Accessible Surface Area and the Prediction of the Phenotypes of Missense Mutations

Eshel Faraggi*

Research and Information Systems, LLC,
1620 E. 72nd ST., Indianapolis, IN 46240, USA and
Physics Department, Indiana University Purdue
University Indianapolis, Indianapolis, IN 46202, USA

Robert L. Jernigan†

Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology,
Iowa State University, Ames, Iowa 50011, USA

Andrzej Kloczkowski‡

Battelle Center for Mathematical Medicine,
The Research Institute at Nationwide Children's Hospital, Columbus, OH 43205, USA and
Department of Pediatrics, The Ohio State University, Columbus, OH 43205, USA

(Dated: April 7, 2022)

Distinguishing between harmful and benign genetic variations is fundamental to our understanding of the relationship between genome and disease in general and for personalized medicine in particular. We investigated the relationship between predicted change in RASA and the phenotype of a missense mutation (MM). The ASAquick program was used to obtain RASA predictions for the original and mutated sequence and a parameter, δ, was introduced to assess the change in RASA for a given MM. We find that predicted RASA shows a robust, intricate signal with respect to genetic variation and that changes in RASA between variants can form a basis for a simple and quick predictor of the effect of MMs. Furthermore, we find that for hydrophobic residues, increase in the RASA corresponds to an increase in the likelihood that a MM would be harmful. For hydrophilic residues we find that a decrease in the RASA corresponds to a likelihood that a MM would be harmful. We also find that the size of the change in predicted RASA plays a role in determining the effect of a given MM. In future work we plan to use these results in developing more sophisticated forms of MM phenotype predictors.

Keywords: accessible surface area, ASAquick program, protein mutation, missense mutation, neutral mutation, deleterious mutation, phenotype prediction

I. INTRODUCTION

Understanding the physical manifestations of genetic variations is a prerequisite for personalized medicine.[1–4] Distinguishing computationally, between variation that are associated with damage (deleterious) and those that are not (benign or neutral) is one aspect of this research.[5–24] In this work we will consider Missense Mutation (MM) types: a single nucleotide mutation resulting in a codon that codes for a different amino acid. It is a type of nonsynonymous substitution. That is, a nucleic acid mutation that conserves the length of the protein and results in a single amino acid change. We will use the term Single Amino acid Variant (SAV) to describe such circumstances.[25]

The Accessible Surface Area (ASA), sometimes called Solvent-Accessible Surface Area, is the surface area of a protein (usually measured in Å^2) accessible to a solvent. It was proposed in 1971 by Lee and Richards[26] as one of major parameters for the description of protein structure. The ASA is usually computed by the rolling ball algorithm of Shrake and Rupley[27] by using a sphere of the radius of a solvent (typically 1.4 Åfor a water molecule) and rolling it along the van der Waals surface of a protein). Another quantity used to describe the surface of proteins is the Relative Accessible Surface Area (RASA). This quantity is obtained by dividing the ASA of a given residue by a maximum ASA value for that residue type. The maximum value is either obtained from a large dataset of protein structures or is calculated for that residue type in a linear peptide chain.

ASA and its change upon mutation have been useful tools in trying to predict if a given SAV is Deleterious (D) or Neutral (N). PolyPhen-2[14] uses the ASA and other information to predict a phenotype. Specifically, it uses the change in accessible surface propensity, which is a knowledge-based hydrophobic scale. In LS-SNP[5] both the Relative Surface Area (RSA) and the change in accessible surface propensity are used to predict the phenotype (D or N) of a given SAV. In both cases the ASA was calculated from three-dimensional model structures for a given sequence. These results indicate the importance of ASA in determining the phenotype of a given SAV.

Since most proteins do not have experimentally solved structures, and since creating computational models of three-dimensional protein structures is both time consuming and prone to errors, we would like to investigate here the ability of a dedicated ASA predictor in discriminating between D and N phenotypes.

We have two types of ASA predictors to choose from: those that use a Position Specific Scoring Matrix (PSSM)[28,29] from multiple alignments of the query sequence, and those that do not. Besides requiring significant computational time, PSSMs are obtained from sequence alignments. Since changing a single residue in a sequence will not typically alter the set of aligned sequences, the PSSM will not change for a SAV. ASA predictors that depend only on the specific sequence will have an advantage for predicting the effect of a SAV.

ASAquick[30] was designed by us to give quick ASA predictions. It does that by eliminating the use of PSSMs. Instead, it uses a physiochemical representation of the sequence[31] coupled with the BLOSUM62 substitution matrix representations of its residues.[32] It also captures some of the global nature of the protein: the residue length of the chain divided by 1000, the residue type composition of the whole chain (25 values), and the directional two-residue composition (625 values). 25 residue types are used to allow for the various characters reported by DSSP[33], these include atypical residues, unknowns and chain gaps in the structure file. ASAquick also uses an expanded training set with a PISCES[34] list of non-homologous protein chains with resolution better than 3 Åand sequence identity lower than 40%. Since numerous genetic mutations are possible, a faster way to predict their phenotype may find many uses.

In what follows we will test the relationship between the changing of ASA upon a single residue mutation and its effect on the phenotype of a given SAV. We specifically compare mutations on identical genes and compare between SAVs in those genes that are predicted by ASAquick to cause a large ASA change in the three-dimensional structure and those that are predicted to cause a smaller change. The main question we asked is: for a given gene, are those SAVs that cause more change in ASA associated with a D phenotype? As we shall see below the answer is yes.

II. MATERIALS AND METHODS

We use the PROVEAN dataset of human variants[18] as our test dataset. This file was downloaded from: http://provean.jcvi.org/downloads/uniprot.human.variants.tsv.gz, it contains a list of Uniprot gene IDs[35] along with a position and residue mutation, and a letter identifier, N or D, to indicate a neutral or deleterious mutation, respectively. There were 58,685 such variants reported. We chose this file because it is a well curated file from a substantiated method. From this dataset we found 1,226 genes for which there was both a neutral and a deleterious mutation reported, not necessarily in the same position. For some of these reported mutations we found discrepancy between the reported wild type residue at a mutation site and the residue found in Uniprot. We discarded these cases.

We also discarded mutations that occurred at the ends of the sequence since they were missing nearest-neighbors required for this study. For the remaining 1,209 genes we found a total of 6,070 neutral mutations and 17,017 deleterious ones. We will refer to this dataset as SA. Since there is a large discrepancy between the number of neutral and deleterious genes in SA we also constructed a more balanced set by randomly selecting 6,070 D variations. First, we selected a random D variation for each gene, then we randomly selected the remaining 4,861 variations from those D variations not selected in the first step. Combining this with all the 6,070 N variations in our possession, this set has a total of 12,140 variations split equally between the two phenotypes, across 1,209 genes. We will refer to this set as SB. Additionally, we also constructed a more balanced set by selecting all the genes that have at least 10 variants for both the neutral and the deleterious phenotypes. This set has 77 genes, with 1,644 neutral and 3,974 deleterious variants. We named this set SC.

To aid in visualizing the distribution of phenotypes per gene for set SA we did the following. We started by counting the number neutral (N) and deleterious (D) phenotypes for each of the genes. We then sorted separately each of these lists of the number of phenotypes and ploted that data indexed on the x-axis. This process is presented in Fig. 1. We first note that the x-axis is an arbitrary index in the sense that two points on the graph that share an x-value are not necessarily variations of the same gene. Second, it is worth noting that the area under the curve represents the total number of variation, i.e., 6070 neutral and 17017 deleterious. From this graph we can observe that there are approximately 50 more genes with a single N variation as those with a single D variation, around 100 more genes with two N variations as those with two D variations, and so forth. Note that for higher numbers of variations per gene we eventually find cases with more D variations as expected by the overall normalization. Also, we have restricted the y-axis range to aid in visualization.

We would like to study the relationship between phenotypic effect of a SAV (neutral or deleterious) and the change in ASA due to the residue mutation. We would like to address two questions: 1) Are D mutations more abundant in regions close to the surface since these are more prone to be functional sites, and 2) Do D mutations correlate with a grater change in ASA as compared to a neutral variants. As may be expected, the Relative Accessible Surface Area (RASA) is a better variable to work with. The RASA is the ratio between the ASA and a residue dependent maximum value. The maximum value is a dependent of the geometry of the sidechain. We use the same approach to normalization used in ASAquick[30]. To obtain predictions for the ASA we will use ASAquick.

The investigation into the first question is pretty straight forward. To that end, we averaged over samples of D and N type mutations separately. These were selected from the set SA. Each time, we selected 1000 mutations of a given type and calculated the average RASA as predicted by ASAquick. We repeated this for ten times and calculated the average and standard-deviation of this average. We find that for D type mutations the RASA is 0.127 ± 0.003, while for N type mutations the RASA is 0.140 ± 0.004. This shows that actually mutations at more buried site have a greater deleterious effects. This indicates that interfering with the stability of proteins may cause more harmful effect than interfering with functional sites. One should also consider that the number of residues on the surface of proteins is significantly lower than the number of buried ones. The rest of this work will focus on addressing the second question.

For a given variation we first predict the RASA for the wild type and mutated gene.

Since we are dealing with nsSNP, the mutated gene differs from the wild type by a change of a single residue along the chain. We record the RASA for the mutated site along with its two nearest neighbor residues along the sequence. We record the two neighboring residues to provide an additional test for consistency of the results. We will use the symbol R_i^W and R_i^M for the RASA at residue i of the wild type and mutated genes, respectively. Then we define the ratio between the mutated and wild type gene as:

$$\gamma_i = \frac{R_i^M}{R_i^W}. \tag{1}$$

The justification for using the ratio in Eq. (1), rather than the difference, is that in this way γ_i is more sensitive to changes between exposed and buried states. In addition, in this way the parameter γ_i is made more uniformly dimensionless. Note that a prediction of zero RASA is not possible because of the specific neural network architecture of ASAquick. Hence, $R_i^M > 0$ and $R_i^W > 0$.

For a given gene, we then average γ_i over all deleterious variants at all positions i in the gene, to obtain γ^D. We take a similar average to obtain γ^N for all neutral variants of a gene. Finally we take the difference between the two and define

$$\Delta\gamma = \gamma^D - \gamma^N, \tag{2}$$

as a measure of the importance of RASA change in determining the phenotype of a gene variation. A large positive $\Delta\gamma$ for a gene indicates a strong positive correlation between the effect on RASA and whether a variation is deleterious.

III. RESULTS

We started by averaging $\Delta\gamma$ for all genes in the sets SA, SB, and SC. We find a value of $\Delta\gamma$ of 0.292 for SA, 0.302 for SB, and 0.270 for SC. The p-value for the null hypothesis, that $\Delta\gamma$ for set SA averages to zero, is less than 0.01. This indicates a statistically significant relationship between RASA change and phenotypic outcome of a given variation.

In Fig. 2 we plot $\Delta\gamma$ for all genes in SA (red) and SB (green). The inset shows result for SC. In the inset, genes in SC were sorted along the x axis by their value of $\Delta\gamma$. This results in the smooth curve presented. In the main figure, genes were sorted by their $\Delta\gamma$ value obtained on SA and we plot both the results for SA and SB. We note some fluctuations in the values of $\Delta\gamma$ in set SB as compared to $\Delta\gamma$ values in SA. However, the major trend exhibited in SA is followed in SB. In SC we find similarly robust results. For SA, $\Delta\gamma$ is positive for 65% of the genes, for SB and SC the values are 62% and 79%, respectively.

Breaking up these results in terms of AA types reveals significant details. We performed the following analysis to obtain results per AA type. For each of the 1209 genes in the dataset we selected a random D SAV and a random N SAV and calculated $\Delta\gamma$ for this pair. We then collected all instances where the original residue (before mutation) of either the D or N variant, is of a given type and calculated the average $\Delta\gamma$ for this set. We repeated this process four times to obtain a statistical distribution over the random choice of representative mutations as a measure of the robustness of the results. The choice of four trials was arbitrary but dictated by the effort in generating each trial. In Fig. 3 we show the results of this analysis. Figure 3A shows the results where the initial type of the D SAVs is grouped and Fig. 3B shows the results considering the type of the N SAVs. Unfortunately we lack the data here to investigate combinations of AA types, e.g., initial final residue type, future work may address such issues. However, these results already suggest a robust relationship between AA type, ASA, and phenotype of a mutation. We will come back to this point in the discussion.

IV. DISCUSSION

We find that on average, D type SAVs tend to increase the ASA of the permutated residue, whereas N type SAVs tend to keep the predicted ASA of the residue unchanged. This is seen from Fig. 2 and also from the average γ. For deleterious SAVs we find $\gamma^D = 1.3 \pm 0.7$, while for neutral SAVs we find $\gamma^N = 1.0 \pm 0.4$. The standard deviations here are given as error estimates. The wide standard deviations for D type SAVs results from some D type mutations lowering the predicted ASA as we will discuss more shortly. Regardless, these results indicate that generally speaking D type SAVs give rise to a larger change in the ASA value, in the direction of more exposure. Intuitively, these results are reasonable as a harmful mutation may decrease the stability of a protein and hence increase its exposure to the solvent.

With respect to AA type we find a strong relationship between predicted change of ASA value and phenotype. The general trend of having an increase in predicted ASA for deleterious SAVs is overall maintained by a larger numerical value for increased RASA for D versus N type SAVs. However, as presented in Fig. 3, a deeper look into AA types reveals that for certain residues this trend is actually reversed: predicted RASA values are larger for N type SAV than for D type SAVs. Burial of a charged residue could be an intuitive example where a decrease in ASA may be more destabilizing than an increase.

For a given residue type, we calculate the difference between values of $\Delta\gamma$ in Figs. 3A and 3B, and plot the result in Fig. 4. To do this we calculated an average curve from the four different random trials in each panel. In this way, Fig. 4 represents the difference between predicted change of RASA for D and N type SAVs. In other words, a positive value indicates that the D type SAV changed the ASA more significantly than an N type SAV and vice versa for negative values. The error bars are the sum of the two standard deviations obtained from taking the averages over random trials in Figs. 3A and B. We can immediately recognize the charged and polar residues on the right side of the figure indicating a decrease of ASA values for them is more destabilizing, while on the left side of the figure we find more hydrophobic residues and those involved in specific spatial structures as cystine, for which an increase in ASA values is more destabilizing.

It is conceptually interesting to find out the ability of these initial findings to actually predict the phenotype of a given SAV. We present it here for furthering the understanding of the relationship between predicted ASA change and predicted phenotype. In future work we plan to include this understanding in a bigger scheme to predict the phenotype of a given

SAV.

The following scheme was applied. We first predicted the RASA of the original and permutated residue and recorded the change in RASA for the mutation. Then, based on Fig. 4, deleterious phenotype was selected for an increase in RASA for cases where the original residue was C, L, I, G, A, W, M, V, and H. Deleterious phenotype was also selected for a decrease in the RASA for cases where the original residue was R, P, D, Q, E, K, N, T, Y, S, and F. Neutral phenotype was selected in all other cases. We estimate the accuracy of our prediction for a given original residue type from the value of the y-axis in Fig. 4.

Table I gives the confusion matrix for this crude predictor. We see relatively bad N type prediction, with only 1617/6070 correctly predicted neutral SAVs. We suspect this is because of the unbalanced nature of the dataset. The Matthews correlation coefficient (MCC) we find for this case is 0.1 which is indicative of the poor quality of prediction of this simple transparent predictor, especially for neutral variants. For comparison, if we assume all increase in RASA prediction corresponds to a deleterious SAV, we arrive at a similar MCC. Finally, a somewhat more sophisticated predictor was constructed based on the magnitude of the change in RASA. We set the cutoff for change in predicted RASA at 0.01 and assigned any RASA change greater in magnitude as D type and the rest as N type. In this case the MCC increases by a factor of 2, indicating the importance of the magnitude of the predicted RASA change in determining the phenotype of a given SAV. While these results indicate that the simple predictors outlined here cannot compete with state-of-the-art predictors, they provide further evidence for the link between ASA change and SAV phenotype, and also provides clues for future designs of phenotype predictors.

We also selected a few cases where the structure of the relevant parts of the wild type protein were experimentally solved and available in the PDB[36]. We selected the first available structure with the greatest change in predicted RASA. Table II gives the details of the four cases we selected. Figure 5 gives the space filling representations of these four cases. We include all chains and ligands in the corresponding PDB file and color them green, blue, red, and yellow, depending on the number of chains and ligands. The permutated site is colored pink in all chains. This was done to check if the permutated site is involved in chain to chain interactions.

The first case is of P10828, thyroid hormone receptor beta. This is a nuclear hormone receptor, acting as a repressor or activator of transcription, and has a high affinity for thyroid hormones, including triiodothyronine and thyroxine. The mutation K342I is deleterious, implicated in generalized thyroid hormone resistance, a disease associated with goiter, abnormal mental functions, increased susceptibility to infections, abnormal growth and bone maturation, tachycardia, deafness, attention deficit-hyperactivity disorders (ADHD) and language difficulties. In the context of the isolated protein the wild-type residue has an RASA value of 0.1. Indicating that the charged Lysine residue is relatively buried and acts to balance the structure in that form. Replacing the charged LYS with the hydrophobic Isoleucine will most likely disrupt this balance. However, ASAquick does not detect this effect and focuses on the properties of the individual residues instead, erroneously predicting a reduction in RASA upon a change from a charged to a hydrophobic residue.

The second case is of O00584, ribonuclease T2. It is involved in processing mitochondrial and non-coding RNA, and in the innate immune response by recognizing and degrading RNAs from pathogens. The mutation C184R is deleterious, implicated in infantile-onset syndrome of cerebral leukoencephalopathy. Newborns with this disease develop significant impairments, including abnormal brain scans. The mutation site is almost completely buried and in this case ASAquick, while overestimating the RASA of the wild type, predicts a significant increase in the RASA upon mutation, and hence points towards instability and deleterious effects. In this case the original residue is Cystine which is overwhelmingly involved in stabilizing proteins through the formation of Cys-Cys disulfide bonds. Hence, this was an easy prediction of ASAquick to make.

The third case is of P48643, T-complex protein 1 subunit epsilon. It is a component of the chaperonin-containing T-complex that assists in the folding of proteins upon ATP hydrolysis. The mutation E146V is categorized as neutral. Interestingly, the proximal mutation H147R is deleterious and is associated with neuropathy. ASAquick here predicts relatively well that the site is exposed and points in the direction of stability upon mutation which would indicate it is neutral.

The final case is of P08246, neutrophil elastase. This protein is involved in the regulation of the function of white blood cells. The mutation C71R is actually categorized as neutral in the dataset we used. However, since its publishing new evidence point that this mutation is actually deleterious and is involved in neutropenia.[38,39] ASAquick actually points in that direction, predicting well for both the RASA of the wild-type protein and indicating a large increase in RASA upon mutation.

V. CONCLUSIONS

In this work we investigated the relationship between genetic variations and predicted ASA. As other studies have found, the results indicate that the change in ASA contains significant information about the phenotype of a SAV. We used the ASAquick program to obtain predictions. ASAquick is significantly faster than sequence alignments based predictions, and since it does not include sequence profiles generated from sequence alignments it is more sensitive to genetic variations that involve a mutation of a single amino acid.

A parameter, $\Delta\gamma$, was introduced to assess the change in RASA for a given SAV. We find that predicted RASA shows a robust, intricate signal with respect to genetic variation and that changes in RASA between variants are most important. We also find that for hydrophobic residues, an increase in the RASA increases the likelihood that a genetic mutation would be deleterious, while for hydrophilic residues a decrease in RASA increases the likelihood that a genetic mutation would be harmful. We find that the size of the change in predicted RASA plays a role in determining the effect (phenotype) of a given genetic mutation. In general, we find that D type SAVs tend to change the ASA more than N type SAVs which tend to keep the predicted ASA of the permutated residue unchanged. In future work we plan to use these results to help improve phenotype prediction.

ACKNOWLEDGMENTS

This research was supported by NSF grant DBI-1661391, and NIH grant R01GM127701.

---

[*] efaraggi@gmail.com

[†] jernigan@iastate.edu

[‡] Andrzej.Kloczkowski@nationwidechildrens.org

[1] John L Hartman, Barbara Garvik, and Lee Hartwell. Principles for the buffering of genetic variation. *science*, 291(5506):1001–1004, 2001.

[2] Thomas Mitchell-Olds, John H Willis, and David B Goldstein. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics*, 8(11):845–856, 2007.

[3] Kelly A Frazer, Sarah S Murray, Nicholas J Schork, and Eric J Topol. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251, 2009.

[4] Haiming Tang and Paul D Thomas. Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics*, 203(2):635–647, 2016.

[5] Rachel Karchin, Mark Diekhans, Libusha Kelly, Daryl J Thomas, Ursula Pieper, Narayanan Eswar, David Haussler, and Andrej Sali. Ls-snp: large-scale annotation of coding non-synonymous snps based on multiple information sources. *Bioinformatics*, 21(12):2814–2820, 2005.

[6] Lei Bao and Yan Cui. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, 21(10):2185–2190, 2005.

[7] Richard J Dobson, Patricia B Munroe, Mark J Caulfield, and Mansoor AS Saqi. Predicting deleterious nssnps: an analysis of sequence and structural attributes. *BMC bioinformatics*, 7(1):217, 2006.

[8] Pauline C Ng and Steven Henikoff. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.*, 7:61–80, 2006.

[9] Matthew A Care, Chris J Needham, Andrew J Bulpitt, and David R Westhead. Deleterious snp prediction: be mindful of your training data! *Bioinformatics*, 23(6):664–672, 2007.

[10] Gregory M Cooper and Jay Shendure. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, 12(9):628–640, 2011.

[11] Jian Tian, Ningfeng Wu, Xuexia Guo, Jun Guo, Juhua Zhang, and Yunliu Fan. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. *BMC bioinformatics*, 8(1):450, 2007.

[12] S Teng, E Michonova-Alexova, and E Alexov. Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions. *Current pharmaceutical biotechnology*, 9(2):123–133, 2008.

[13] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7):1073, 2009.

[14] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.

[15] Tao Huang, Ping Wang, Zhi-Qiang Ye, Heng Xu, Zhisong He, Kai-Yan Feng, LeLe Hu, WeiRen Cui, Kai Wang, Xiao Dong, et al. Prediction of deleterious non-synonymous snps based on protein interaction network and hybrid properties. *PLoS One*, 5(7):e11900, 2010.

[16] Emidio Capriotti and Russ B Altman. Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC bioinformatics*, 12(S4):S3, 2011.

[17] Emidio Capriotti and Russ B Altman. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. *Genomics*, 98(4):310–317, 2011.

[18] Yongwook Choi, Gregory E Sims, Sean Murphy, Jason R Miller, and Agnes P Chan. Predicting the functional effect of amino acid substitutions and indels. *PloS one*, 7(10):e46688, 2012.

[19] Margarida C Lopes, Chris Joyce, Graham RS Ritchie, Sally L John, Fiona Cunningham, Jennifer Asimit, and Eleftheria Zeggini. A combined functional annotation score for non-synonymous variants. *Human heredity*, 73(1):47–51, 2012.

[20] Jiaxin Wu and Rui Jiang. Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. *The Scientific World Journal*, 2013, 2013.

[21] Tikam Chand Dakal, Deepak Kala, Gourav Dhiman, Vinod Yadav, Andrey Krokhotin, and Nikolay V Dokholyan. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms in il8 gene. *Scientific reports*, 7(1):1–18, 2017.

[22] Mansi Desai and JB Chauhan. Computational analysis for the determination of deleterious nssnps in human mthfr gene. *Computational biology and chemistry*, 74:20–30, 2018.

[23] Mansi Desai and Jenabhai B Chauhan. Predicting the functional and structural consequences of nssnps in human methionine synthase gene using computational tools. *Systems Biology in Reproductive Medicine*, 65(4):288–300, 2019.

[24] Hind Bouafi, Sara Bencheikh, Mehdi AL Krami, Imane Morjane, Hicham Charoute, Hassan Rouba, Rachid Saile, Fouad Benhnini, and Abdelhamid Barakat. Prediction and structural comparison of deleterious coding nonsynonymous single nucleotide polymorphisms (nssnps) in human lep gene associated with obesity. *BioMed research international*, 2019, 2019.

[25] Yunhui Peng and Emil Alexov. Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding. *Proteins: Structure, Function, and Bioinformatics*, 84(2):232–239, 2016.

[26] Byungkook Lee and Frederic M Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–IN4, 1971.

[27] Andrew Shrake and John A Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of molecular biology*, 79(2):351–371, 1973.

[28] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[29] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang,

Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[30] Eshel Faraggi, Yaoqi Zhou, and Andrzej Kloczkowski. Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins: Structure, Function, and Bioinformatics*, 82(11):3170–3176, 2014.

[31] Eshel Faraggi, Yuedong Yang, Shesheng Zhang, and Yaoqi Zhou. Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure*, 17(11):1515–1527, 2009.

[32] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[33] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[34] Guoli Wang and Roland L Dunbrack Jr. Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.

[35] UniProt Consortium et al. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 46(5):2699, 2018.

[36] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[37] UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.

[38] Phil J Ancliff, Rosemary E Gale, Michael J Watts, Ri Liesner, Ian M Hann, Stephan Strobel, and David C Linch. Paternal mosaicism proves the pathogenic nature of mutations in neutrophil elastase in severe congenital neutropenia. *Blood*, 100(2):707–709, 2002.

[39] Manuela Germeshausen, Sabine Deerberg, Yvonne Peter, Christina Reimer, Christian P Kratz, and Matthias Ballmaier. The spectrum of elane mutations and their implications in severe

congenital and cyclic neutropenia. *Human mutation*, 34(6):905–914, 2013.

TABLE I: Confusion matrix for residue dependent model

		Predicted		
		D	N	Total
Actual	D	13563	3454	17017
	N	4453	1617	6070
	Total	18016	5071	23087

TABLE II: Structure Cases

PDB-ID[a]	Gene-ID[b]	Mutation	Phenotype	ASA[c]	RASA[d]	W-RASA[e]	P-RASA[f]	ΔRASA[g]
1N46	P10828	K342I	D	32	0.10	0.35	0.10	-0.25
3T0O	O00584	C184R	D	3	0.01	0.13	0.42	0.29
5UYX	P48643	E146V	N	114	0.42	0.33	0.08	-0.24
1H1B	P08246	C71R	N	8	0.04	0.07	0.29	0.23

[a] PDB ID for structure.[36] [b] Uniprot ID for gene.[37] [c] ASA in $Å^2$ at site of mutation obtained from PDB file. [d] RASA at site of mutation obtained from PDB file. [e] Predict RASA for wild-type gene at site of mutation. [f] Predict RASA for permutated gene at site of mutation. [g] Change in predicted RASA upon mutation.

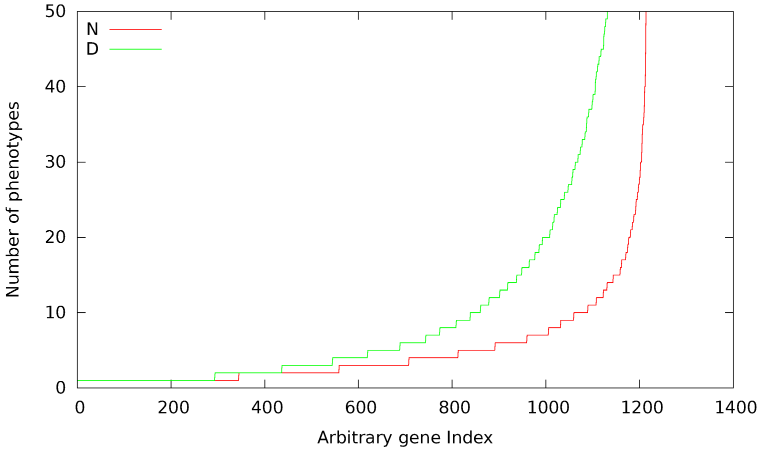


FIG. 1: Sorted number of neutral (green) and deleterious (red) SAVs versus an arbitrary index for the PROVEAN human gene variation dataset. Note that the area under the curve represents the total number of variations in the dataset and that we have restricted the y-axis to aid in visualization.

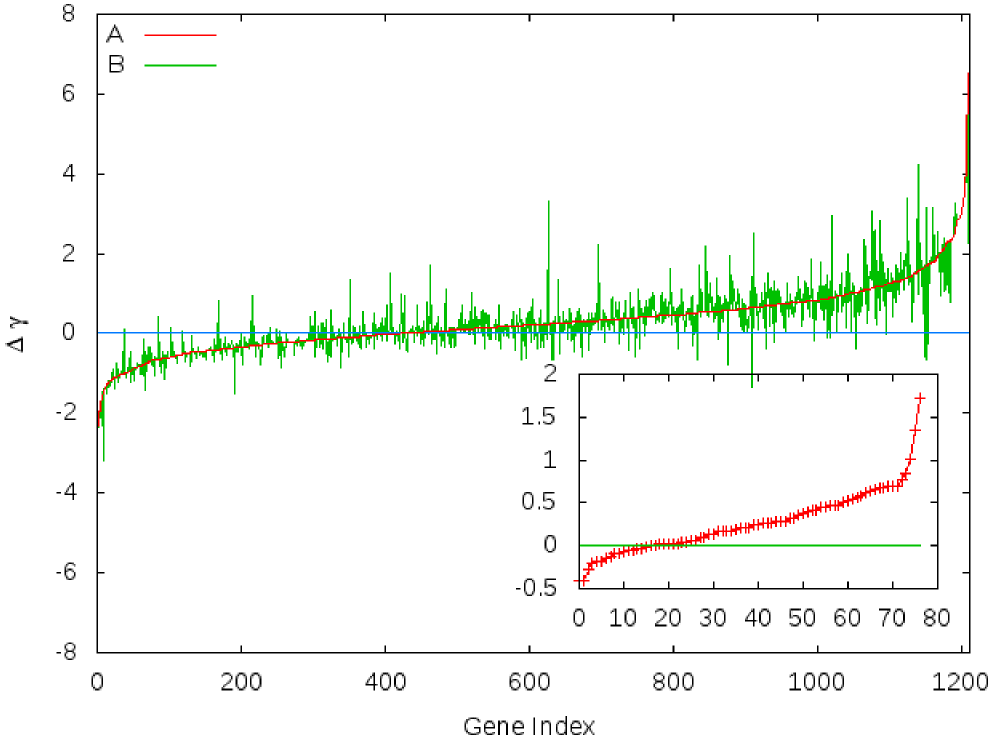


FIG. 2: The value of $\Delta\gamma$ for all genes in the set SA (red) and set SB (green). The inset shows result for SC. In the inset, genes in SC were sorted along the x axis by their value of $\Delta\gamma$. This results in the smooth curve presented. In the main figure, genes were sorted by their $\Delta\gamma$ value obtained on SA and we plot both the results for SA and SB based on this ordering.

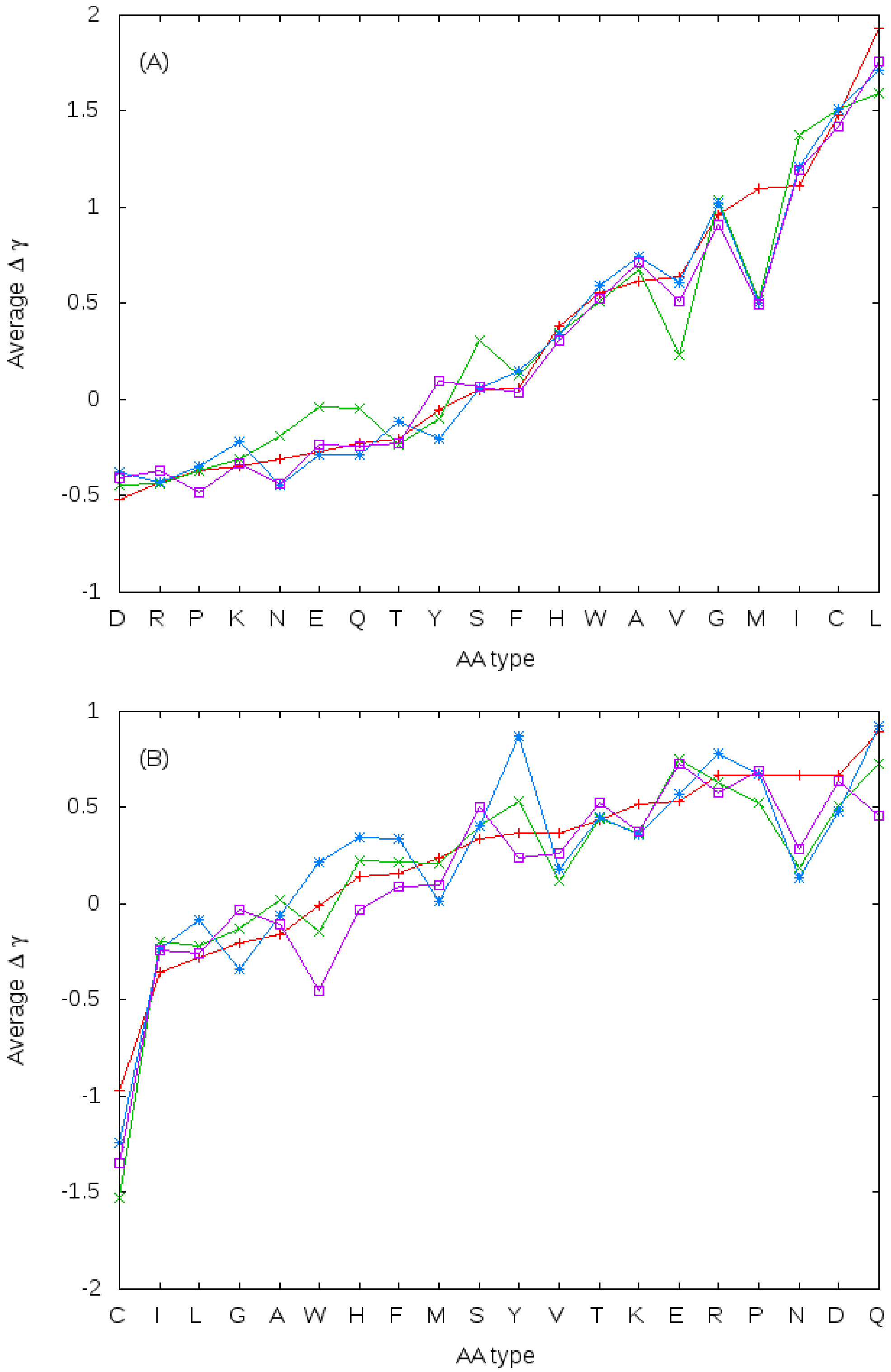


FIG. 3: Results per AA type. Panel A gives results where the initial type of the D SAV is grouped. Panel B gives results where the initial type of the N SAV is grouped. These results suggest a robust relationship between AA type, ASA, and phenotype of a mutation.

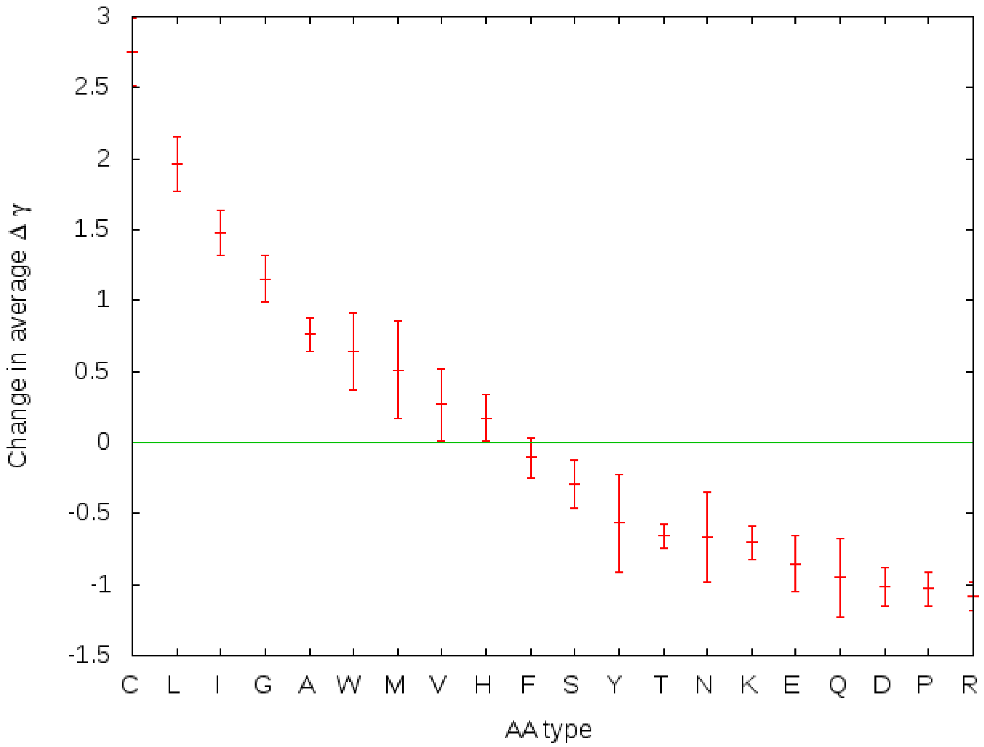
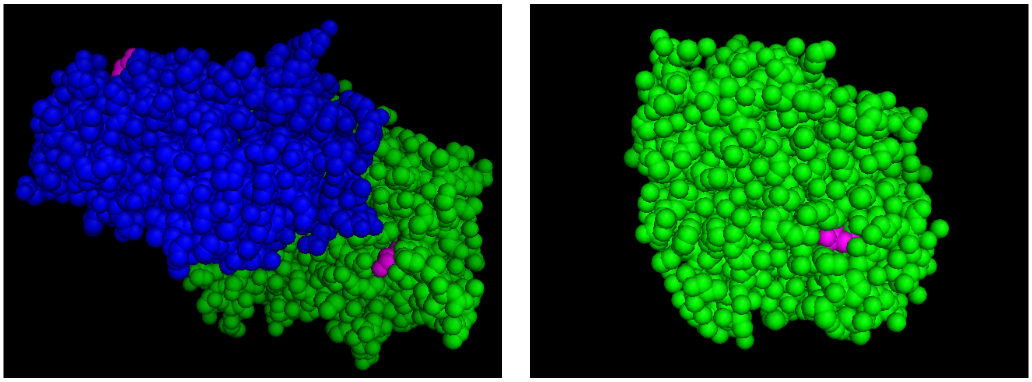
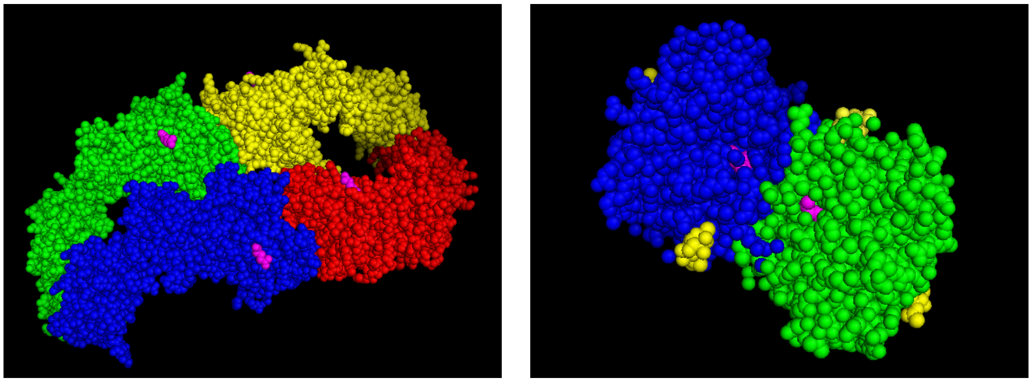


FIG. 4: Relative change in predicted RASA between D type SAVs and N type. Results per AA type. A positive value indicates that the D type SAV changed the ASA more significantly than an N type SAV and vice versa for negative values. We add a line at zero change to help guide the eye. The error bars are the sum of the two standard deviations obtained from taking the averages over random trials in Figs. 3A and B.



(a) 1N46 (b) 3T0O



(c) 5UYX (d) 1H1B

FIG. 5: Space filling representations of four example cases for experimentally known structures. First available structure with the greatest change in predicted RASA were selected. Table II gives the details for these cases. We included all chains and ligands and color them green, blue, red, and yellow, depending on the number of chains and ligands. The permutated site is colored pink in all chains. (A) P10828, thyroid hormone receptor beta. The mutation K342I is deleterious, implicated in generalized thyroid hormone resistance. (B) O00584, ribonuclease T2. The mutation C184R is deleterious, implicated in infantile-onset syndrome of cerebral leukoencephalopathy. (C) P48643, T-complex protein 1 subunit epsilon. The mutation E146V is categorized as neutral. (D) P08246, neutrophil elastase. The mutation C71R is deleterious and is involved in neutropenia.