scientific question for understanding the mechanisms of tumorigenesis, but is also
essential to the development of personalized cancer therapies.

Here we described the state-of-the-art PD-based approaches for a systematic identification, classification, and investigation of cancer-associated mutations. Altogether, these strategies promise an accurate prioritization of driver mutations in cancer genomes. Despite PD-based approaches enable the transfer of functional information from characterized variants to unknown ones, further efforts are warranted to finally confirm that all mutations clustered on specific residues of PDs do have the same functional consequences. This concept is particularly important for therapeutically actionable mutations. Indeed, PD-based analyses have the potential to accelerate the choice of patient-specific targeted drugs. Also, they set the bases for the development of new cancer drugs. Altogether, these efforts will guarantee a more efficacious application of precision oncology.

**Author contributions**

E.G, C.R., M.C. drafted the manuscript; E.G. prepared figures; E.G., L. Z., S.M. edited and revised the manuscript.

**Declaration of Competing Interest**

None.

**Acknowledgments**

This work was supported by grants from AIRC to S.M. (IG17276). E. G. was supported by Fondazione Umberto Veronesi Fellowship.


M.S. Cline, R. Karchin, Using bioinformatics to predict the functional impact of SNVs, Bioinformatics 27 (2011) 441–448.


