Bioinformatics, 34, 2018, i795–i801
doi: 10.1093/bioinformatics/bty569
ECCB 2018

Analysis of single amino acid variations in singlet hot spots of protein–protein interfaces

E. Sila Ozdemir[1], Attila Gursoy[2,3,*] and Ozlem Keskin[1,3,*]

[1]Department of Chemical and Biological Engineering and [2]Department of Computer Engineering and [3]Research Center for Translational Medicine (KUTTAM), Koc University, Istanbul, 34450, Turkey

[*]To whom correspondence should be addressed.

Abstract

Motivation: Single amino acid variations (SAVs) in protein–protein interaction (PPI) sites play critical roles in diseases. PPI sites (interfaces) have a small subset of residues called hot spots that contribute significantly to the binding energy, and they may form clusters called hot regions. Singlet hot spots are the single amino acid hot spots outside of the hot regions. The distribution of SAVs on the interface residues may be related to their disease association.

Results: We performed statistical and structural analyses of SAVs with literature curated experimental thermodynamics data, and demonstrated that SAVs which destabilize PPIs are more likely to be found in singlet hot spots rather than hot regions and energetically less important interface residues. In contrast, non-hot spot residues are significantly enriched in neutral SAVs, which do not affect PPI stability. Surprisingly, we observed that singlet hot spots tend to be enriched in disease-causing SAVs, while benign SAVs significantly occur in non-hot spot residues. Our work demonstrates that SAVs in singlet hot spot residues have significant effect on protein stability and function.

Availability and implementation: The dataset used in this paper is available as Supplementary Material. The data can be found at http://prism.ccbb.ku.edu.tr/data/sav/ as well.

Contact: agursoy@ku.edu.tr or okeskin@ku.edu.tr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent developments in sequencing created a vast amount of information about single amino acid variations (SAVs) in human genome (Gonzalez-Perez *et al.*, 2013; Stranger *et al.*, 2011). Missense variations resulting in alterations in amino acid sequences are one of the main reasons of Mendelian diseases or complex genetic disorders (Collins *et al.*, 1998; Risch and Merikangas, 1996). Distinguishing disease-causing SAVs from benign SAVs has become a very hot topic for diagnostic and therapeutic purposes. However, associating distribution of SAVs with their effects still needs significant effort.

Proteins are responsible for diverse functions in all cellular machineries. Most, if not all, biological processes such as proliferation, cell signaling and apoptosis are mediated by protein–protein interactions (PPIs) (Alexov, 2008; Jones and Thornton, 1996; Keskin *et al.*, 2008, Keskin, *et al.*, 2016). SAVs that disrupt PPIs most likely affect many protein functions, which may lead to diseases (Gao *et al.*, 2015; Kar *et al.*, 2009; Stefl *et al.*, 2013; Yates and Sternberg, 2013). Recent studies on mechanistic effects of SAVs suggest that disease-causing variations prefer to be located on protein–protein interaction sites (interfaces) compared to other regions in the surface (Butler *et al.*, 2015; David *et al.*, 2012; David and Sternberg, 2015; Jubb *et al.*, 2016; Sahni *et al.*, 2015; Schuster-Böckler and Bateman, 2008). Also, it has been observed that significant physicochemical changes occur in proteins upon SAVs in interfaces compared to variations in non-interface residues, and a significant number of SAVs located in interfaces are predicted as disease-causing (Nishi *et al.*, 2013).

SAVs in interface residues can interfere with PPIs by affecting their stability and/or their binding properties (Alexov and Sternberg, 2013; Yates and Sternberg, 2013). The effect of a SAV on protein binding can be measured with binding free energy changes ($\Delta\Delta G$). Recent studies show that SAVs, especially disease-causing SAVs, destabilize proteins and their interactions by effecting the electrostatic component of binding energy (Nishi *et al.*, 2013; Petukh *et al.*, 2015). However, disease-causing SAVs may also increase the stability of interaction of protein complexes, therefore the effect of stabilizing SAVs cannot be ignored (Kucukkal *et al.*, 2015; Nishi *et al.*, 2013). Indeed, SAVs on suppressors and activators can have opposing phenotypes (Engin *et al.*, 2016; Yates and Sternberg, 2013). The changes in binding free energy upon SAVs serve as a good indicator for disease association (Peng and Alexov, 2016).

The contribution of each amino acid in interfaces to the formation and stability of the interaction is not the same. Using

experimental methods, it has been shown that even a single residue can provide a large binding free energy change compared to other residues of an interface, which makes it crucial for the interaction. These critical residues that contribute more to the binding free energy than the others are called 'hot spots' (Bogan and Thorn, 1998; Clackson and Wells, 1995). Hot spots are identified *in vitro* and *in silico* as those residues that cause at least 2.0 kcal/mol increase in the binding free energy upon changing them into alanine (Bogan and Thorn, 1998; Tuncbag *et al.*, 2009). Hot spots are not randomly distributed at interfaces and but rather tightly packed, they are clustered to form hot regions (Cukuroglu *et al.*, 2012; Keskin *et al.*, 2005a,b). Hot spots not involved in any hot regions are called singlet hot spots. David and Sternberg showed that disease-causing and destabilizing SAVs have a tendency to occur in hot spot residues (David and Sternberg, 2015).

Despite the studies showing the significance of SAVs occurring in energetically important interface residues, the correlation between the location of SAVs in interfaces and their disease-association is not fully revealed, yet. In this study, we provide statistical analyses and case studies of the correlation between SAV locations and their disease-association as well as their effects on protein binding energy by using a large dataset (SKEMPI 2.0). As opposed to similar studies using computationally predicted free energies, this is the first study using experimental free energies for associating effect of SAVs with their distribution within singlet hot spots and hot regions. Our aim is to understand how SAVs are distributed within interface residues and how this distribution can be associated with their effect on protein function and binding. We observed that destabilizing SAVs have a statistically significant tendency to occur in hot spots, specifically in singlet hot spots. Interestingly, analysis of disease-causing and benign SAVs showed that disease-causing SAVs are more likely to occur in singlet hot spot residues.

2 Materials and methods

2.1 Mapping SAVs to PPI interface residues

The SAV data used in this study is taken from SKEMPI 2.0 (https://life.bsc.es/pid/skempi2/) (Jankauskaite *et al.*, 2018; Moal and Fernández-Recio, 2012). This dataset includes experimentally measured affinities of wild-type complexes and affinities upon SAVs collected from scientific literature. In our analysis, insertion and deletion variations and entries with multiple variations are excluded. Around one-third of SAVs is obtained from alanine scanning experiments (Moal and Fernández-Recio, 2012).

PPI interfaces are extracted from our previous dataset, Piface (Cukuroglu *et al.*, 2014). Piface is a structurally non-redundant protein–protein interface dataset. This dataset was generated by structurally comparing and clustering 130 209 protein–protein interfaces extracted from the protein data bank, PDB (Berman *et al.*, 2000). This interface set has been also used in template-based modeling (Muratcioglu *et al.*, 2015, Baspinar, *et al.*, 2014).

A PPI interface is described as the contact region between two interacting proteins. The distance between any two heavy atoms of the two residues from two different chains should be less than the sum of their Van der Waals radii plus 0.5 Å in order to define them as contacting residues, nearby residues are any residues within the 6 Å distance from alpha carbon atoms of contacting residues on the same chain. The interface scaffold is built by combining both contacting and nearby residues (Cukuroglu *et al.*, 2014; Keskin *et al.*, 2005b; Li *et al.*, 2004). Other residues which are not core and interface residues are classified as surface residues. Hot spot residues having smaller distances than 6 Å between their Cα atoms are considered as in the same hot region, at least three hot spot residues are required to form a hot region (Keskin *et al.*, 2005a,b).

In this study, SAVs from SKEMPI 2.0 are mapped to contacting residues of Piface interfaces and core and surface residues of the proteins which have interfaces in Piface. However, our main focus is SAVs in contacting residues. Since some of the interfaces have high sequence similarity, the redundancy of interfaces and SAVs in contacting residues are checked after mapping. It is seen that more than 10% of the SAVs are coming from the redundant sequences. We did not exclude these SAVs, since they may be involved in different interactions.

Our mapping procedure consists of four steps: i) First PDB IDs obtained from SKEMPI 2.0 dataset and PDB IDs of interfaces from Piface are compared and matching PDB IDs are listed. ii) Then, SAV residues are matched with contacting residues of PPI interfaces. iii) Using HotRegion (Cukuroglu *et al.*, 2012), KFC2a (Zhu and Mitchell, 2011) and Robetta (Kim *et al.*, 2004) web servers, computational hot spots are found on contacting residues. Majority voting is performed using the hot spot predictions of these three servers to determine hot spots. For some interface residues, Robetta results could not be obtained, therefore KFC2b results are used in majority voting in such cases. Then, hot regions are determined using the approach explained above. iv) Finally, hot spot residues and hot regions are matched with SAV residues to identify singlet hot spot SAVs, hot region SAVs and non-hot spot SAVs.

2.2 Determining changes in binding free energy upon SAVs

In order to calculate the binding free energy changes in proteins upon SAVs, the experimental affinity data (K_d) from SKEMPI 2.0 is used. Binding free energy is calculated for each SAV (ΔG_{SAV}) and wild type (ΔG_{WT}) residue (Eq. 1) where R is the universal gas constant. Then, $\Delta\Delta G$ values are calculated with these ΔG values for experiments conducted at different temperatures (*Eq.* 2). For most of the cases, temperature is equal to 298 K. However, for some cases, experimental temperatures vary between 273 and 323 K. Temperature (T) in the following relations is changed according to different experimental temperatures.

$$\Delta G = \ln(K_d) \tag{1}$$

$$\Delta\Delta G = \Delta G_{SAV} - \Delta G_{WT} \tag{2}$$

where $R = 8.314/4184$ kcal/K.mol and $T = 273.15 + 25$ K. There are some entries with several experimental data for the same SAV in SKEMPI 2.0, for these variations the average of the $\Delta\Delta G$ values are taken. If these several $\Delta\Delta G$ values for the same SAV differ more than 1.5 kcal/mol, these outlier cases are removed before calculating the averages. Therefore, 2 cases are removed in our analysis. Removed cases are variation of aspartic acid to alanine in the 39th residue of D chain of 1BRS and variation of threonine to glutamic acid in the 17th residue of I chain of 1CHO.

A threshold is applied to categorize the effect of SAVs as destabilizing, stabilizing and neutral. SAVs with $\Delta\Delta G$ values larger than 0.5 kcal/mol are considered as destabilizing while $\Delta\Delta G$ values smaller than −0.5 kcal/mol indicate stabilizing SAVs. Neutral SAVs have $\Delta\Delta G$ between −0.5 and 0.5 kcal/mol (Zhao *et al.*, 2014).

The data used in our analysis including PDB ID, chain, residue number, original amino acid, variation, type of the residue [hot spot (H) or non-hot spot (NH)], hot region information and effect

of SAVs on binding free energy information are provided in Supplementary Table S1.

2.3 Determining functional effects of variations on proteins

PolyPhen2 web server is used to evaluate the disease-association of SAVs in contacting residues (Adzhubei *et al.*, 2013; Adzhubei, 2010). PolyPhen2 evaluates a variation as probably damaging (more confident prediction), possibly damaging (less confident prediction) or benign (harmless) based on HumDiv model. HumDiv is compiled from all damaging alleles with known effects on the molecular function causing human Mendelian diseases from the UniProtKB database. In our analysis, protein structures which belong to other organisms rather than human are excluded and both probably and possibly damaging SAVs are considered as disease-causing.

Additionally, HGMD (Stenson *et al.*, 2017), dSysMap (Mosca *et al.*, 2015), PinSNP (Lu *et al.*, 2016), COSMIC (Forbes *et al.*, 2015) and human disease mutation data from Sahni *et al.* (2015) datasets are used to match human SKEMPI 2.0 SAVs with known disease variations to strength the analyses utilizing PolyPhen2 predictions. In SKEMPI 2.0, PDB locations of SAVs are provided, while in these datasets listed above the UniProt location of SAVs are given. Therefore, after converting PDB residues to UniProt residues, SKEMPI 2.0 SAVs are matched with variations in these datasets. In order to obtain the energy distribution profile of disease-causing and benign SAVs, ΔΔG values of 298 disease-causing and 191 benign SAVs. Statistical analysis of the distributions is obtained by two tailed, unpaired t-test.

2.4 Calculating χ^2 and odds ratio

χ^2 and *P*-values are calculated using the number of destabilizing, stabilizing and neutral SAVs in singlet hot spot residues, hot regions and non-hot spot residues to test the dependence of change in binding free energy on SAV location. In order to find the preferences of SAVs to be located in the different regions or residues of proteins, odds ratios are calculated according to

$$OR_{ij} = \frac{x_i / (1-x_i)}{x_j / (1-x_j)} \quad (3)$$

$$x_i = n_i / N_i \quad (4)$$

where x_i is the probability of observing a variant in the specific region *i*, N_i is the total residue number in the specific region *i* and n_i is the number of variations in the specific region *i*. Two-tailed *P*-values are calculated with Fisher's exact test to test statistical significance of the OR values using the statistical packages in R version 3.3.1 (https://www.r-project.org/).

3 Results

In this study, the distribution of SAVs effecting protein interaction stability and protein function within singlet hot spot, non-hot spot and hot region residues is analyzed. Figure 1 explains the differences between hot spot, singlet hot spot, hot region and non-hot spot residues of a protein. Hot regions are hot spot clusters which meet the conditions described in Figure 1. Singlet hot spots are hot spot residues which are not part of any hot region. Hot spots include both singlet hot spots and hot regions. Residues which are not hot spots are classified as non-hot spots.

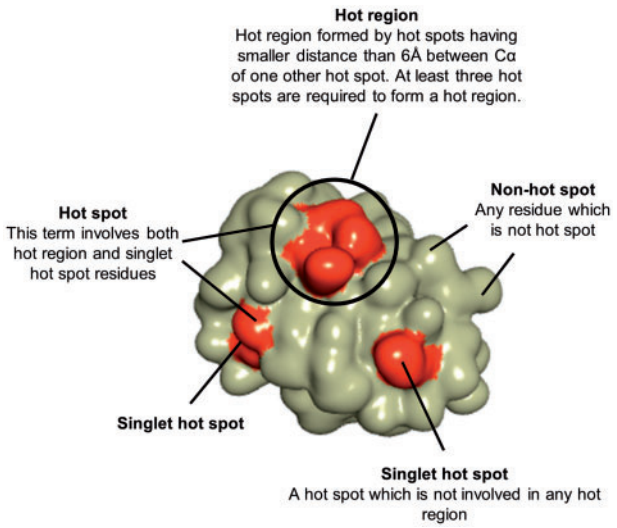

Fig. 1. Schema of hot spots, hot region, singlet hot spots and non-hot spot residues. Red residues are hot spots, however the residues inside the black circle cluster and form a hot region. Other hot spot residues which cannot form hot region are singlet hot spots. Grey residues are non-hot spot residues

Table 1. Classification of SAVs in interface residues

Classification	Description
Location	
Singlet hot spot SAVs	SAVs located in singlet hot spot residues
Hot region SAVs	SAVs located in hot regions
Non-hot spot SAVs	SAVs located in residues which are not hot spots.
Effect on binding energy	
Destabilizing SAVs	SAVs with ΔΔG values larger than 0.5 kcal/mol
Stabilizing SAVs	SAVs with ΔΔG values smaller than −0.5 kcal/mol
Neutral SAVs	SAVs having ΔΔG between −0.5 kcal/mol and 0.5 kcal/mol
Disease-association	
Benign SAVs	Human SAVs predicted as benign based on HumDiv model of PolyPhen2
Disease-causing SAVs	Human SAVs predicted as probably damaging (more confident prediction) and possibly damaging (less confident prediction) based on HumDiv model of PolyPhen2

The categories of SAVs which are used in our distribution analyses are summarized in Table 1. We classify SAVs in terms of their location, their effect on binding free energy and their disease-association. Details are provided in Section 2. Below we provide the results of our analysis.

3.1 Destabilizing SAVs are more likely to be found in singlet hot spots

Since SKEMPI 2.0 dataset consists of mutations mostly found in interface residues (Jankauskaite *et al.*, 2018; Moal and Fernández-Recio, 2012), analyses are focused on the contacting residues of interfaces. Initially 4441 SAVs (with 5079 different ΔG entries) from 320 different PDB structures are collected from the SKEMPI 2.0 database and are mapped onto Piface interfaces (Cukuroglu

Table 2. General statistics on the data and number of SAVs, hot spots and hot spot SAVs on PPI interfaces

Single SAVs from SKEMPI	4441	Contacting residues in Piface interfaces	3612[a]
PDB structures/chains[b]	172/234	SAVs in contacting residues	1643
Hot spots in interfaces	1294	Hot spot SAVs	679
Hot regions clusters/hot region residues in interfaces	151/855	Hot region SAVs	356
Singlet hot spots in interfaces	439	Singlet hot spot SAVs	323

[a]This number obtained by counting interface residues in 234 chains of 172 structures which have at least one SAV in.

[b]With SAV(s) in their interfaces.

Table 3. Odds ratios of destabilizing, stabilizing and neutral SAVs for hot region residue distribution

	Destabilizing SAVs		Stabilizing SAVs		Neutral SAVs	
Hot region versus Non-hot spot	273 542	**1.54**†	30 86	0.94	53 336	**0.39**†
Singlet hot spot versus Non-hot spot	291 542	**6.44**†	15 86	0.92	17 336	**0.24**†
Singlet hot spot versus Hot region	291 273	**4.19**†	15 30	0.97	17 53	0.61
Hot spot versus Non-hot spot	273+291 542	**2.53**†	30+15 86	0.94	53+17 336	**0.34**†

Note: The number of SAVs in corresponding groups are given in the second, fourth and sixth columns. The corresponding odd ratios are in the following columns. All calculations are based on the total number of residues in each group. Hot region residues = 855, Non-hot spot residues = 2318, Singlet hot spot =439, Hot-spot = 1294.

†Statistically significant *P*-value (<0.05) was obtained for OR values.

et al., 2014). We observe that 1643 out of 4441 SAVs correspond to contacting residues of 172 PDB structures consisting of 234 different chains. The number of SAVs in these 234 chains varies between 1 and 114 with a median of 5. The maximum number of 114 SAVs comes from one chain (PDB ID: 3SGB, chain I) whose several residues were mutated to all possible other amino acids. In these chains, there are 679 hot spot SAVs and 356 of them are in hot region (Table 2). Supplementary Table S2 shows detailed information about the number of contacting residues, SAVs, hot spot residues and hot regions for each chain forming interfaces. Supplementary Table S2 indicates that our hot spots and hot regions are obtained from 161 and 99 chains, respectively, but not from a few chains.

First, destabilizing, stabilizing and neutral SAVs are mapped to hot regions, non-hot spots and singlet hot spots of protein complexes to find the number of SAVs in each of these locations. We aim to see whether there is a relationship between SAV localization and binding free energy change. Therefore, we construct a contingency table (Supplementary Table S3A) and test the dependence of the change in binding free energy on SAV distribution within hot regions, non-hot spots and singlet hot spots with chi square (χ^2) test. The result of the χ^2 test indicates that the binding free energy changes depend on the distribution of SAVs within these regions (*P*-value <2.2e−16). Therefore, in order to reveal the nature of this dependence, the correspondence of destabilizing, stabilizing and neutral SAVs to hot regions, non-hot spots and singlet hot spots is further analyzed with odds ratio (OR) calculations (Eq. 3 and Eq. 4). Remember that a hot region is formed if at least three hot spots tightly cluster spatially together. An interface might contain one or a few hot regions. It should also be kept in mind that some hot spot residues are not involved in any hot region, therefore they are considered as singlet hot spot residues. In our interfaces, there are totally 855 residues in 151 hot regions, therefore there are 5.66 residues in each hot region on average (Table 2). The maximum number of hot regions in an interface is 3, the average number of hot regions per interface is 1.52 (total number of hot regions/number of interfaces with hot regions), while some interfaces do not have any hot regions (Supplementary Table S2). In Table 3, the distribution of these SAVs within hot spots, hot regions, singlet hot spots and non-hot spots is analyzed. It is observed that destabilizing SAVs are 2.53, 1.54 and 6.44 times more likely to occur in hot spots, hot regions and singlet hot spots rather than in non-hot spot residues. Moreover, destabilizing SAVs significantly prefers to be in singlet hot spots rather than hot regions (OR = 4.19, $P < 0.05$). The preference of destabilizing SAVs to be located mostly in hot spots is expected as the definition of hot spots is based on the destabilizing energies ($\Delta\Delta G \geq 2.0$ kcal/mol). However, the preference to be in singlet hot spots is striking. We also show that, although it is not statistically significant, stabilizing SAVs are distributed with almost the same preference over non-hot spots, hot regions and singlet hot spots. An opposite trend can be seen for neutral SAVs, they are more likely to occur in non-hot spot residues compared to hot region residues (OR = 0.39, $P < 0.05$) and compared to singlet hot spots (OR = 0.24, $P < 0.05$). Our results indicate that any SAV that is destabilizing is more likely to be found in hot spots, especially in singlet hot spots, while a neutral SAV is probably located in non-hot spots. Overall, the distribution profile of SAVs having different effects on the binding free energies shows that hot regions, singlet hot spots and non-hot spots are enriched in different SAV profiles. Therefore, we continue to analyze the distribution profile of SAVs within these residues to differentiate them.

3.2 Singlet hot spots are enriched in disease-causing SAVs

Disease-association of SAVs is found by PolyPhen2 web server (Adzhubei, 2013). PolyPhen2 prediction is performed for only human SAVs. After excluding SAVs from non-human organisms, 489 human SAVs out of 1643 SKEMPI 2.0 SAVs are left. PolyPhen2 data belonging to these human SAVs as well as hot spot, hot region information and effect on binding free energy can be seen in Supplementary Table S4. The numbers of residues in hot region, singlet hot spot and non-hot spot are recalculated excluding non-human PDBs. Totally, 181 hot region, 176 singlet hot spot and 766 non-hot spot residues are found and OR calculations are performed based on these numbers. First, in order to associate binding free energy change with disease-association of SAVs, binding free energy change distribution of disease-causing and benign SAVs is analyzed. Figure 2 shows distributions of ΔΔGs of disease-causing and benign SAVs. x-axis shows ΔΔG values, while y-axis presents frequencies of these observations for disease-causing (blue bars) and benign (orange bars) SAVs. Distribution of ΔΔG values caused by disease-causing SAVs is broader and shifts towards more positive values indicating their destabilizing effect. On the other hand, ΔΔG values of benign SAVs show a skewer distribution indicating a more neutral effect. The mean and standard deviation of ΔΔGs caused by disease-causing SAVs are 1.42 kcal/mol and 1.35 kcal/mol, while the mean and standard deviation of ΔΔGs caused by benign SAVs

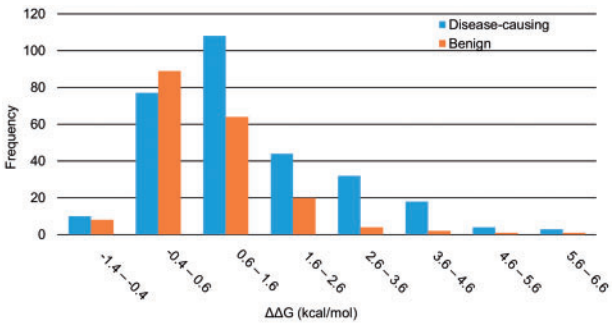

Fig. 2. The distribution of binding free energy change in disease-causing and benign SAVs. ΔΔG values of disease-causing SAVs (blue bars) and benign SAVs (orange bars) are in *x*-axis and frequencies of these values are in *y*-axis

0.84 kcal/mol and 1.16 kcal/mol, respectively. The difference between these two distributions is statistically highly significant ($P \ll 0.05$).

We, then, test and confirm the dependence of disease-association on location of SAVs (hot region, singlet hot spots, non-hot spots) with χ^2 calculation by constructing a contingency table (Supplementary Table S3B, *P*-value = 8.86e−05). Then, we further analyze the distribution of disease-causing and benign SAVs within singlet hot spot, hot region or non-hot spot residues with OR calculations. Table 4 shows how benign SAVs are distributed within hot regions, singlet hot spots and non-hot spots. As expected, benign SAVs are 2.03 times more likely (OR = 0.49, $P < 0.05$) to occur in non-hot spot residues compared to singlet hot spots. Disease-causing SAVs (both more confident and less confident PolyPhen2 predictions were considered as diseae-causing) in Table 4 prefer to occur in singlet hot spot residues compared to non-hot spot residues (OR = 2.32, $P < 0.05$). Also, singlet hot spot and hot regions comparison shows that disease-causing SAVs are 1.81 times more likely to be found in singlet hot spots (OR = 1.81, $P < 0.05$) rather than in hot regions. Considering only more confident predictions, we obtained similar results with a higher OR. Disease-causing SAVs with more confident predictions prefer to be in singlet hot spot residues compared to both non-hot spot residues and hot regions (OR = 2.61, $P < 0.05$ and OR = 1.84, $P < 0.05$, respectively). This highlights that disease-causing SAVs specifically prefer to be located in singlet hot spots, which further differentiates SAV distribution on singlet hot spots from hot regions. This result can be explained by the epistasis of hot spots in the same hot region. If a SAV exists in a hot region, other hot spots in the same hot region may compensate for its effect. However, if a SAV corresponds to a singlet hot spot, its effect might be more severe, since their effect on binding free energy is additive. The epistasis and additivity phenomena is explained in detailed in Case Studies (Supplementary Material).

Since our classification as disease-causing and benign comes from PolyPhen2 predictions, we repeat the analyses in Table 4, after matching human SKEMPI 2.0 SAVs with known disease variations from HGMD (Stenson *et al.*, 2017), dSysMap (Mosca *et al.*, 2015), COSMIC (Forbes *et al.*, 2015), PinSNP (Lu *et al.*, 2016) which covers dbSNP, OMIM and Sahni *et al.* (2015) datasets. Thirty five human SKEMPI 2.0 SAVs match with disease variations from these datasets. Supplementary Table S5 gives information about disease types, gene names and datasets of these matched SAVs. According to the analysis with these 35 SAVs, disease-causing SAVs are more likely to be located in singlet hot spots rather than hot region with a higher OR in Table 4 (4.14, *P*-value = 0.055. All calculations are based on the total number of residues in each group. Singlet hot spot residues = 23, hot region residues = 27). Although these numbers are not statistically significant due to small numbers of observations, they provide some confirmations about our results using PolyPhen2 predictions.

Table 4. Odds ratios of benign and disease-causing SAVs for hot spot residue distribution for human proteins

	Benign SAVs		Disease-causing SAVs			
			More and less confident PolyPhen2 predictions		More confident PolyPhen2 predictions	
Hot region versus	29	0.83	50	1.27	35	1.41
Non-hot spot	144		176		111	
Singlet hot spot versus	18	**0.49**†	72	**2.32**†	54	**2.61**†
Non-hot spot	144		176		111	
Singlet hot spot versus	18	0.60	72	**1.81**†	54	**1.84**†
Hot region	29		50		35	
Hot spot versus	29+18	**0.65**†	50+72	**1.74**†	35+54	**1.96**†
Non-hot spot	144		176		111	

Note: All calculations are based on the total number of residues in each group. Hot region residues = 181, Non-hot spot residues = 766, Singlet hot spot = 176, Hot-spot = 357.

†Statistically significant *P*-value (<0.05) was obtained for OR value.

4 Discussion

Our detailed analysis of SAVs obtained from a large experimental dataset provides a link between SAV distribution within the different interface residues and their effects. We show that destabilizing SAVs are more likely to be in hot spot residues, specifically in singlet hot spots, while neutral SAVs tend to be located in non-hot spot residues. Since, hot spots are the major contributors to the binding free energy, they are considered to be crucial in establishing the stability and affinity of the protein complexes (Agius *et al.*, 2013; Tuncbag *et al.*, 2009). Our results show that SAVs being located in hot spots, especially in singlet hot spots are more likely to change binding affinity of PPI compared to SAVs non-hot spot residues. This indicates the importance of hot spots (singlet hot spots and hot regions) to provide stability to the protein complex. Hot spots are likely evolutionary conserved residues. Therefore, it is expected that SAVs in these residues affect protein function and stability (Moreira *et al.*, 2007, An, *et al.*, 2013).

We also analyze the relationship between disease-association of SAVs and their distribution within non-hot spot, singlet hot spot residues and hot regions. We show that benign SAVs significantly tend to occur in non-hot spot residues compared to hot regions and singlet hot spots. Also, we observe that singlet hot spot residues are significantly enriched in disease-causing SAVs. This may be explained by additivity of singlet hot spots and epistatic effect of hot spots in hot regions on binding free energy. When a SAV occurs in one of the hot region other residues forming hot region may compensate the effect of this SAV on interaction, since they are epistatic. However, SAVs in singlet hot spots can directly affect PPIs due to their additive effect on binding free energy. Since our classification as disease-causing and benign comes from predictions, we further confirmed our findings by matching SKEMPI 2.0 SAVs with known disease variation data from various datasets. We obtain similar results to our previous results; disease-causing SAVs are more likely to be found in singlet hot spots. Although the number of the matching SAVs is small, we can still show similar distribution trends in both prediction and experimental data.

Singlet hot spots and hot regions are important in pharmacological studies. Designing drugs to target PPIs is a promising area for drug discovery and drug repurposing. Studies on this area reveal that there is a close relationship between PPI specific drugs and hot spots in interfaces. Also, computational methods confirm the relationship between hot spots and druggability (Hall *et al.*, 2015). Drugs targeting hot spots in interfaces increase their possibility to bind to their target interface and establish stable interaction. The major characteristics of druggable hot spots are about their location preferences, they have tendency to form clusters. Hot spots present a binding platform for PPI. These regions usually coevolve with hot regions of their binding partners which also give a critical insight for drug repurposing (Li *et al.*, 2004). Therefore, it is crucial to establish a link between disease-association of SAVs and their distribution within contacting residues. We showed the SAVs in singlet hot spots cannot be easily compensated and they can destabilize the interaction, therefore drugs designed to target singlet hot spots may be more effective to eliminate the interactions.

In case studies (Supplementary Material), we observe SAVs in different singlet hot spots may lead to different phenotypes. Also, we exemplify the epistasis and additivity phenomena by showing the binding free energy change caused by SAVs in different hot regions are additive and binding free energy change upon SAVs in the same hot regions is epistatic. This phenomena also help to explain and understand the distribution of disease-causing SAVs in singlet hot spots.

Parallel to our results described above, similar studies previously show hot spots are enriched in destabilizing and disease-causing SAVs, however they do not reach statistical significance because of the small number of observations (David and Sternberg, 2015). Therefore, it is the first time to our knowledge that experimentally obtained large binding energy data is used to relate the effect of SAVs to their distribution within singlet hot spot residues.

To conclude, revealing the relationship between disease-association of SAVs and their distribution within interface residues will provide important knowledge for targeting interfaces by therapeutic drugs. In this study, with the use of experimentally obtained binding free energy data and PPI interfaces, we performed a comprehensive analysis of SAV distribution. We mainly studied the relationship between the effects of SAVs and their distribution within different interface residues. Our analysis showed the importance of this relation in order to predict the impact of SAVs in PPIs and to evaluate their disease-association. Predicting and determining these impacts are crucial to understand nature of SAVs and their overall influence to biological processes.

Acknowledgements

E.S.O acknowledges TUBITAK (The Scientific and Technological Research Council of Turkey) for financial support (Scholarship 2211-E). We thank Dr. Iain Moal for kindly sharing SKEMPI 2.0 dataset with us.

Conflict of Interest: none declared.

References

Adzhubei,I. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, Chapter 7, Unit7 20.

Adzhubei,I.A. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Agius,R. *et al.* (2013) Characterizing changes in the rate of protein–protein dissociation upon interface mutation using hotspot energy and organization. *PLoS Comput. Biol.*, **9**, e1003216.

Alexov,E. (2008) Protein–protein interactions. *Curr. Pharm. Biotechnol.*, **9**, 55–56.

Alexov,E. and Sternberg,M. (2013) Understanding molecular effects of naturally occurring genetic differences. *J. Mol. Biol.*, **425**, 3911–3913.

An,O. *et al.* (2013) Structural and functional analysis of perforin mutations in association with clinical data of familial hemophagocytic lymphohistiocytosis type 2 (FHL2) patients. *Protein Sci.*, **22**, 823–839.

Baspinar,A. *et al.* (2014) PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. *Nucleic Acids Res.*, **42**, W285–W289.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **280**, 1–9.

Butler,B.M. *et al.* (2015) Conformational dynamics of nonsynonymous variants at protein interfaces reveals disease association. *Proteins*, **83**, 428–435.

Clackson,T. and Wells,J.A. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science*, **267**, 383–386.

Collins,F.S. *et al.* (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.*, **8**, 1229–1231.

Cukuroglu,E. *et al.* (2012) HotRegion: a database of predicted hot spot clusters. *Nucleic Acids Res.*, **40**, D829–D833.

Cukuroglu,E. *et al.* (2014) Non-redundant unique interface structures as templates for modeling protein interactions. *PloS One*, **9**, e86738.

David,A. *et al.* (2012) Protein–protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum. Mutat.*, **33**, 359–363.

David,A. and Sternberg,M.J. (2015) The contribution of missense mutations in core and rim residues of protein–protein interfaces to human disease. *J. Mol. Biol.*, **427**, 2886–2898.

Engin,H.B. *et al.* (2016) Structure-based analysis reveals cancer missense mutations target protein interaction interfaces. *PloS One*, **11**, e0152929.

Forbes,S.A. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.

Gao,M. *et al.* (2015) Insights into disease-associated mutations in the human proteome through protein structural analysis. *Structure*, **23**, 1362–1369.

Gonzalez-Perez,A. *et al.* (2013) Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods*, **10**, 723–729.

Hall,D.R. *et al.* (2015) Lessons from hot spot analysis for fragment-based drug discovery. *Trends Pharmacol. Sci.*, **36**, 724–736.

Jankauskaite,J. *et al.* (2018) SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, doi: 10.1093/bioinformatics/bty635.

Jones,S. and Thornton,J.M. (1996) Principles of protein–protein interactions. *P. Natl. Acad. Sci. USA*, **93**, 13–20.

Jubb,H.C. *et al.* (2016) Mutations at protein–protein interfaces: small changes over big surfaces have large impacts on human health. *Prog. Biophys. Mol. Biol*, **128**, 3–13.

Kar,G. *et al.* (2009) Human cancer protein–protein interaction network: a structural perspective. *PLoS Comput. Biol.*, **5**, e1000601.

Keskin,O. *et al.* (2008) Principles of protein–protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.*, **108**, 1225–1244.

Keskin,O. *et al.* (2005a) Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.*, **345**, 1281–1294.

Keskin,O. *et al.* (2005b) Protein–protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach. *Phys. Biol.*, **2**, S24–S35.

Keskin,O. *et al.* (2016) Predicting protein-protein interactions from the molecular to the proteome level. *Chem. Rev.*, **116**, 4884–4909.

Kim,D.E. *et al.* (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.*, **32**, W526–W531.

Kucukkal,T.G. *et al.* (2015) Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. *Curr. Opin. Struct. Biol.*, **32**, 18–24.

Li,X. *et al.* (2004) Protein–protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. *J. Mol. Biol.*, **344**, 781–795.

Lu,H.C. *et al.* (2016) PinSnps: structural and functional analysis of SNPs in the context of protein interaction networks. *Bioinformatics*, **32**, 2534–2536.

Moal,I.H. and Fernández-Recio,J. (2012) SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics*, **28**, 2600–2607.

Moreira,I.S. *et al*. (2007) Hot spots–a review of the protein–protein interface determinant amino-acid residues. *Proteins*, **68**, 803–812.

Mosca,R. *et al*. (2015) dSysMap: exploring the edgetic role of disease mutations. *Nat. Methods*, **12**, 167–168.

Muratcioglu,S. *et al*. (2015) Advances in template-based protein docking by utilizing interfaces towards completing structural interactome. *Curr. Opin. Struct. Biol.*, **35**, 87–92.

Nishi,H. *et al*. (2013) Cancer missense mutations alter binding properties of proteins and their interaction networks. *PloS One*, **8**, e66273.

Peng,Y. and Alexov,E. (2016) Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding. *Proteins*, **84**, 232–239.

Petukh,M. *et al*. (2015) On human disease-causing amino acid variants: statistical study of sequence and structural patterns. *Hum. Mutat.*, **36**, 524–534.

Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Sahni,N. *et al*. (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, **161**, 647–660.

Schuster-Böckler,B. and Bateman,A. (2008) Protein interactions in human genetic diseases. *Genome Biol.*, **9**, R9.

Stefl,S. *et al*. (2013) Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.*, **425**, 3919–3936.

Stenson,P.D. *et al*. (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.*, **136**, 665–677.

Stranger,B.E. *et al*. (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, **187**, 367–383.

Tuncbag,N. *et al*. (2009) Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, **25**, 1513–1520.

Yates,C.M. and Sternberg,M.J.E. (2013) The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein protein interactions. *J. Mol. Biol.*, **425**, 3949–3963.

Zhao,N. *et al*. (2014) Determining effects of non-synonymous SNPs on protein–protein interactions using supervised and semi-supervised learning. *PLoS Comput. Biol.*, **10**, e1003592.

Zhu,X. and Mitchell,J.C. (2011) KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins*, **79**, 2671–2683.