Structural Location of Disease-Associated Single Nucleotide Polymorphisms

N. O. Stitziel¹, J. Tseng¹, D. Pervouchine², D. Goddeau², S. Kasi², J. Liang¹*

¹Department of Bioengineering
University of Illinois at Chicago
MC063, 851 S. Morgan St
Chicago, IL 60607

and

²Department of Biomedical Engineering
Boston University
Boston, MA 02215

*To whom reprint requests should be addressed. Email: jliang@uic.edu, Ph: (312)355-1789, Fax: (312)996-5921.
SUMMARY

Nonsynonymous single nucleotide polymorphism (nsSNP) of genes introduces amino acid changes to proteins, and plays an important role in providing genetic functional diversity. To understand the structural characteristics of disease-associated SNPs, we have mapped a set of nsSNPs derived from the Online Mendelian Inheritance in Man (OMIM) database to the structural surfaces of encoded proteins. These nsSNPs are disease-associated or have distinctive phenotypes. As a control dataset, we also mapped a set of nsSNPs derived from SNP database dbSNP to the structural surfaces of those encoded proteins. Using the alpha shape method from computational geometry, we examine the geometric locations of the structural sites of these nsSNPs. We classify each nsSNP site into three categories of geometric locations: those in a pocket or a void (type P), those on a convex region or a shallow depressed region (type S), and those that are completely buried in the interior (type I). We find that the majority (88%) of disease associated nsSNPs are located in voids or pockets, and they are infrequently observed in the interior of proteins (3.2% in the data set). We find that nsSNPs mapped from dbSNP are less likely to be located in pockets or voids (68%). We further introduce a novel application of Hidden Markov Models for analyzing sequence homology of SNPs on various geometric sites. For SNPs on surface pocket or void, we find that there is no strong tendency for them to occur on conserved residues. For SNPs buried in the interior, we find that disease associated mutations are more likely to be conserved. The approach of classifying nsSNPs with alpha shape and Hidden Markov model developed in this study can be integrated with additional methods and to improve the accuracy of predictions of whether a given nsSNP is likely to be disease associated.

Key words: single nucleotide polymorphism, alpha shape, hidden markov model, surface pockets.
Introduction

Single Nucleotide Polymorphisms (SNPs) are the most common form of human genetic variation. The coding regions of the human genome contain about 500,000 single nucleotide polymorphisms (SNPs)\(^1\). Among these, the non-synonymous SNPs (nsSNPs) cause changes in the amino acid residues, and are likely to be an important factor contributing to the functional diversity of encoded proteins in human population\(^2\). There are well known examples where nsSNPs affect the functional roles of proteins in signal transduction of visual, hormonal and other stimulants\(^3\)\(^-\)^4, in gene regulation by altering DNA and transcription factor binding\(^5\), and in maintaining the structural integrity of cells and tissues\(^6\). In addition, by affecting drug-target proteins such as G-protein coupled receptors\(^7\), enzymes\(^8\), ion channels\(^9\), and proteins involved in the detoxification pathways\(^10\), nsSNPs play important roles in the diverse responses in efficacy and toxicity of human population to therapeutic agents.

nsSNPs can affect human physiology through many different mechanisms. nsSNPs may inactivate functional sites of enzymes\(^11\), or alter splice sites and thereby form defective gene products\(^12\). They may also destabilize proteins, or reduce protein solubility\(^13\). To understand the mechanism of phenotypic variations due to nsSNPs, it is important to assess the structural consequences of the alteration of amino acid residue. A classical example is sickle-cell anemia, the first molecular disease discovered\(^14\). First studied by Sir Kendrew 50 years ago, sickle-cell anemia results from a single base change and residue V is changed to E at position 6 of the beta chain of hemoglobin. This residue is located at the interface of the alpha and beta chains, and the E6V mutation markedly reduces the solubility of the deoxygenated form of hemoglobin. The knowledge of structural role of this mutation is essential for understanding the disease mechanism of sickle cell anemia.

With the advent of high-throughput SNP detection techniques, the number of known nsSNPs is growing rapidly, providing an important source of information for studying the relationship between genotypes and phenotypes of human diseases. An important study has shown recently that there is a strong correlation between disease-associated polymorphism and sites of low solvent accessibility\(^15\). In this study, we introduce new geometric classifications for characterizing disease associated SNPs. Here we attempt to align SNPs to protein surface pockets and voids that may be potential functional binding regions.

Materials and Methods

Geometric Locations of Mutation Sites. Amino acid residues are located at different geometric locations. Some of them are located in the interior of a protein and have zero solvent accessibility. Others may be on the outer boundary surface of the protein, or on the wall surface of an interior void. In this study, we formally classify amino acid residues altered by nsSNPs to be located at three different geometric sites: (1) in the interior of proteins (type I); (2) on the wall of a surface pocket or an interior void (Class P); and (3) on the rest of the boundary surface regions which are either convex regions or shallow depressed regions (Class S) (Figure 1). A residue is located in the interior if it is in contact with many other residues and is fully buried such that it is inaccessible to a water molecule (modeled as a probe ball of radius 1.4 Å). Voids are unfilled
spaces inside the protein that are fully enclosed by atoms. A void is sufficiently enclosed if a solvent probe ball is too large to escape. Voids are traps for probe balls. Pockets are caverns that open to the outside of the protein through mouths that are small relative to cavern dimensions but big enough that the probe ball has access to the outside of the molecule. The mouth of a pocket is narrower than at least one cross section of the interior of the pocket. Depressions are concave regions on protein surfaces that have no constriction at the mouth. From the deepest part toward the outside, a depression widens monotonically. Residues in the interior may be part of the folding core important for structural stability and folding. Surface pockets and interior voids are frequently involved in molecular interactions. Convex regions and shallow depressed regions may be involved in protein-protein and protein-membrane interfaces. These topographic descriptions of the geometric locations of mutant residues therefore may provide useful information about the mechanism of nsSNPs affecting the functions of the protein.

These three types of geometric sites are computed using the alpha shape method from computational geometry. Alpha shape theory is a powerful geometric concept that formalizes the intuitive notion of “shapes”. It has been applied in a variety of studies of proteins, including protein folding, protein packing, enzyme functions, ligand recognition, and protein electrostatics. Based on the weighted Delaunay triangulation and the dual simplicial complex, it provides a mathematical framework for studying the topological, combinatorial, and metric properties of molecular shapes. Its theory and implementations has been described extensively elsewhere. Here we use alpha shape software Delcx, Mkalf, Volbl (downloadable from NCSA: http://www.ncsa.uiuc.edu), and castP (http://cast.engr.uic.edu) to characterize the structural locations of a set of disease-associated nsSNP. Delcx and Mkalf compute the Delaunay triangulation and the alpha shape of the molecules. castP identifies and measures protein voids and pockets from the alpha shapes, and Volbl is used to calculate the solvent accessible surface area.

Selection of disease associated nsSNPs. We select variant alleles of genes that are known to be disease producing or having distinctive phenotypes from the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/omim/). Only variants that are SNPs are included. Insertions, deletions, and other variant types are excluded. Because the OMIM database does not contain explicit sequence information, we further restrict the data to the subset of variant alleles where links to corresponding SwissProt Database entries exist. The positions of the SNPs on a gene in its SwissProt entry are deduced by measuring the relative distances in residue numbers between the OMIM alleles, and then by identifying the corresponding pairs of SNPs in SwissProt entry with the same relative distances in residue numbers. The nsSNPs and the full sequences of the genes are then extracted from the SwissProt database. Following this procedure, we are able to construct a data set of 2,128 variants on 310 genes from an original set of 5,467 nsSNPs in 1,061 alleles from the OMIM database.

Selection of control nsSNPs. As a control data set, we select variant alleles of genes that are annotated as nsSNP from dbSNP. Since there is no experimental conformation that none of these nsSNPs are associated with disease, this is not a perfect control data set. However, we assume that a smaller fraction of nsSNPs from dbSNP will be disease-associated when compared to nsSNPs extracted from OMIM database, where each entry contains annotation of disease phenotype based on
Structural Mapping of nsSNPs. It is a challenging task to determine whether a particular SNP is located on a protein surface region that may be functionally important. Existing computational methods such as docking at current stage are not well-suited for automated high-throughput analysis. The alpha shape method we introduce here provides a rapid and objective approach, which allows automatic structural mapping and provides classification of SNPs to different geometric sites on protein structures. We follow the links in the SwissProt entries of the nsSNP genes to the corresponding Protein Data Bank structures. A semiglobal pair-wise sequence alignment using dynamic programming was performed so the residue number in the SwissProt entry is mapped to the residue number in the PDB entry. Not all nsSNP alleles can be mapped, since there are occasionally missing residues in the PDB structure.

In the disease-associated dataset, we found that 924 variants in 82 alleles can be mapped to 129 PDB structures (some SwissProt entries have multiple PDB structures associated with them). For the control set, we found that 558 variants in 339 alleles can be mapped to 263 PDB structures. The structural locations of these nsSNPs are then classified into three categories: those that are in a surface pocket or an interior void (type P), those located on the convex regions or depressed regions of the protein (type S), and those that are mapped to the buried interior of the protein (type I) \(^{23}\). Among the 924 structural sites where the 2,128 OMIM-derived nsSNPs are mapped, 814 (88\%) are located in a surface pocket or a void (type P), 80 (9\%) are located on a convex regions or a shallow depression of the surface (type S), and 30 (3\%) are buried in the interior of the proteins (type I) (see Table 1). In the control set, out of the 558 structural sites, 381 (68\%) are of type P, 150 (27\%) are found to be type S, and 27 (5\%) are type I (see Table 2). All raw data can be found in Supplemental Table 3 and Supplemental Table 4. Although there are limitations in experimental resolutions of the protein structures, alpha shape computation provides exhaustive, quantitative, and precise identification, as well as measurement of these geometric sites \(^{18}\). These data will be made available at http://gila.bioengr.uic.edu/snp.

Calculating Conservation at Variant Sites. To perform this classification we use probabilistic models of protein family derived from hidden Markov models. Hidden Markov models (HMMs) have become a valuable method with wide applications in protein modeling, homology detection, and functional classification of putative genes. First introduced for protein modeling in Krogh et al.\(^{29}\) and Delcher et al.\(^{30}\), HMMs can be viewed as special cases in the more general framework of probabilistic Bayesian networks. In a typical HMM model for a protein family, a residue in a protein sequence can be in a match state, an insertion state, or a deletion state. A match state corresponds to a column in the well-aligned region of a multiple sequence alignment of members of the protein family, with position specific characteristic probabilities (called
emission probabilities) for each of the twenty amino acid residues. An insertion state corresponds to highly variable regions in the multiple sequence alignment. A deletion state corresponds to gaps in a few sequences at a column of the multiple alignment. An HMM model architecture includes a number of match states, insertion states, and deletion states, each with connections to other states. The 20 amino acid residues appear in each of these states with different characteristic probabilities (emission probabilities). Each state has its own probabilities of transiting to another state along the connections in the architecture (transition probabilities). If a protein sequence is given, the state to which each residue belongs is not directly known. That is, the state is hidden and needs to be computed. The emission probabilities and the transition probabilities also need to be estimated. Dynamic programming methods, including the Viterbi algorithm, are well suited to estimate these probabilities and align a sequence to the states in a protein family HMM model. Introductory overviews of HMMs can be found elsewhere. 

In this study we systematically examine positions of nsSNPs in motif regions of proteins. For this purpose, we use the PFAM database of probabilistic models of protein domains and families derived using the HMM method. PFAM has been used extensively in many bioinformatics studies and has played a seminal role in popularizing the HMM methodology, as well as allowing the evaluation of the strengths and weaknesses of HMM as a protein classification device. The specific alignment architecture featured in the PFAM database is described in. The analysis software used for our SNP analysis can be obtained by sending email to kasif@bu.edu. For each nsSNP, we first align the sequence of the protein to the protein HMM model. The decoding Viterbi algorithm aligns the SNP position to either a match state or an insert state of the HMM protein model. Once the SNP position is aligned to an HMM state, they can be classified into different categories based on the degree of conservation of the aligned position.

In our study, each entry in both OMIM- and dbSNP-derived nsSNP databases contain information about the protein sequence, the position of the nsSNP, the original amino acid residue at this position (we call it SNP amino acid), and the amino acid it was mutated to. We aligned every sequence from each database against the PFAM HMM database containing a library of protein family HMM models using HMMER 2.1.1, a hidden Markov alignment tool, with default parameters. We use an E-value of $10^{-3}$ as the significance threshold. To assess the extent of conservation, we compare the HMM consensus amino acid residue at position corresponding to the nsSNP. The consensus residue is the one with highest emission probability at that position. We define an amino acid residue to be highly conserved if its emission probability is 0.5 or greater. In general, based on degree of conservation and whether the SNP residue corresponds to the consensus amino-acid in the alignment, we can define multiple classification categories given below.

**Statistical Confidence Intervals by Bootstrap.** In order to assess the confidence intervals of parameters such as percentage of nsSNP at various geometric sites, we used the bootstrap technique. Let the true value of the distribution be $\theta$. Our estimator $T$ takes the value $t$, which is the estimated value for $\theta$. Our goal is to calculate a $(1-2\alpha)$ confidence interval for $\theta$. If we sample independently $R$ times from the distribution with replication, we have a simulated data set of $Y_1^*, \ldots, Y_R^*$. We estimate the parameter of interest from each of the $R$ samples, and obtain $T_1^*, \ldots, T_R^*$. 
A simple approach to estimate confidence intervals of \( \theta \) is to use the bootstrap estimates of quantiles for \( T-\theta \). The definition of probability implies:

\[
\Pr(a < T - \theta < b) \Rightarrow \Pr(T - b \leq \theta \leq T - a)
\]

For an equitailed \((l-2\alpha)\) confidence interval, we have \(\Pr(T - b \leq \theta \leq T - a) = 1 - 2\alpha\), and the following basic bootstrap confidence limits:

\[
t - (t_{(l,\alpha)}(1-\alpha) - t), \quad t - (t_{(l,\alpha)}(\alpha) - t)
\]

In our calculation, \( R \) is chosen as 10000, and \( \alpha \) is 2.5\% for estimation of 95\% confidence intervals.

**Results**

**Many Disease-associated nsSNPs Are Located in Pockets or Voids.** Compared to control nsSNPs, disease-associated nsSNPs derived from the OMIM database are more likely to be located in well-formed surface pocket or void locations (88\% vs. 68\%, with confidence intervals 77\%-100\% and 55\%-83\% at 95\% confidence intervals for disease and non-disease SNPs, respectively). An example of this type of nsSNP is insulin receptor tyrosine kinase. Its enzyme activity is essential in insulin-stimulated glucose transport in adipose, muscle and liver cells. In the disease-associated database derived from OMIM database, several nsSNP variant alleles of insulin receptor kinase are mutations of residues A1134 and M1153. Residue A1134 (red) (pdb code: 1ir3, Figure 2a) is a highly conserved residue located in a consensus sequence found in most tyrosine kinases \(^{38}\). This residue is mapped to a large binding pocket (green) for 5'-adenyly-imido triphosphate (ANP, yellow) and Mg++ (blue), and is close to both the Mg++ ion and the ANP ligand. Residue M1153 (red) is located in a smaller pocket (purple) near the ANP binding site. A M1153 mutation causes a defect in receptor internalization relative to normal receptors, and was demonstrated to cause insulin resistance \(^{39,40}\). There are additional examples of disease-associated nsSNPs located in pockets or voids. These examples indicate that when an nsSNP causes a mutation in an important protein surface pocket, there is an increased probability that such an nsSNP may alter protein function, leading to various disease phenotypes.

**Disease-associated nsSNPs on Convex Regions and Shallow Depressed Region.** If a convex region or a shallow depressed region of a protein participates in binding with other protein or membrane, nsSNPs on these regions may also cause diseases. For example, polycystic kidney disease (PKD) is an autosomal dominant disorder leading to renal cysts, liver cysts, intracranial aneurysm, and hypertension. The PKD1 gene encodes a membrane protein polycystin-1, which is essential for cell-to-cell interactions. A disease associated nsSNP variant allele of the PKD1 gene was identified as a missense R324L mutation in exon 5 \(^6\). Residue R324 (red) (pdb code: 1b4r, Figure 2d) is located on a convex region of the surface of the PKD domain. It is likely that this region is important for heterodimerization, and the R324L mutant form of PKD1 is incapable of
such oligomerization, resulting in the loss of channel activity. Disease nsSNPs are far less likely to be located in shallow depressed regions or convex regions (8.6% vs. 27%, at 95% confidence intervals of 4.9%-12.9% and 19%-36% for disease and non-disease SNPs, respectively. See Tables 1 and 2).

**nsSNPs in Buried Protein Interior.** The nsSNPs in our data set are less likely to be fully buried in the core of the proteins. Only 30 out of 924 OMIM nsSNP sites are located in the buried interior of the protein. An example is the estrogen receptor (ESR), which is a nuclear receptor. Breast cancer patients possessing estrogen receptor typically have a lower risk of relapse and better overall survival \(^4\). The variant in our database, a C447A mutant of ESR displayed a dose-response shift for estradiol in transactivation studies \(^2\). C447 (red) (pdb code: 3ert, Figure 2g) is located in a tightly packed region of the protein. It has 30 atomic contacts with 12 residues, including two ionizable residues (E443, E444), two polar residues (S450, T485), and one aromatic residue (F445). The substitution of a Cys by a short Ala presumably leads to the loss of many favorable contacts and hence the loss of thermal stability and binding affinity.

Why are disease-associated nsSNPs infrequently observed in truly buried sites? One reason may be that many buried residues are not accessible for molecular recognition, and mutations on these sites do not directly affect the binding events of the protein. In addition, frequently mutations in the protein core do not strongly affect the stability of a protein \(^4\). Theoretical studies of design patterns of protein folding show that proteins with stable structures often are tolerant to mutation \(^4\). An additional reason could be that if a residue in the protein core is critical for protein folding stability or for folding kinetic accessibility, it is likely that a mutation at this site will be fatal and such genotype may be eliminated early in the stages of evolution, and hence not observed in current human population.

**SNP Locations and Conserved Residues.** In this study we attempt to provide an evolutionary perspective on disease SNPs. We address this question by introducing a novel technique using Hidden Markov models (HMMs) for classifying SNPs. An important application of HMM models is to align protein subsequences to conserved HMM motif regions. For positions in conserved motif regions (called a match state or an alignment state in an HMM model), the relative entropy is low. Regions of a protein that are not well-conserved are aligned to high entropy positions (called insert states in an HMM model). One intuitive theory might suggest that a disease SNP is likely to be associated with a highly conserved residue, since evolution may potentially select against frequent mutations in sites where changes might be harmful. In this study we attempt to gain an insight into this question by correlating geometric locations of disease SNPs and the degree of their conservation in the protein family.

We divide all SNP’s into the following five categories:

1. SNP amino acid is the same as HMM consensus amino acid in the aligned position (AP) and the later is highly conserved.
2. SNP amino acid is the same as HMM consensus amino acid in AP and the later is not highly conserved.
3. SNP amino acid is different from HMM consensus amino acid in AP that is conserved.
4. SNP amino acid is different from HMM consensus amino acid in AP that is not conserved.
The first category corresponds to our intuition about disease SNPs that many disease associated SNPs might be highly conserved residues. The biological reality appears more complex. The second category includes SNPs that are not located in conserved positions or SNPs in positions where the conservation is relatively weak. It is well known that amino acid residues at many positions can be substituted and the protein still maintains a given function. The third category includes SNPs aligned to relatively conserved positions but the SNP residue is not the same as the conserved residue. This category is not expected to occur often. Indeed, if the HMM "decides" to align the pre-mutation residue to a position where a different and highly conserved amino-acid is usually located, then the alignment score is decreased as a result. This misalignment can only be compensated by scores from strong alignment at other positions. The fourth category is a "noise" category that could be a result of an incorrect alignment or other features of the probabilistic model (Table 1 and 2).

These four HMM-based classifications can be contrasted and correlated to the geometric structural locations of SNPs obtained from the alpha shape calculations. SNPs can be classified into four groups according to the geometric locations predicted by the alpha shape: pocket, surface, interior, and not predicted. The intersection of these categories generates $5 \times 4 = 20$ different groups. The results for 15 of the 20 groups where alpha shape calculations are available are summarized in Tables 1 and 2.

Several interesting observations emerge from our analysis using HMM models and alpha shape calculations. First, it appears that for disease-associated SNPs located in the interior of proteins, they are more likely to be aligned to the most probable residue of a position in a conserved region of the HMM model (category 1). It is possible that if a disease SNP falls within the interior of a protein, and if it is at a well-conserved location, it is likely that the damage would be dramatic. Since residues at most interior positions are unlikely to be directly involved in protein function, the conserved residues may be important for protein stability or folding accessibility, and mutations there are likely to have severe phenotypic changes and may be eliminated quickly by evolution. Second, for disease-associated SNPs located in protein surface pocket and surface regions, we found that they are evenly distributed among the first three HMM categories, and are less likely to be from the fourth “noise” category. That is, for disease nsSNP located on pockets and surfaces, there is a significant fraction of them that do not matched to highly conserved sites (not category 1). This is in contrast to our anticipation that the majority of disease SNPs would align to highly conserved residues. It is interesting to note, however, that if the SNP and the consensus residue are different (category 3 and 4) the position is much more likely to be conserved.

The data from the control set of nsSNPs that lack evidence of disease and other significant phenotypic changes shows a different pattern. These nsSNPs are less likely to be located on well-formed existing protein surface pockets. In addition, it appears that for all structural classes of non disease-associated SNPs, they are most likely to differ from the consensus residue (categories 3 and 4). After removing SNPs where HMM alignments are not available, we find that for non-disease associated nsSNPs the percentage of residues that are not the consensus residues is 70.9%, 49.5% of which are located in a conserved region, and 33.3% are not located in a conserved region. In contrast, about 39.7% of disease associated nsSNPs
are not the same as the consensus residues, of which 33.3% are located in a conserved region and 6.4% are located in a variant region. The difference is especially significant for the SNPs located in the interior of proteins. Where it was most likely to lie in a conserved position for the disease-associated SNPs, almost no interior SNPs are found to be conserved in the control set.

**Discussion**

In this study, we have described a new approach for SNP classification. For SNPs that can be mapped to protein structures, we classify them into three geometric sites: those in a pocket or a void, those on a convex region or a shallow depressed region, and those buried in the interior. Specifically, we find that the majority of disease-associated nsSNPs are located in voids or pockets on proteins, and only a small number of SNPs are completely buried in the interior (Fig 3). For disease SNPs on surface pocket or void, there is no strong tendency for them to occur on conserved residues. For SNPs buried in the interior, we find that disease associated mutations are more likely to be conserved. This is quite different from the control set, in which very few interior SNPs are conserved. Our geometric descriptions and variant allele information are different from annotations contained in other database, such as PFAM, where fold information from SCOP and active site information from SwissProt are also provided.

A fundamental challenge in analyzing disease SNPs is the relative scarcity of alleles that can be mapped to three-dimensional protein structures. To gain some understanding of the robustness of the observed statistics of disease and non-disease nsSNPs at different geometric location with different conservation, we employed the bootstrap technique to assess the 95% confidence intervals. This allows us to point to some tentative observation with the safe-guard of statistical evidence in the form of fairly conservative confidence intervals. Although ultimately large amount of future data will be needed to sharpen results obtained in this study, we believe our analysis provides a useful picture of the molecular structural nature of disease nsSNPs.

Non-disease associated nsSNPs are apparently under different selection pressure. Although in this study we have not employed detailed phylogenetic analysis and a rigorous conclusion cannot be drawn, we found that nsSNPs not associated with disease are more frequently occurring in positions of genes whose wild types residue is already different from that of the consensus residue (categories 3 and 4). That is, if one assumes that the consensus residue represents the protein family well, nsSNPs that do not impact phenotypes frequently occur in sites of genes that have already diverged from the consensus residue, the latter may be similar to the residue in the ancestral gene. In contrast, for disease associated nsSNPs, they are more likely to be mutations from the consensus residues (categories 1 and 2), indicating that consensus residues at these sites are phenotypically important.

Disease nsSNPs are found to be far less likely to be located in shallow depressed regions or convex regions of protein. This may indicate that protein-protein interaction and membrane binding interface is not a great source of disease causing nsSNPs.
This observation is consistent with current understanding that generic hydrophobic interactions plays important role in protein-protein interactions \cite{45,46}. Indeed, the number of hot-spot residues that are critical for stabilizing protein-protein interactions is small, and the majority of mutations in the interface have little effects on the stability of protein-protein interactions \cite{47}.

The control data set extracted from dbSNP is important for this study, because it contains mostly alleles with no evidence of disease causing or strong phenotypic changes. Although it is possible some of the nsSNPs in this set may turn out to be diseases related if more vigorous biochemical and clinical studies are carried out, we expect that alleles contained in dbSNPs are mostly neutral markers of mutations with little phenotypic changes. We expect that if a true negative control data set can be established, that is, if the neutrality of each of the nsSNPs in such a set can be established by comprehensive biochemical and clinical studies, the observations can be made more significant.

In a recent study \cite{48}, a set of generic structure and sequence-derived features are developed based on lac repressor and lysozyme mutation data, and statistical F-test and chi-square test are applied to identify those that are predictive of functional effects of mutations. The most discriminating structural features are found to be solvent accessibility and experimental B-factors. Our study examines a large number of protein structures derived comprehensively from the OMIM database, and goes beyond solvent accessibility description and introduces geometric features that are likely to be related to protein functions \cite{16}.

The integrated structural/sequence methodology described in this study can be further developed into a computational method for predicting whether any given SNP is likely to be disease associated. This prediction requires using a standard application of Bayes’ Rule in computing the probability of a SNP being a disease SNP given its structural/sequence classification from the statistics computed here, namely the conditional probability of structure/sequence class given disease SNP and the general statistical distribution of disease SNPs and structure/sequence classes. These statistics must be carefully combined with the more established statistical analysis of SNPs with respect to their polymorphism across different populations \cite{49}. As SNP projects rapidly proceed and sufficient data has been accumulated, the new structure/sequence methodology described in this paper promises to provide an improved diagnostic capability of SNPs to be disease associated.
Table 1. Sequence conservation and geometric location of disease-associated nsSNP sites. The 95% confidence interval is given in brackets for each value.

<table>
<thead>
<tr>
<th>Category</th>
<th>Alpha shape predicted</th>
<th>Totals</th>
<th>Percentages after removing Cat. 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pocket</td>
<td>Surface</td>
<td>Interior</td>
</tr>
<tr>
<td></td>
<td>22% [19%-24%]</td>
<td>3% [2%-4%]</td>
<td>2% [0.8%-2.4%]</td>
</tr>
<tr>
<td></td>
<td>27%[24%-30%]</td>
<td>3% [2%-4%]</td>
<td>0.76% [0.22%-1.4%]</td>
</tr>
<tr>
<td></td>
<td>28% [25%-31%]</td>
<td>2% [1%-3%]</td>
<td>0.87% [0.32%-1.5%]</td>
</tr>
<tr>
<td></td>
<td>5% [4%-7%]</td>
<td>0.76% [0.21%-1.2%]</td>
<td>0.00% [0%-0.1%]</td>
</tr>
<tr>
<td></td>
<td>6% [4%-7%]</td>
<td>0.54% [0.1%-1%]</td>
<td>0.11% [0%-0.32%]</td>
</tr>
<tr>
<td></td>
<td>88% [77%-100%]</td>
<td>8.6% [4.9%-12.9%]</td>
<td>3.2% [1.3%-5.7%]</td>
</tr>
</tbody>
</table>

Table 2. Sequence conservation and geometric location of nsSNP sites from dbSNP. For each cell, the 95% confidence interval is given in brackets while the percentage of the total distribution is given in parenthesis.

<table>
<thead>
<tr>
<th>Category</th>
<th>Alpha shape predicted</th>
<th>Totals</th>
<th>Percentages after removing Cat. 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pocket</td>
<td>Surface</td>
<td>Interior</td>
</tr>
<tr>
<td></td>
<td>7% [5%-9%]</td>
<td>2% [7%-3%]</td>
<td>0.18% [0%-0.54%]</td>
</tr>
<tr>
<td></td>
<td>10% [7%-12%]</td>
<td>4% [2%-5%]</td>
<td>0.54% [0.1%-1.6%]</td>
</tr>
<tr>
<td></td>
<td>27% [23%-30%]</td>
<td>10% [8%-13%]</td>
<td>2% [0.9%-3%]</td>
</tr>
<tr>
<td></td>
<td>10% [8%-13%]</td>
<td>6% [4%-8%]</td>
<td>0.72% [0.18%-1.4%]</td>
</tr>
<tr>
<td></td>
<td>15% [12%-18%]</td>
<td>5% [4%-7%]</td>
<td>1% [0.54%-2.5%]</td>
</tr>
<tr>
<td></td>
<td>68% [55%-83%]</td>
<td>27% [19%-36%]</td>
<td>4.8% [2%-9%]</td>
</tr>
</tbody>
</table>
Figure 1. Any amino acid residue of a protein can be classified into three geometric locations: those located in a pocket or a void (type P), those located on a convex region or a shallow depressed region of protein surface (type S), and those that are completely buried in the interior (type I). These locations and their geometric types are illustrated here. Voids are completely enclosed and have no access to the outside bulk. Pockets are connected to the outside by narrow neck(s).

Figure 2. Geometric locations of nsSNPs. About 88% of the structural sites of nsSNPs are located in a pocket or void of the protein (type P), about 8.6% are on the rest of the outer surfaces (type S), which are either convex regions or shallow depressed regions. About 3.2% are fully buried in the interior (type I). Examples of geometric sites of nsSNPs: (Top level from left to right) (a) Insulin receptor tyrosine kinase (1ir3). nsSNP sites are colored red. Residue A1134 (red) is located in the large binding pocket (green) of ANP ligand (yellow) and Mg ion (blue). M1153 (red) is in a smaller pocket (purple) near the ANP binding site. (b) Alcohol dehydrogenase (1htb). Residue R47 (red) is located in the NAD (yellow) binding pocket (green), and R369 (blue) is also near this pocket. (c) Isovaleryl-CoA dehydrogenase (1ivh). G170 (red) is located in a small pocket (green) away from the FAD (yellow) and CoA persulfide binding site. (d) PKD domain from polycystin-1 (1b4r). R324 (red) is located on a convex region of the protein. (e) Fructose-bisphosphate aldolase (4ald). Residue 128 (red) is located on a convex region of the protein surface. (f) Carbonic anhydrase I (1azm). Residue 253 (red) is at an exposed convex region away from the substrate binding site. (g) Estrogen receptor (3ert). C447 (red) is located in a tightly packed region of the protein with 30 atomic contacts with 12 residues.
REFERENCES


Figure 1.
Figure 2.
Fig. 3. The distribution of disease and non-disease nsSNPs at different geometric locations.