Cancer-related single amino acid variation prediction

Jia-Jun Liu[1], Chin-Sheng Yu[2], Hsiao-Wei Wu[1], Yu-Jen Chang[1], Chih-Peng Lin[3], and Chih-Hao Lu[1]

[1]China Medical University
[2]Feng Chia University
[3]Yourgene Health

July 20, 2020

Abstract

Single amino acid variation (SAV) is an amino acid substitution of the protein sequence and might influence the whole protein structure, binding affinity, or functional domain and related to disease, even cancer. However, to clarify the relationship between SAV and cancer using traditional experiments is time and resource consuming. Though there are some SAVs predicted methods using the computational approach, most of them predict the protein stability changed caused by SAV. In this work, all of the SAV characteristics generated from protein sequences, structures, and micro-environment would be converted into feature vectors and fed into an integrated predicting system by using Support Vector Machine and genetic algorithm. The critical features were used to estimate the relationship between their properties and cancer caused by SAVs. In the results, we have developed a prediction system based on protein sequence and structure, which could distinguish the SAV is related to cancer or not, and the accuracy, the Matthews correlation coefficient, and the F1-score yield to 90.88%, 0.77 and 0.83, respectively. Moreover, an online prediction server called CanSavPre was built (http://bioinfo.cmu.edu.tw/CanSavPre/), which will be a useful, practical tool for cancer research and precision medicine.

Cancer-related single amino acid variation prediction

Jia-Jun Liu[1] | Chin-Sheng Yu[2,3] | Hsiao-Wei Wu[4]| Yu-Jen Chang[1] | Chih-Peng Lin[5] | Chih-Hao Lu[1,4,6]

[1]The Ph.D. Program of Biotechnology and Biomedical industry, China Medical University, Taichung, Taiwan

[2]Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan

[3]Master's Program in Biomedical Informatics and Biomedical Engineering, Feng Chia University, Taichung, Taiwan

[4]Graduate Institute of Biomedical Sciences, China Medical University, Taichung, Taiwan

[5]Yourgene Health, New Taipei City, Taiwan

[6]Department of Medical Laboratory Science and Biotechnology, China Medical University, Taichung, Taiwan

Correspondence

Chih-Hao Lu, Graduate Institute of Biomedical Sciences, China Medical University, No.91, Hsueh-Shih Road, Taichung, 40402, Taiwan

Email: *chlu@mail.cmu.edu.tw*

Funding information

Ministry of Science Technology, Taiwan, Grant Number: MOST 108-2221-E-039-013- China Medical University, Taiwan, Grant Number: CMU-108-MF-121

Abstract

Single amino acid variation (SAV) is an amino acid substitution of the protein sequence and might influence the whole protein structure, binding affinity, or functional domain and related to disease, even cancer. However, to clarify the relationship between SAV and cancer using traditional experiments is time and resource consuming. Though there are some SAVs predicted methods using the computational approach, most of them predict the protein stability changed caused by SAV. In this work, all of the SAV characteristics generated from protein sequences, structures, and micro-environment would be converted into feature vectors and fed into an integrated predicting system by using Support Vector Machine and genetic algorithm. The critical features were used to estimate the relationship between their properties and cancer caused by SAVs. In the results, we have developed a prediction system based on protein sequence and structure, which could distinguish the SAV is related to cancer or not, and the accuracy, the Matthews correlation coefficient, and the F1-score yield to 90.88%, 0.77 and 0.83, respectively. Moreover, an online prediction server called CanSavPre was built (*http://bioinfo.cmu.edu.tw/CanSavPre/*), which will be a useful, practical tool for cancer research and precision medicine.

KEYWORDS

single amino acid variation, human cancer, support vector machine, genetic algorithm

1 | INTRODUCTION

Single amino acid variation (SAV) is one amino acid substitution resulting from genetic polymorphisms. The non-synonymous encoding variant would alter the protein sequence. However, in some extreme cases, this slight difference might affect the whole protein structure or function. Due to the unique physicochemical properties of each amino acid, the mutation in different positions of the sequence causes various effects for the whole protein conformation and its function. It is vital to understand how the single amino acid variation could influence protein and clarify the links between genetic variation and human disease. In previous studies, most disease-related SAVs occur in the structurally or functionally essential positions (Juritz et al., 2012; Sunyaev, Ramensky, & Bork, 2000; Yue, Li, & Moult, 2005). Just like some cases, as we have known, some conserved residues mutated, it could directly damage the native protein folding. These mutation residues might affect protein structure or the complex aggregation. Protein destabilization is a primary factor in many Mendelian diseases (Guo et al., 2011; Redler, Das, Diaz, & Dokholyan, 2016; Teng, Srivastava, Schwartz, Alexov, & Wang, 2010).

Further, structural dynamics are correlated to protein function because the missense-folding structure may result in protein dysfunction (Bromberg & Rost, 2009; Ponzoni & Bahar, 2018). If missense variants occur at the functional sites, it will change protein activity and binding affinity, resulting in diseases. Moreover, SAVs occurring at interfaces are also related to diseases, since it might ruin the network of protein-protein interaction (David, Razali, Wass, & Sternberg, 2012; Yates & Sternberg, 2013). At present, there are many large-scale studies; however, most of them focus on human genetic diseases (X. Wang et al., 2012).

Additionally, there are more and more studies indicating that SAVs are also associated with several cancers (Lori et al., 2013; Niroula & Vihinen, 2015; Song et al., 2014). Cancer, which is caused by a particular change to chromosome, is often regarded as a genetic disease. However, the mechanism is distinct from Mendelian diseases, and little do we know this complicated network. Until now, a large number of studies reveal massive radical changes in cancer patient genomic sequences. Investigated mutation spots are often biomarkers or targets for treatment (Ma et al., 2018; Nie et al., 2014; Renaud et al., 2016). Previous studies reported that

a set of missense variations that disrupt protein function was associated with cancer (B. Li et al., 2009). Recent studies suggested that the accumulation of somatic mutation is a vital factor in carcinogenic progress. Some variations seem neutral, but they might contribute to cancer progressing, known as driver mutation (McFarland et al., 2017). However, identifying the trigger point preciously is still not easy, but observing the accumulation of possibly threatening is very helpful. In the proteome level, amino acid substitution caused by genetic codon transition might be the reason for human cancer (Son, Kang, Kim, & Kim, 2017). The chemical properties of replaced amino acid could lose or gain protein function. Besides, amino acid alteration seems to follow special rules. For example, arginine has a positive charge that is important to balance the charges of protein and DNA binding; however, it is highly mutated in various cancer types. The loss of function of cancer-associated proteins is frequently due to the loss of arginine.

On the other hand, if a protein gains cysteine, an active and reducing agent, this might enhance its capability to neutralize ROS in tumor environment (Anoosha, Sakthivel, & Michael Gromiha, 2016; Halasi et al., 2013; Tsuber, Kadamov, Brautigam, Berglund, & Helleday, 2017). Proteomic changes by the protein carrying missense mutation may help the cancer cells adapt to environmental pressure (Szpiech et al., 2017). Even though different types of cancers have their properties, they might share some substitution patterns. Between breast and digestive tract cancers, the amino acid substitution spectrum is similar, dominated by glutamic acid altered to lysine (Tan, Bao, & Zhou, 2015).

Nowadays, machine learning is a favorite tool for data analysis and has solved some conundrums. Thus, many predictors utilized machine learning as algorithms for SAVs had been developed. Initially, most tools got on with the protein stability and functional changes caused by missense variations (Radusky et al., 2018; Schaefer & Rost, 2012; Sim et al., 2012; Vaser, Adusumalli, Leng, Sikic, & Ng, 2016). Further, some predictors are to distinguish benign and pathogenic variations (Sundaram et al., 2018; Yates, Filippis, Kelley, & Sternberg, 2014). Hitherto, some web software demonstrated the connection between SAVs and diseases; however, those are biased towards genetic diseases (I. Adzhubei, Jordan, & Sunyaev, 2013; I. A. Adzhubei et al., 2010; Ferrer-Costa et al., 2005; Lopez-Ferrando, Gazzo, de la Cruz, Orozco, & Gelpi, 2017; Reeb, Hecht, Mahlich, Bromberg, & Rost, 2016). Few predictors are established for cancer, but most of them focus on the specific purpose or particular cancer (B. Wang et al., 2018). Some tools are designed for classifying driver and passenger mutation (Carter et al., 2009; Kaminker, Zhang, Watanabe, & Zhang, 2007; Shihab, Gough, Cooper, Day, & Gaunt, 2013). Though they are useful, a comprehensive predictor in cancer biology research is in pressing demand. In this work, we developed a prediction model that recognizes whether the SAV is cancer-related or neutral. Though numerous predictors have been developed, the critical question is how to build the prediction models and what descriptors of SAV are used (Care, Needham, Bulpitt, & Westhead, 2007). Not only to figure out for each SAV change physically in protein function and structure but also to estimate how it simultaneously contributes to cancer progression. To take into account every kind of SAVs might be a vital feature for cancer, we perform an integrated system to discriminate the cancer-related residues in sequence from multiple predicting models utilizing spread information extracted from the fundamental of protein in this work. We would provide a novel way for cancer research, not only for the clinical outcome but also for prognostic biomarker, and a breakthrough for precision medicine. Besides setting up this predictor for cancer-related variations, it would be helpful to figure out the relationship between SAV and cancer and the underlying mechanisms.

2 | MATERIALS AND METHODS

2.1 | Dataset of SAVs

All of the SAV data were collected from CanProVar 2.0 (J. Li et al., 2011; Zhang et al., 2017), a human Cancer Proteome Variation Database. Single amino acid alterations, including both germline and somatic variations in the human proteome, are stored, notably including those related to the genesis or development of human cancer based on the published literature. Until now, there are 156,671 cancer-related SAVs and 967,017 neutral SAVs in the CanProVar 2.0. In order to find out the exact protein structure of SAV sequence,

protein BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) was used via searching Protein Data Bank proteins. There were five criteria in searching as following: 1. The e-value of alignment results should be smaller than 1e-50; 2. The alignment coverage of query protein should be higher than 95%; 3. The organism of the aligned target protein should be homo sapiens; 4. The experimental method of aligned target protein structures should be X-ray Diffraction; 5. The SAV position should be identically aligned between the wild type of SAV sequence and the aligned target protein. Then, CD-HIT Suite (Huang, Niu, Gao, Fu, & Li, 2010) was used to filter out the homologous proteins by the sequence identity cut-off 0.3. After that, 2,894 cancer-related SAVs and 7,668 neutral SAVs were remained and separated into twenty groups by the representative wild amino acid type of SAV. For each wild amino acid type, the number of cancer-related and neutral SAVs were listed in Table 1, and δ, the ratio of cancer to neutral was from 22.49% to 65.89%.

2.2 | Prediction systems

In this work, two cancer-related SAV prediction systems were built by the machine learning method. The first system, $\text{CanSavPre}_{\text{w}}$, contained twenty individual prediction models constructed from twenty groups according to the wild amino acid type of SAV. In the second prediction system, $\text{CanSavPre}_{\text{wm}}$, every twenty groups were divided into smaller sub-groups by its mutated amino acid type of SAV. For example, an alanine should have a different prediction model with an acidic (e.g., aspartate or glutamate) and a basic (e.g., arginine, lysine, or histidine) mutated amino acid type due to their essential factors of SAV should be distinct. Finally, 100 prediction models were built in the second prediction system.

Each prediction model was a two-level Support Vector Machine (SVM) (Chang & Lin, 2011) classifier modules. The first level SVM comprised twelve SVM classifiers based on the three specific feature sets, as sequence-based, structure-based, and micro-environment-based feature sets, which described in the next section, respectively. For each feature set, four fitness functions (Equations 1-4) were used for feature selection and performance optimization using the genetic algorithm (Lu, Chen, Yu, & Hwang, 2007; Yu & Lu, 2011).

Four informative measures for predictive performance were used as the fitness functions, which were accuracy (Acc), Matthews correlation coefficient (MCC), F1 score (F1) and summation of sensitivity and weighted specificity (Hybrid) and were calculated by true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values as follows:

$Acc = \frac{TP+TN}{TP+TN+FP+FN}$, (1)$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, (2)$F1 = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity}$, (3)$Hybrid = Sensitivity + \delta \times Specificity$, (4) where $Precision = \frac{\text{TP}}{TP+FP}$,$Sensitivity = \frac{\text{TP}}{TP+FN}$,$Specificity = \frac{\text{TN}}{TN+FP}$, TP is the true positives, TN is the true negatives, FP is the false positives, FN is the false negatives and δ is the ratio of the number of cancer-related to neutral SAV which listed in Table 1. All of the descriptors of SAV were fed into SVM, and the five-fold cross-validation was performed when the model training and testing.

The second level of SVM classifiers was used to process the prediction results generated from twelve classifiers (three feature sets was multiplied by four fitness functions) in the first level to produce the final probability distribution of the relationship with cancer-related or neutral. The relationship with the largest probability was used as the final prediction. The two-level SVM system is shown schematically in Figure 1.

FIGURE 1 The two-level SVM prediction system.

2.3 | Feature vectors sets

The descriptors of SAVs used for machine learning were classified into three classes, the sequence-based, structure-based, and micro-environment-based features sets. For the sequence-based feature set, 44 descriptors were extracted from the protein sequence and partitioned into three groups listed in Table 2. The first group was from the most generally used substitution index of wild type residue to mutation for the SAV residue. Three kinds of substitution index were used included the BLOSUM62 (Choi, Sims, Murphy, Miller, & Chan, 2012; Henikoff & Henikoff, 1992), PAM250 (D. T. Jones, Taylor, & Thornton, 1992), and

position-specific scoring matrix (PSSM), which derived from PSI-BLAST (Altschul et al., 1997). The second group represented the conservation for each residue comparing to homologs. The fifteen evolutional entropy values derived from PSI-BLAST were used to denote a sliding window of length 15 centered on the SAV. Then the average entropy values for the window of length 15 and 5 centered on the SAV were also calculated. The third group was the amino acid compositions (AAC) (Chou, 2001) of fifteen residues peptide used to represent the composition of the neighbor residues for centered SAV. According to the physicochemical properties of residues, we used the following classification schemes (Yu, Chen, Lu, & Hwang, 2006) of amino acid compositions: H for polar (RKEDQN), neutral (GASTPHY), and hydrophobic (CVLIMFW); V for small (GASCTPD), medium (NVEQIL), and large (MHKFRYW); Z for low polarizability (GASDT), medium (CPNVEQIL), and high (KMHFRYW); P for low polarity (LIFWCMVY), neutral (PATGS), and high polarity (HQRKNED); F for acidic (DE), basic (HKR), polar (CGNQSTY), and nonpolar (AFILMPVW); E for acidic (DE), basic (HKR), aromatic (FWY), amide (NQ), small hydroxyl (ST), sulfur-containing (CM), aliphatic 1 (AGP), and aliphatic 2 (ILV). For clarity, these sequence-based descriptors were summarized in Table S1.

In the structure-based feature sets, there were thirteen descriptors extracted from PDB and DSSP (Cheng, Randall, Sweredoski, & Baldi, 2005; Kabsch & Sander, 1983). The b-factor value of Cα atom of SAVs was used as the first structure-based descriptor, which was the displacement of atoms from their mean position in a crystal structure diminishes the scattered X-ray intensity. The displacement may be the result of temperature-dependent atomic vibrations or static disorder in a crystal lattice. Additionally, the critical information of the related solvent accessibility, eight DSSP defined secondary structures element (e.g., H, B, E, G, I, T, S, and others), the energy of backbone hydrogen bonds for acceptor and donor, and disulfide bonding or not gathered from DSSP were also used. These structure-based descriptors were summarized in Table S2.

In the third feature set, the weighted contact number (WCN) model (Lin et al., 2008) was used to describe the micro-environment properties of SAVs. The weighted contact number model was a local packing density profile, and it was reported that the WCN profile has a high correlation with the sequence conservation profile (Shih, Chang, Lin, Lo, & Hwang, 2012). The WCN value of atom i was calculated by$\mathrm{WCN}_i = \sum_{j \neq i}^{N} \frac{1}{r_{\mathrm{ij}}^2}$, where r_{ij} was the distance between the atom i and other atom j, N was the number of calculated atoms. In this work, atom i was defined as the C_α atom of SAV, and the different micro-environment properties were represented by calculated different atom type or source of atom j. The atom type of $j$could be C_α atoms, nitrogen atoms or oxygen atoms of an amino acid. The source of atom j could also be from the same protein chain with SAV or whole protein to represent the packing density of SAV. Moreover, the source could also be from the other protein chain or molecules such as DNA, RNA, ligand, or metal ion to represent the protein-protein or protein-molecule interaction. The packing density of SAV could be divided into different classification represented the micro-environment properties where the SAV located in, e.g. polar, hydrophobic, acidic or basic *et al* . according to the physicochemical properties of residues where C_α atom $j$belongs to. The same classification schemes were used as described in the sequence-based feature set, and the micro-environment-based descriptors were listed in Table S3.

3 | RESULTS

3.1 | Comparison of different feature sets

Table 2 compare the predictive performance of two prediction systems based on three different feature sets and the combined by second-level SVM, which are all optimized by using MCC as the fitness function. In our experiment, the individual prediction model by using the sequence-based feature scheme outperforms the other two. And then the model by using the micro-environment-based feature better than the structure-based feature scheme. The combined model by second-level SVM procedure with outstanding performance shows further information is indeed and very helpful to understand and determine the cancer-related factors.

On the other hand, between two systems, $\text{CanSavPre}_{\text{wm}}$performs better than $\text{CanSavPre}_{\text{w}}$ in three individual feature sets and the combined. That is because the distinct training and predicting models are built from the specific sub-group according to the wild and mutated amino acid type of SAV. Our best prediction system, $\text{CanSavPre}_{\text{wm}}$ with two-level SVM that combined sequence-, structure-, and micro-environment-based features, could distinguish the SAVs related to cancer or not, and the accuracy, the Matthews correlation coefficient, and F1-score yield to 90.88%, 0.77 and 0.83, respectively. The predictive performance for each wild type of SAV of system $\text{CanSavPre}_{\text{wm}}$ is illustrated in Table 3.

3.2 | Case Study: PI3K

The phosphatidylinositol-3-kinase (PI3K) signal pathway contributed to several cellular processes, such as metabolism, proliferation, differentiation, and activation. The PI3k/AKT/mammalian target of rapamycin (mTOR) signal pathway is one of the most vital intracellular pathways. However, it is also the most frequently dysregulated pathway correlated to almost all human cancer (Asati, Mahapatra, & Bharti, 2016; Benetatos, Voulgaris, & Vartholomatos, 2017; Dong et al., 2014; Hemmings & Restuccia, 2012). Amino acid mutation of PI3K is closely related to oncogenic transformation, and numerous SAVs have been recorded as cancer-related, such as P57S, Q75K, K111E, P134L, S361F, N380H, L634F, H677R, E713K, A723V, I776T, G890R, and L977I (Beadling et al., 2011; Hou et al., 2007; S. Jones et al., 2010; Kinross et al., 2012; Kuo et al., 2009; Pita, Figueiredo, Moura, Leite, & Cavaco, 2014). Figure 2 is illustrated the protein structure of PI3K and p85α complex (PDB ID: 5DXU) (Heffron et al., 2016), fourteen amino acids including a neutral and thirteen cancer-related SAVs are drawn as spheres. These cancer-related SAVs are all correctly predicted by our prediction system. It should be noted that another SAV, R104C, has been marked as neutral SAV and is also predicted correctly. The predicted results of PI3K are listed in Table 4.

3.4 | Case Study: D227Y of CD23

CD23 is the low-affinity receptor for IgE. It is expressed in several hematopoietic cells surface (Acharya et al., 2010), such as lymphocytes (Delespesse et al., 1991), monocytes (Vercelli et al., 1988), follicular dendritic cells (Krauss, Mayer, Rank, & Rieber, 1993; Rieber, Rank, Kohler, & Krauss, 1993), and bone marrow stromal cells (Fourcade et al., 1992). Several stimuli regulate the CD23 expression, which is the critical factor for B-cell activation, growth, and IgE production (OMIM#151445). The D227Y mutation generated from *FCER2* genetic altered had been reported in head and neck squamous cell carcinoma (HNSCC) (Stransky et al., 2011) and the colorectal neuroendocrine carcinomas mutational analyses project (Woischke et al., 2017). D227 located in one of the conserved double-loop, which is the interface between CD23 and the carbohydrate protein, Fcε3-4. Moreover, Ca^{2+} is a regulated ligand for CD23 binding affinity. With Ca^{2+} binding, the loop1 and loop4 would change the conformation and increase the binding affinity. D227 (loop1) and D258 (loop4) would form the additional salt bridges between CD23 and Fcε3-4 (Dhaliwal et al., 2013; Yuan et al., 2013). Though there are other bounds involved in CD23 and Fcε3-4 binding, the D227Y would affect the binding affinity and affect the IgE antitumor function (Figure 3).

Figure 4 shows the boxplot of the micro-environment descriptors in ASP altered to TYR sub-group. The distribution of cancer-related SAVs in several descriptors has a significant difference comparing to the neutral SAVs, and get a 95% confidence interval by z-test. The cancer-related SAVs are located in the relatively low packing density region, whether C_α atoms, nitrogen, or oxygen in a single SAV chain and whole protein. In the case of D227Y in CD23, it also has low WCN value in a single SAV chain but has relatively high WCN value in whole protein or other chains (Figure 4 a, b, and c). It is because D227 is located in the interface of CD23 and Fcε3-4 and is involved in the binding. Subsequently, the cancer-related SAVs has the lower distributions of specific tendency of H -neutral (AGPHY),V -small (GASCTPD), V -large(MHKFRYW), Z -low polarizability(GASDT), Z -high polarizability (KMHFRYW),P -neutral polarity (PATGS), F -basic (HKR),F -nonpolar (AFILMPVW), E -basic (HKR), E -aromatic (FWY) and E -aliphatic1 (AGP) of neighboring amino acid. This unique surrounding pattern is also found in the cases D227Y of CD23 (Figure 4 d, e, f, g, h, and i).

3.5 | Case Study: E194G of CASQ

The calsequestrin (CASQ) is the Ca^{2+} buffering protein, which could store large amounts of Ca^{2+} in the cardiac and skeletal muscles. Ca^{2+} is an essential molecular that could regulate diverse cellular processes, such as gene transcription, cell proliferation, or migration (Kim, Tam, Siems, & Kang, 2005; MacLennan, Abu-Abed, & Kang, 2002; Manno et al., 2017). Though most researches of CASQ are focus on the cardiac muscle, CASQ in the Ca^{2+} signal pathway is also vital in cancer research (Terentyev et al., 2003). It is reported that the Ca^{2+} signaling pathway is highly correlated to tumor growth or metastatic (Stewart, Yapa, & Monteith, 2015), and E194G of CASQ has been found in glioblastoma patients (Parsons et al., 2008). In CASQ, T189, E194, and D196 would form a pack harboring Ca^{2+} (Sanchez, Lewis, Danna, & Kang, 2012). Hence, this substitution, E194G, would lose it functional and destroy the Ca^{2+} binding (Figure 5).

Although no micro-environment descriptor has a significant difference at 95% confidence interval between the distribution of cancer-related and neutral SAVs in GLU altered to GLY sub-group, several relevant descriptors are found in the case of E194G in CASQ. E194 has higher WCN values of oxygen in a single SAV chain and atoms in other molecular due to CASQ is a GLU and ASP rich and Ca^{2+} buffering protein (Figure 6 a, b). Furthermore, for the micro-environment around E194, higher WCN values are found than the third quartile of cancer-related SAVs in H -polar (RKEDQN), V -medium (NVEQIL), Z -low polarizability(GASDT), P -high polarity (HQRKNED), F -acidic (DE), and E -acidic (DE) descriptors and lower than the first quartile in E -sulfur-containing (CM). The boxplot of the micro-environment descriptors in GLU altered to GLY sub-group is shown in figure 6.

4 | DISCUSSION

We have developed a two-level SVM system CanSavPre to predict cancer-related single amino acid variation. Not only protein sequences but also structures are used for descriptors extracted for model training. Our experiment showed much better improvement in the two-level prediction system, and it means more adequate information is necessary for identifying cancer-related SAVs from the divergent sequence of promiscuous protein function in an extensive network of cells. Even though without structure resolved for many sequences, the precise structure information can still be extracted with the help of the homologous search on the PDB database, like homology modeling method. To take into account the properties of the conformation and environment surrounding SAVs, the performance of the result in this work significantly enhanced obviously. Furthermore, the algorithm picked up the optimized the best combination feature vectors using for each kind of variation for specific amino acid type. Therefore, the difference is distanced feasibly.

In this work, we found that it is essential to divide the training data into proper subsets according to the wild and mutated type of SAVs when the model is trained. Moreover, by the feature selection procedure, the critical descriptors could be figured out. The relationship between the mutated residues and the interaction changed could be studied and characterized, primarily by analyzing the micro-environment-based feature set. Although further study is needed to reveal out the mechanism of cancer in most selected features, our results indicate that it is possible to predict cancer-related SAV reliably. Our work will provide a useful, practical tool for cancer research and precision medicine.

FIGURE LEGEND

FIGURE 1 The two-level SVM prediction system.

FIGURE 2 The protein structure of PI3K and p85 α complex . The PI3K and p85αcomplex (PDB ID: 5DXU) is drawn in the cartoon by PyMOL (Schrödinger, 2015). PI3K is colored wheat, and the p85a is colored in gray. The ARG104 presented in green spheres is a neutral SAV when mutated to CYS. The other residues presented in pink spheres are all cancer-related SAVs.

FIGURE 3 The superimposed of the structure of CD23 apo form and holo form from the complex of CD23 bound to Ca^{2+} and Fc ε3-4. The structure of the Ca^{2+} free wild type CD23 lentic

domain (PDB ID:4G96) (Yuan et al., 2013) is represented in the green cartoon. The structure of CD23 holo form bound to Ca^{2+} complexed with Fcε3-4 (PDB ID: 4GKO) (Yuan et al., 2013) is drawn in gray and wheat cartoons. Ca^{2+} is shown in a yellow bubble, and a close-up view shows the interface of CD23 and Fcε3-4. The D227 of the CD23 apo form is shown in the green stick. The salt-bridges forming residues in the CD23 holo form and Fcε3-4 complex, are also highlighted with sticks.

FIGURE 4 The boxplot of the micro-environment descriptors in the ASP altered to the TYR sub-group. All micro-environment descriptors are divided in nine groups, which are (a) atoms in SAV chain, (b) atoms in whole protein, (c) atoms in other chains or molecules, (d) H -group, (e) V -group, (f)Z -group, (g) P -group, (h) F -groups, and (i)E -groups. The white and grey boxes represented the distribution of cancer-related and neutral SAVs. The boxes have the red frame if the significant difference is found at a 95% confidence interval by z-test between cancer-related and neutral SAVs. The label of selected descriptors by the genetic algorithm are bold in the x -axis. The symbol stars are noted as the cases D227Y of CD23.

FIGURE 5 The protein structure of the human skeletal calsequestrin. The structure of CASQ (PDB ID:3UOM) (Sanchez et al., 2012) is drawn in the cyan cartoon by PyMOL. All of the yellow bubbles are Ca^{2+} in CASQ. Three Ca^{2+} binding residues are highlighted with sticks in deep pink and the SAV, E194G is a cancer-related SAV.

FIGURE 6 The boxplot of the micro-environment descriptors in GLU altered to GLY sub-group. All micro-environment descriptors are divided in nine groups, which are (a) atoms in SAV chain, (b) atoms in whole protein, (c) atoms in other chains or molecules, (d) H -group, (e) V -group, (f) Z -group, (g) P -group, (h) F -groups, and (i) E -groups. The white and grey boxes represented the distribution of cancer-related and neutral SAVs. The label of selected descriptors by the genetic algorithm are bold in the x -axis. The symbol stars are noted as the cases E194G of CASQ.

ACKNOWLEDGEMENTS

This work is supported by Ministry of Science Technology, Taiwan, Grant Number: MOST 108-2221-E-039-013- and China Medical University, Taiwan, Grant Number: CMU-108-MF-121.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

Study concept and protocol design : Jia-Jun Liu, Chih-Peng Lin and Chih-Hao Lu.

***Analysis and interpretation of data*: Jia-Jun Liu and Hsiao-Wei Wu.**

***Drafting of the manuscript*: Jia-Jun Liu and Yu-Jen Chang.**

***Server development*: Chin-Sheng Yu and Chih-Hao Lu.**

***Study supervision*: Chih-Hao Lu.**

REFERENCE

Acharya, M., Borland, G., Edkins, A. L., Maclellan, L. M., Matheson, J., Ozanne, B. W., & Cushley, W. (2010). CD23/FcepsilonRII: molecular multi-tasking. *Clin Exp Immunol, 162* (1), 12-23. doi:10.1111/j.1365-2249.2010.04210.x

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet, Chapter 7* , Unit7 20. doi:10.1002/0471142905.hg0720s76

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations.*Nat Methods, 7* (4), 248-249. doi:10.1038/nmeth0410-248

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215* (3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25* (17), 3389-3402.

Anoosha, P., Sakthivel, R., & Michael Gromiha, M. (2016). Exploring preferred amino acid mutations in cancer genes: Applications to identify potential drug targets. *Biochim Biophys Acta, 1862* (2), 155-165. doi:10.1016/j.bbadis.2015.11.006

Asati, V., Mahapatra, D. K., & Bharti, S. K. (2016). PI3K/Akt/mTOR and Ras/Raf/MEK/ERK signaling pathways inhibitors as anticancer agents: Structural and pharmacological perspectives. *Eur J Med Chem, 109* , 314-341. doi:10.1016/j.ejmech.2016.01.012

Beadling, C., Heinrich, M. C., Warrick, A., Forbes, E. M., Nelson, D., Justusson, E., . . . Corless, C. L. (2011). Multiplex mutation screening by mass spectrometry evaluation of 820 cases from a personalized cancer medicine registry. *J Mol Diagn, 13* (5), 504-513. doi:10.1016/j.jmoldx.2011.04.003

Benetatos, L., Voulgaris, E., & Vartholomatos, G. (2017). The crosstalk between long non-coding RNAs and PI3K in cancer. *Med Oncol, 34* (3), 39. doi:10.1007/s12032-017-0897-2

Bromberg, Y., & Rost, B. (2009). Correlating protein function and stability through the analysis of single amino acid substitutions. *BMC Bioinformatics, 10 Suppl 8* , S8. doi:10.1186/1471-2105-10-S8-S8

Care, M. A., Needham, C. J., Bulpitt, A. J., & Westhead, D. R. (2007). Deleterious SNP prediction: be mindful of your training data! *Bioinformatics, 23* (6), 664-672. doi:10.1093/bioinformatics/btl649

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., . . . Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res, 69* (16), 6660-6667. doi:10.1158/0008-5472.CAN-09-1133

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A Library for Support Vector Machines. *Acm Transactions on Intelligent Systems and Technology, 2* (3). doi:Artn 27

10.1145/1961189.1961199

Cheng, J., Randall, A. Z., Sweredoski, M. J., & Baldi, P. (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res, 33* (Web Server issue), W72-76. doi:10.1093/nar/gki396

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One, 7* (10), e46688. doi:10.1371/journal.pone.0046688

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins, 43* (3), 246-255.

David, A., Razali, R., Wass, M. N., & Sternberg, M. J. (2012). Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum Mutat, 33* (2), 359-363. doi:10.1002/humu.21656

Delespesse, G., Suter, U., Mossalayi, D., Bettler, B., Sarfati, M., Hofstetter, H., . . . Dalloul, A. (1991). Expression, structure, and function of the CD23 antigen. *Adv Immunol, 49* , 149-191. doi:10.1016/s0065-2776(08)60776-2

Dhaliwal, B., Pang, M. O., Yuan, D., Yahya, N., Fabiane, S. M., McDonnell, J. M., . . . Sutton, B. J. (2013). Conformational plasticity at the IgE-binding site of the B-cell receptor CD23. *Mol Immunol, 56* (4), 693-697. doi:10.1016/j.molimm.2013.07.005

Dong, P., Konno, Y., Watari, H., Hosaka, M., Noguchi, M., & Sakuragi, N. (2014). The impact of microRNA-mediated PI3K/AKT signaling on epithelial-mesenchymal transition and cancer stemness in endometrial cancer. *J Transl Med, 12* , 231. doi:10.1186/s12967-014-0231-0

Ferrer-Costa, C., Gelpi, J. L., Zamakola, L., Parraga, I., de la Cruz, X., & Orozco, M. (2005). PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics, 21* (14), 3176-3178. doi:10.1093/bioinformatics/bti486

Fourcade, C., Arock, M., Ktorza, S., Ouaaz, F., Merle-Beral, H., Mentz, F., . . . Mossalayi, M. D. (1992). Expression of CD23 by human bone marrow stromal cells. *Eur Cytokine Netw, 3* (6), 539-543.

Guo, W., Chen, Y., Zhou, X., Kar, A., Ray, P., Chen, X., . . . Wu, J. Y. (2011). An ALS-associated mutation affecting TDP-43 enhances protein aggregation, fibril formation and neurotoxicity. *Nat Struct Mol Biol, 18* (7), 822-830. doi:10.1038/nsmb.2053

Halasi, M., Wang, M., Chavan, T. S., Gaponenko, V., Hay, N., & Gartel, A. L. (2013). ROS inhibitor N-acetyl-L-cysteine antagonizes the activity of proteasome inhibitors. *Biochem J, 454* (2), 201-208. doi:10.1042/BJ20130282

Heffron, T. P., Heald, R. A., Ndubaku, C., Wei, B., Augistin, M., Do, S., . . . Olivero, A. G. (2016). The Rational Design of Selective Benzoxazepin Inhibitors of the alpha-Isoform of Phosphoinositide 3-Kinase Culminating in the Identification of (S)-2-((2-(1-Isopropyl-1H-1,2,4-triazol-5-yl)-5,6-dihydrobenzo[f]imidazo[1,2-d][1 ,4]oxazepin-9-yl)oxy)propanamide (GDC-0326). *J Med Chem, 59* (3), 985-1002. doi:10.1021/acs.jmedchem.5b01483

Hemmings, B. A., & Restuccia, D. F. (2012). PI3K-PKB/Akt pathway. *Cold Spring Harb Perspect Biol, 4* (9), a011189. doi:10.1101/cshperspect.a011189

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A, 89* (22), 10915-10919.

Hou, P., Liu, D., Shan, Y., Hu, S., Studeman, K., Condouris, S., . . . Xing, M. (2007). Genetic alterations and their relationship in the phosphatidylinositol 3-kinase/Akt pathway in thyroid cancer. *Clin Cancer Res, 13* (4), 1161-1170. doi:10.1158/1078-0432.CCR-06-1125

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics, 26* (5), 680-682. doi:10.1093/bioinformatics/btq003

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci, 8* (3), 275-282.

Jones, S., Wang, T. L., Shih Ie, M., Mao, T. L., Nakayama, K., Roden, R., . . . Papadopoulos, N. (2010). Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science, 330* (6001), 228-231. doi:10.1126/science.1196333

Juritz, E., Fornasari, M. S., Martelli, P. L., Fariselli, P., Casadio, R., & Parisi, G. (2012). On the effect of protein conformation diversity in discriminating among neutral and disease related single amino acid substitutions. *BMC Genomics, 13 Suppl 4* , S5. doi:10.1186/1471-2164-13-S4-S5

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers, 22* (12), 2577-2637. doi:10.1002/bip.360221211

Kaminker, J. S., Zhang, Y., Watanabe, C., & Zhang, Z. (2007). CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res, 35* (Web Server issue), W595-598. doi:10.1093/nar/gkm405

Kim, E., Tam, M., Siems, W. F., & Kang, C. (2005). Effects of drugs with muscle-related side effects and affinity for calsequestrin on the calcium regulatory function of sarcoplasmic reticulum microsomes. *Mol Pharmacol, 68* (6), 1708-1715. doi:10.1124/mol.105.016253

Kinross, K. M., Montgomery, K. G., Kleinschmidt, M., Waring, P., Ivetac, I., Tikoo, A., . . . Phillips, W. A. (2012). An activating Pik3ca mutation coupled with Pten loss is sufficient to initiate ovarian tumorigenesis in mice. *J Clin Invest, 122* (2), 553-557. doi:10.1172/JCI59309

Krauss, S., Mayer, E., Rank, G., & Rieber, E. P. (1993). Induction of the low affinity receptor for IgE (Fc epsilon RII/CD23) on human blood dendritic cells by interleukin-4.*Adv Exp Med Biol, 329* , 231-236. doi:10.1007/978-1-4615-2930-9_39

Kuo, K. T., Mao, T. L., Jones, S., Veras, E., Ayhan, A., Wang, T. L., . . . Shih Ie, M. (2009). Frequent activating mutations of PIK3CA in ovarian clear cell carcinoma. *Am J Pathol, 174* (5), 1597-1601. doi:10.2353/ajpath.2009.081000

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., . . . Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics, 25* (21), 2744-2750. doi:10.1093/bioinformatics/btp528

Li, J., Su, Z., Ma, Z. Q., Slebos, R. J., Halvey, P., Tabb, D. L., . . . Zhang, B. (2011). A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Mol Cell Proteomics, 10* (5), M110 006536. doi:10.1074/mcp.M110.006536

Lin, C. P., Huang, S. W., Lai, Y. L., Yen, S. C., Shih, C. H., Lu, C. H., . . . Hwang, J. K. (2008). Deriving protein dynamical properties from weighted protein contact number.*Proteins, 72* (3), 929-935. doi:10.1002/prot.21983

Lopez-Ferrando, V., Gazzo, A., de la Cruz, X., Orozco, M., & Gelpi, J. L. (2017). PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update.*Nucleic Acids Res, 45* (W1), W222-W228. doi:10.1093/nar/gkx313

Lori, C., Lantella, A., Pasquo, A., Alexander, L. T., Knapp, S., Chiaraluce, R., & Consalvi, V. (2013). Effect of single amino acid substitution observed in cancer on Pim-1 kinase thermodynamic stability and structure. *PLoS One, 8* (6), e64824. doi:10.1371/journal.pone.0064824

Lu, C. H., Chen, Y. C., Yu, C. S., & Hwang, J. K. (2007). Predicting disulfide connectivity patterns.*Proteins, 67* (2), 262-270. doi:10.1002/prot.21309

Ma, Y. S., Huang, T., Zhong, X. M., Zhang, H. W., Cong, X. L., Xu, H., . . . Fu, D. (2018). Proteogenomic characterization and comprehensive integrative genomic analysis of human colorectal cancer liver metastasis. *Mol Cancer, 17* (1), 139. doi:10.1186/s12943-018-0890-1

MacLennan, D. H., Abu-Abed, M., & Kang, C. (2002). Structure-function relationships in Ca(2+) cycling proteins. *J Mol Cell Cardiol, 34* (8), 897-918. doi:10.1006/jmcc.2002.2031

Manno, C., Figueroa, L. C., Gillespie, D., Fitts, R., Kang, C., Franzini-Armstrong, C., & Rios, E. (2017). Calsequestrin depolymerizes when calcium is depleted in the sarcoplasmic reticulum of working muscle. *Proc Natl Acad Sci U S A, 114* (4), E638-E647. doi:10.1073/pnas.1620265114

McFarland, C. D., Yaglom, J. A., Wojtkowiak, J. W., Scott, J. G., Morse, D. L., Sherman, M. Y., & Mirny, L. A. (2017). The Damaging Effect of Passenger Mutations on Cancer Progression. *Cancer Res, 77* (18), 4763-4772. doi:10.1158/0008-5472.CAN-15-3283-T

Nie, S., Yin, H., Tan, Z., Anderson, M. A., Ruffin, M. T., Simeone, D. M., & Lubman, D. M. (2014). Quantitative analysis of single amino acid variant peptides associated with pancreatic cancer in serum by an isobaric labeling quantitative method. *J Proteome Res, 13* (12), 6058-6066. doi:10.1021/pr500934u

Niroula, A., & Vihinen, M. (2015). Harmful somatic amino acid substitutions affect key pathways in cancers.*BMC Med Genomics, 8* , 53. doi:10.1186/s12920-015-0125-x

Parsons, D. W., Jones, S., Zhang, X., Lin, J. C., Leary, R. J., Angenendt, P., . . . Kinzler, K. W. (2008). An integrated genomic analysis of human glioblastoma multiforme.*Science, 321* (5897), 1807-1812. doi:10.1126/science.1164382

Pita, J. M., Figueiredo, I. F., Moura, M. M., Leite, V., & Cavaco, B. M. (2014). Cell cycle deregulation and TP53 and RAS mutations are major events in poorly differentiated and undifferentiated thyroid carcinomas. *J Clin Endocrinol Metab, 99* (3), E497-507. doi:10.1210/jc.2013-1512

Ponzoni, L., & Bahar, I. (2018). Structural dynamics is a determinant of the functional significance of missense variants. *Proc Natl Acad Sci U S A, 115* (16), 4164-4169. doi:10.1073/pnas.1715896115

Radusky, L., Modenutti, C., Delgado, J., Bustamante, J. P., Vishnopolska, S., Kiel, C., . . . Turjanski, A. (2018). VarQ: A Tool for the Structural and Functional Analysis of Human Protein Variants. *Front Genet, 9* , 620. doi:10.3389/fgene.2018.00620

Redler, R. L., Das, J., Diaz, J. R., & Dokholyan, N. V. (2016). Protein Destabilization as a Common Factor in Diverse Inherited Disorders. *J Mol Evol, 82* (1), 11-16. doi:10.1007/s00239-015-9717-5

Reeb, J., Hecht, M., Mahlich, Y., Bromberg, Y., & Rost, B. (2016). Predicted Molecular Effects of Sequence Variants Link to System Level of Disease. *PLoS Comput Biol, 12* (8), e1005047. doi:10.1371/journal.pcbi.1005047

Renaud, S., Seitlinger, J., Falcoz, P. E., Schaeffer, M., Voegeli, A. C., Legrain, M., . . . Massard, G. (2016). Specific KRAS amino acid substitutions and EGFR mutations predict site-specific recurrence and metastasis following non-small-cell lung cancer surgery. *Br J Cancer, 115* (3), 346-353. doi:10.1038/bjc.2016.182

Rieber, E. P., Rank, G., Kohler, I., & Krauss, S. (1993). Membrane expression of Fc epsilon RII/CD23 and release of soluble CD23 by follicular dendritic cells. *Adv Exp Med Biol, 329* , 393-398.

Sanchez, E. J., Lewis, K. M., Danna, B. R., & Kang, C. (2012). High-capacity Ca2+ binding of human skeletal calsequestrin. *J Biol Chem, 287* (14), 11592-11601. doi:10.1074/jbc.M111.335075

Schaefer, C., & Rost, B. (2012). Predict impact of single amino acid change upon protein structure.*BMC Genomics, 13 Suppl 4* , S4. doi:10.1186/1471-2164-13-S4-S4

Schrödinger, L. (2015). The PyMOL Molecular Graphics System, Version 1.8.

Shih, C. H., Chang, C. M., Lin, Y. S., Lo, W. C., & Hwang, J. K. (2012). Evolutionary information hidden in a single protein structure. *Proteins, 80* (6), 1647-1657. doi:10.1002/prot.24058

Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N., & Gaunt, T. R. (2013). Predicting the functional consequences of cancer-associated amino acid substitutions.*Bioinformatics, 29* (12), 1504-1510. doi:10.1093/bioinformatics/btt182

Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins.*Nucleic Acids Res, 40* (Web Server issue), W452-457. doi:10.1093/nar/gks539

Son, H., Kang, H., Kim, H. S., & Kim, S. (2017). Somatic mutation driven codon transition bias in human cancer. *Sci Rep, 7* (1), 14204. doi:10.1038/s41598-017-14543-1

Song, C., Wang, F., Cheng, K., Wei, X., Bian, Y., Wang, K., . . . Zou, H. (2014). Large-scale quantification of single amino-acid variations by a variation-associated database search strategy. *J Proteome Res, 13* (1), 241-248. doi:10.1021/pr400544j

Stewart, T. A., Yapa, K. T., & Monteith, G. R. (2015). Altered calcium signaling in cancer cells.*Biochim Biophys Acta, 1848* (10 Pt B), 2502-2511. doi:10.1016/j.bbamem.2014.08.016

Stransky, N., Egloff, A. M., Tward, A. D., Kostic, A. D., Cibulskis, K., Sivachenko, A., . . . Grandis, J. R. (2011). The mutational landscape of head and neck squamous cell carcinoma. *Science, 333* (6046), 1157-1160. doi:10.1126/science.1208130

Sundaram, L., Gao, H., Padigepati, S. R., McRae, J. F., Li, Y., Kosmicki, J. A., . . . Farh, K. K. (2018). Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet, 50* (8), 1161-1170. doi:10.1038/s41588-018-0167-z

Sunyaev, S., Ramensky, V., & Bork, P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet, 16* (5), 198-200.

Szpiech, Z. A., Strauli, N. B., White, K. A., Ruiz, D. G., Jacobson, M. P., Barber, D. L., & Hernandez, R. D. (2017). Prominent features of the amino acid mutation landscape in cancer. *PLoS One, 12* (8), e0183273. doi:10.1371/journal.pone.0183273

Tan, H., Bao, J., & Zhou, X. (2015). Genome-wide mutational spectra analysis reveals significant cancer-specific heterogeneity. *Sci Rep, 5* , 12566. doi:10.1038/srep12566

Teng, S., Srivastava, A. K., Schwartz, C. E., Alexov, E., & Wang, L. (2010). Structural assessment of the effects of amino acid substitutions on protein stability and protein protein interaction. *Int J Comput Biol Drug Des, 3* (4), 334-349. doi:10.1504/IJCBDD.2010.038396

Terentyev, D., Viatchenko-Karpinski, S., Gyorke, I., Volpe, P., Williams, S. C., & Gyorke, S. (2003). Calsequestrin determines the functional size and stability of cardiac intracellular calcium stores: Mechanism for hereditary arrhythmia.*Proc Natl Acad Sci U S A, 100* (20), 11759-11764. doi:10.1073/pnas.1932318100

Tsuber, V., Kadamov, Y., Brautigam, L., Berglund, U. W., & Helleday, T. (2017). Mutations in Cancer Cause Gain of Cysteine, Histidine, and Tryptophan at the Expense of a Net Loss of Arginine on the Proteome Level. *Biomolecules, 7* (3). doi:10.3390/biom7030049

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., & Ng, P. C. (2016). SIFT missense predictions for genomes. *Nat Protoc, 11* (1), 1-9. doi:10.1038/nprot.2015.123

Vercelli, D., Jabara, H. H., Lee, B. W., Woodland, N., Geha, R. S., & Leung, D. Y. (1988). Human recombinant interleukin 4 induces Fc epsilon R2/CD23 on normal human monocytes.*J Exp Med, 167* (4), 1406-1416. doi:10.1084/jem.167.4.1406

Wang, B., Li, J., Cheng, X., Zhou, Q., Yang, J., Zhang, M., . . . Li, J. (2018). NIPS, a 3D network-integrated predictor of deleterious protein SAPs, and its application in cancer prognosis. *Sci Rep, 8* (1), 6021. doi:10.1038/s41598-018-24286-2

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., & Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotechnol, 30* (2), 159-164. doi:10.1038/nbt.2106

Woischke, C., Schaaf, C. W., Yang, H. M., Vieth, M., Veits, L., Geddert, H., . . . Horst, D. (2017). In-depth mutational analyses of colorectal neuroendocrine carcinomas with adenoma or adenocarcinoma components. *Mod Pathol, 30* (1), 95-103. doi:10.1038/modpathol.2016.150

Yates, C. M., Filippis, I., Kelley, L. A., & Sternberg, M. J. (2014). SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features.*J Mol Biol, 426* (14), 2692-2701. doi:10.1016/j.jmb.2014.04.026

Yates, C. M., & Sternberg, M. J. (2013). Proteins and domains vary in their tolerance of non-synonymous single nucleotide polymorphisms (nsSNPs). *J Mol Biol, 425* (8), 1274-1286. doi:10.1016/j.jmb.2013.01.026

Yu, C. S., Chen, Y. C., Lu, C. H., & Hwang, J. K. (2006). Prediction of protein subcellular localization.*Proteins, 64* (3), 643-651. doi:10.1002/prot.21018

Yu, C. S., & Lu, C. H. (2011). Identification of antifreeze proteins and their functional residues by support vector machine and genetic algorithms based on n-peptide compositions. *PLoS One, 6* (5), e20445. doi:10.1371/journal.pone.0020445

Yuan, D., Keeble, A. H., Hibbert, R. G., Fabiane, S., Gould, H. J., McDonnell, J. M., . . . Dhaliwal, B. (2013). Ca2+-dependent structural changes in the B-cell receptor CD23 increase its affinity for human immunoglobulin E. *J Biol Chem, 288* (30), 21667-21677. doi:10.1074/jbc.M113.480657

Yue, P., Li, Z., & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol, 353* (2), 459-473. doi:10.1016/j.jmb.2005.08.020

Zhang, M., Wang, B., Xu, J., Wang, X., Xie, L., Zhang, B., . . . Li, J. (2017). CanProVar 2.0: An Updated Database of Human Cancer Proteome Variation. *J Proteome Res, 16* (2), 421-432. doi:10.1021/acs.jproteome.6b00505

Hosted file

`Table.docx` available at https://authorea.com/users/343673/articles/470297-cancer-related-single-amino-acid-variation-prediction