doi:10.1016/S0022-2836(03)00240-7 J. Mol. Biol. (2003) 327, 1021–1030

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Structural Location of Disease-associated Single-nucleotide Polymorphisms

Nathan O. Stitziel[1], Yan Yuan Tseng[1], Dimitri Pervouchine[2] David Goddeau[2], Simon Kasif[2] and Jie Liang[1]*

[1]*Department of Bioengineering SEO, MC-063, University of Illinois at Chicago, Room 218 851, S. Morgan Street, Chicago IL 60607-7052, USA*

[2]*Department of Biomedical Engineering, Boston University Boston, MA 02215, USA*

**Corresponding author*

Non-synonymous 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 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 one of 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 buried completely 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 (HMM) 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 HMM developed in this study can be integrated with additional methods to improve the accuracy of predictions of whether a given nsSNP is likely to be disease-associated.

Keywords: 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 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 the 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 the 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

Abbreviations used: SNP, single-nucleotide polymorphism; nsSNP, non-synonymous SNP; PKD, polycystic kidney disease; ER, estrogen receptor; HMM, hidden Markov models; AP, aligned position.

E-mail address of the corresponding author: jliang@uic.edu

sites and thereby form defective gene products.[12] They may 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 John 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 reduces the solubility of the deoxygenated form of hemoglobin markedly. 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.

Results

Many disease-associated nsSNPs are located in pockets or voids

Compared to control nsSNPs, disease-associated nsSNPs derived from the online mendelian inheritance in man (OMIM) database are more likely to be located in well-formed surface pocket or void locations. Of the disease-associated nsSNPs derived from OMIM, 88% are located in pockets or voids (with 95% confidence interval of 77–100%), while 68% of non-disease control SNPs are located in pockets or voids (with 95% confidence intervals of 55–83%). An example of this type of nsSNP is insulin receptor tyrosine kinase. Its enzyme activity is essential for insulin-stimulated glucose transport in adipose, muscle and liver cells. In the disease-associated database derived from the OMIM database, several nsSNP variant alleles of insulin receptor kinase are mutations of residues A1134 and M1153. Residue A1134 (red) (PDB code 1ir3, Figure 1(a)) is a highly conserved residue located in a consensus sequence found in most tyrosine kinases.[16] This residue is mapped to a large binding pocket (green) for 5′-adenyly-imido triphosphate (ANP, yellow) and Mg^{2+} (blue), and is close to both the Mg^{2+} and the ANP ligand. Residue M1153 (red) is located in a smaller pocket (purple) near the ANP binding site. An M1153 mutation causes a defect in receptor internalization relative to normal receptors, and was demonstrated to cause insulin resistance.[17,18] 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–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 1(d)) 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% *versus* 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 (ER), which is a nuclear receptor. Breast cancer patients possessing ER typically have a lower risk of relapse and better overall survival.[19] The variant in our database, a C447A mutant of ER, displayed a dose-response shift for estradiol in transactivation studies.[20] C447 (red) (PDB code 3ert, Figure 1(g)) is located in a tightly packed region of the protein. It has 30 atomic contacts with 12 residues, including two ionizable residues (E443 and E444), two polar residues (S450 and T485), and one aromatic residue (F445). The substitution of a Cys residue by a short Ala residue presumably leads to the loss of many favorable contacts and hence the loss of thermal stability and binding affinity.

Why are disease-associated nsSNPs observed infrequently 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 affect the binding events of the protein directly. In addition, mutations in the protein core

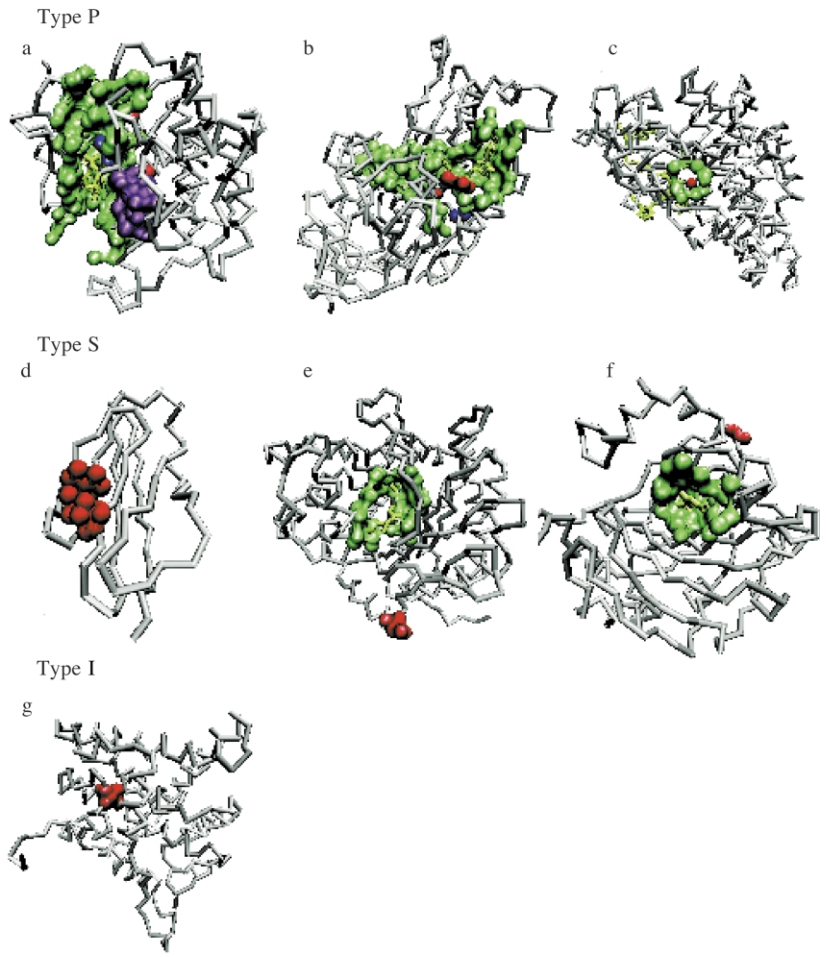

Figure 1. Geometric locations of nsSNPs. About 88% of the structural sites of nsSNPs are located in a pocket or void of the protein (type P), about 8.6% are on the rest of the outer surfaces (type S), which are either convex regions or shallow depressed regions. About 3.2% are fully buried in the interior (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 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.

frequently do not affect the stability of a protein strongly.[21] Theoretical studies of design patterns of protein folding show that proteins with stable structures often are tolerant to mutation.[22] 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 HMMs 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 SNPs into the following five categories:

1. SNP amino acid is the same as HMM consensus amino acid in the aligned position (AP) and the latter is highly conserved;
2. SNP amino acid is the same as HMM consensus amino acid in AP and the latter is not highly conserved;
3. SNP amino acid is different from HMM consensus amino acid in AP that is conserved;

Table 1. Sequence conservation and geometric location of disease-associated nsSNP sites

	Alpha shape predicted				
Category	Pocket	Surface	Interior	Totals	After removing Cat. 5
1. Consensus, conserved	202 (178–226) 22% (19–24%)	23 (14–33) 3% (2–4%)	14 (7–22) 2% (0.8–2.4%)	239 (199–281) 26% (22–30%)	28% (23–32%)
2. Consensus, not conserved	250 (224–276) 27% (24–30%)	26 (17–36) 3% (2–4%)	7 (2–13) 0.76% (0.22–1.4%)	283 (243–325) 31% (26–35%)	33% (28–38%)
3. Not consensus, conserved	261 (234–287) 28% (25–31%)	19 (11–27) 2% (1–3%)	8 (3–14) 0.87% (0.32–1.5%)	288 (248–328) 31% (27–35%)	33% (29–38%)
4. Not consensus, not conserved	48 (35–62) 5% (4–7%)	7 (2–13) 0.76% (0.21–1.2%)	0 (0–1) 0.00% (0–0.1%)	55 (37–82) 6% (4–9%)	6.4% (4.3–9.5%)
5. No HMM alignment	53 (41–68) 6% (4–7%)	5 (1–10) 0.54% (0.1–1%)	1 (0–3) 0.11% (0–0.32%)	59 (42–81) 6% (5–9%)	N/A
Totals	814 (712–925) 88% (77–100%)	80 (45–119) 8.6% (4.9–12.9%)	30 (12–53) 3.2% (1.3–5.7%)	924	865

The 95% confidence interval is given in parentheses for each value.

4. SNP amino acid is different from HMM consensus amino acid in AP that is not conserved;
5. HMM data could not be calculated.

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 be compensated only 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 (Tables 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.

Table 2. Sequence conservation and geometric location of nsSNP sites from dbSNP

	Alpha shape predicted				
Category	Pocket	Surface	Interior	Totals	After removing Cat. 5
1. Consensus, conserved	39 (27–51) 7% (5–9%)	9 (4–15) 2% (7–3%)	1 (0–3) 0.18% (0–0.54%)	49 (31–69) 9% (6–12%)	11% (7–16%)
2. Consensus, not conserved	54 (42–69) 10% (7–12%)	21 (13–30) 4% (2–5%)	3 (0–7) 0.54% (0–1.6%)	78 (55–106) 14% (10–19%)	18% (13–24%)
3. Not consensus, conserved	149 (129–172) 27% (23–30%)	57 (44–72) 10% (8–13%)	11 (5–18) 2% (0.9–3%)	217 (178–262) 39% (32–47%)	50% (41–60%)
4. Not consensus, not conserved	57 (42–71) 10% (8–13%)	33 (22–44) 6% (4–8%)	4 (1–9) 0.72% (0.18–1.4%)	94 (85–124) 17% (15–22%)	21% (19–28%)
5. No HMM alignment	82 (66–98) 15% (12–18%)	30 (21–40) 5% (4–7%)	8 (3–14) 1% (0.54–2.5%)	120 (90–152) 22% (16–27%)	N/A
Totals	381 (306–461) 68% (55–83%)	150 (104–201) 27% (19–36%)	27 (9–51) 4.8% (2–9%)	558	438

The 95% confidence interval is given in parentheses for each value.

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 involved in protein function directly, 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 distributed evenly 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 (categories 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.

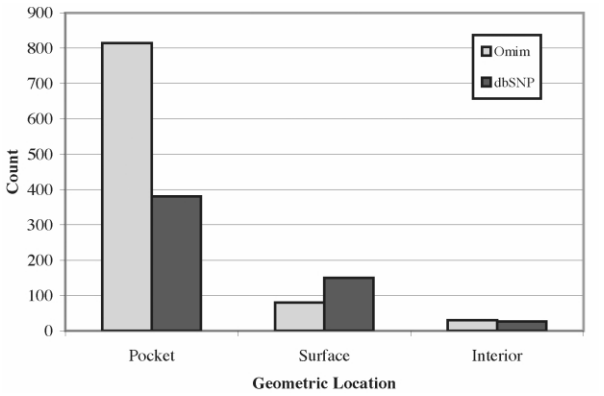

Figure 2. The distribution of disease and non-disease nsSNPs at different geometric locations.

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 buried completely in the interior (Figure 2). 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,[23] where fold information from SCOP and active-site information from SwissProt are 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 amounts 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 occurring more frequently in positions of genes whose wild-type 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 occur frequently 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.[24,25] 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.[26]

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 disease-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,[27] a set of generic structure and sequence-derived features are developed on the basis of 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.[28]

The integrated structural/sequence methodology described in this study can be developed further 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 an 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 combined carefully with the more established statistical analysis of SNPs with respect to their polymorphism across different populations.[29] As SNP projects proceed rapidly and sufficient data has been accumulated, the new structure/sequence methodology described here promises to provide an improved diagnostic capability of SNPs to be disease-associated.

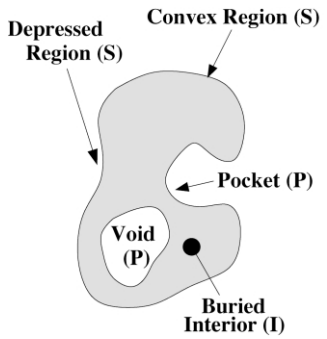

Figure 3. Any amino acid residue of a protein can be classified into three geometric locations: those located in a pocket or a void (type P), those located on a convex region or a shallow depressed region of protein surface (type S), and those that are buried completely in the interior (type I). These locations and their geometric types are illustrated here. Voids are enclosed completely and have no access to the outside bulk. Pockets are connected to the outside by narrow neck(s).

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 (type P); and (3) on the rest of the boundary surface regions, which are either convex regions or shallow depressed regions (type S) (Figure 3). A residue is located in the interior if it is in contact with many other residues and is buried fully, 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 enclosed fully 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.[28,30–36] 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.[28,32,33,37–40] Here, we use alpha shape software Delcx, Mkalf, Volbl (downloadable from NCSA†), and castP‡ 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 OMIM database.§ 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 2128 variants on 310 genes from an original set of 5467 nsSNPs in 1061 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.[41] 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 the OMIM database, where each entry contains annotation of disease phenotype based on experimental data. We exhaustively search dbSNP (release 103) for non-synonymous refSNPs with location information.‖

The full sequences of the corresponding genes are then extracted from the GenBank database. These sequences are used to search the PDB database to extract corresponding structural information. BLASTP is used for this task with the default settings. PDB structures that match nsSNP containing genes with E-value $< 10^{-50}$ are considered exact matches, and we select only the highest-scoring alignment for further consideration. From an original dataset of 9076 variants on 5049 genes, we are able to extract structural information for 504 of genes, which represent 973 variants.

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 semi-global 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 residues missing from 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).[36] Among the 924 structural sites where the 2128 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 the Supplementary Material. 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.[28]¶

Calculating conservation at variant sites

To perform this classification, we use probabilistic models of protein family derived from HMMs, which have become a valuable method with wide applications in protein modeling, homology detection, and functional classification of putative genes. First introduced for protein modeling by Krogh *et al.*[42] and Delcher *et al.*,[43] HMMs can be viewed as special cases in the more general framework of probabilistic Bayesian networks.

† http://www.ncsa.uiuc.edu
‡ http://cast.engr.uic.edu
§ http://www.ncbi.nlm.nih.gov/omim/
‖ ftp://ftp.ncbi.nih.gov/snp/human/

¶ These data will be made available at http://gila.bioengr.uic.edu/snp

In a typical HMM 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 20 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 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 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. Introductory overviews of HMMs can be found elsewhere.[44,45]

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.[23] 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 has been described by Bateman *et al*.[46,47] The analysis software used for our SNP analysis can be obtained by sending e-mail to kasif@bu.edu

In our study, each entry in both OMIM- and dbSNP-derived nsSNP databases contains 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 HMMs 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 the position corresponding to the nsSNP. The consensus residue is that with the 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, on the basis of the degree of conservation and whether the SNP residue corresponds to the consensus amino acid in the alignment, we can define the multiple classification categories given above.

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.[48,49] Let the true value of the distribution be θ. Our estimator T takes the value t, which is the estimated value for θ. Our goal is to calculate a $(1 - 2\alpha)$ confidence interval for θ. 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 θ is to use the bootstrap estimates of quantiles for $T - \theta$. The definition of probability (Pr) implies:

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

For an equitailed $(1 - 2\alpha)$ confidence interval, we have:

$$\Pr(T - b \leq \theta \leq T - a) = 1 - 2\alpha$$

and the following basic bootstrap confidence limits:

$$t - (t^*_{((R+1)(1-\alpha))} - t),\ t - (t^*_{((R+1)\alpha)} - t)$$

In our calculation, R is chosen as 10,000, and α is 2.5 % for estimation of 95% confidence intervals.

Acknowledgements

This work was supported by funding from the National Science Foundation (CAREER DBI0133856, DBI0078270, and MCB998008).

References

1. Collins, F. S., Brooks, L. D. & Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**, 1229–1231.
2. Lander, E. S. (1996). The new genomics: global views of biology. *Science*, **274**, 536–539.
3. Dryja, T. P., McGee, T. L., Hahn, L. B., Cowley, G. S., Olsson, J. E., Reichel, E. *et al.* (1990). Mutations within the rhodopsin gene in patients with autosomal dominant retinitis pigmentosa. *N. Engl. J. Med.* **323**, 1302–1307.
4. Smith, E. P., Boyd, J., Frank, G. R., Takahashi, H., Cohen, R. M., Specker, B. *et al.* (1994). Estrogen resistance caused by a mutation in the estrogen-receptor gene in a man. *N. Engl. J. Med.* **331**, 1056–1061.
5. Barroso, I., Gurnell, M., Crowley, V. E., Agostini, M., Schwabe, J. W., Soos, M. A. *et al.* (1999). Dominant negative mutations in human PPARgamma associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature*, **402**, 880–883.
6. Thomas, R., McConnell, R., Whittacker, J., Kirkpatrick, P., Bradley, J. & Sandford, R. (1999). Identification of mutations in the repeated part of the autosomal dominant polycystic kidney disease type 1 gene PKD1, by long-range PCR. *Am. J. Hum. Genet.* **65**, 39–49.
7. Bonnardeaux, A., Davies, E., Jeunemaitre, X., Fery, I., Charru, A., Clauser, E. *et al.* (1994). Angiotensin II type 1 receptor gene polymorphisms in human essential hypertension. *Hypertension*, **24**, 63–69.
8. Vatsis, K. P., Martell, K. J. & Weber, W. W. (1991). Diverse point mutations in the human gene for polymorphic *N*-acetyltransferase. *Proc. Natl Acad. Sci. USA*, **88**, 6333–6337.
9. Wang, Q., Curran, M. E., Splawski, I., Burn, T. C., Millholland, J. M., VanRaay, T. J. *et al.* (1996). Positional cloning of a novel potassium channel

gene: KVLQT1 mutations cause cardiac arrhythmias. *Nature Genet.* **12**, 17–23.
10. Hassett, C., Aicher, L., Sidhu, J. S. & Omiecinski, C. J. (1994). Human microsomal epoxide hydrolase: genetic polymorphism and functional expression *in vitro* of amino acid variants. *Hum. Mol. Genet.* **3**, 421–428.
11. Yoshida, A., Huang, I. Y. & Ikawa, M. (1984). Molecular abnormality of an inactive aldehyde dehydrogenase variant commonly found in Orientals. *Proc. Natl Acad. Sci. USA*, **81**, 258–261.
12. Jaruzelska, J., Abadie, V., d'Aubenton-Carafa, Y., Brody, E., Munnich, A. & Marie, J. (1995). *In vitro* splicing deficiency induced by a C to T mutation at position-3 in the intron 10 acceptor site of the phenylalanine hydroxylase gene in a patient with phenylketonuria. *J. Biol. Chem.* **270**, 20370–20375.
13. Proia, R. L. & Neufeld, E. F. (1982). Synthesis of beta-hexosaminidase in cell-free translation and in intact fibroblasts: an insoluble precursor alpha chain in a rare form of Tay-Sachs disease. *Proc. Natl Acad. Sci. USA*, **79**, 6360–6364.
14. Stryer, L. (1995). *Biochemistry*, 4th edit., W. H. Freeman, New York.
15. Sunyaev, S., Ramensky, V. & Bork, P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.* **16**, 198–200.
16. Moller, D. E., Yokota, A., Ginsberg-Fellner, F. & Flier, J. S. (1990). Functional properties of a naturally occurring Trp1200–Ser1200 mutation of the insulin receptor. *Mol. Endocrinol.* **4**, 1183–1191.
17. Cama, A., de la Luz Sierra, M., Ottini, L., Kadowaki, T., Gorden, P., Imperato-McGinley, J. & Taylor, S. I. (1991). A mutation in the tyrosine kinase domain of the insulin receptor associated with insulin resistance in an obese woman. *J. Clin. Endocrinol. Metab.* **73**, 894–901.
18. Cama, A., de la Luz Sierra, M., Quon, M. J., Ottini, L., Gorden, P. & Taylor, S. I. (1993). Substitution of glutamic acid for alanine 1135 in the putative catalytic loop of the tyrosine kinase domain of the human insulin receptor. A mutation that impairs proteolytic processing into subunits and inhibits receptor tyrosine kinase activity. *J. Biol. Chem.* **268**, 8060–8069.
19. Clark, G. M. & McGuire, W. L. (1988). Steroid receptors and other prognostic factors in primary breast cancer. *Semin. Oncol.* **15**, 20–25.
20. Reese, J. C. & Katzenellenbogen, B. S. (1992). Characterization of a temperature-sensitive mutation in the hormone binding domain of the human estrogen receptor. Studies in cell extracts and intact cells and their implications for hormone-dependent transcriptional activation. *J. Biol. Chem.* **267**, 9868–9873.
21. Axe, D. D., Foster, N. W. & Fersht, A. R. (1996). Active barnase variants with completely random hydrophobic cores. *Proc. Natl Acad. Sci. USA*, **93**, 5590–5594.
22. Mélin, R., Li, H., Wingreen, N. & Tang, C. (1999). Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study. *J. Chem. Phys.* **110**, 1252–1262.
23. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R. *et al.* (2002). The Pfam protein families database. *Nucl. Acids Res.* **30**, 276–280.
24. Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1997). Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* **6**, 53–64.
25. Tsai, C. J. & Nussinov, R. (1997). Hydrophobic folding units at protein–protein interfaces: implications to protein folding and to protein–protein association. *Protein Sci.* **6**, 1426–1437.
26. Hu, Z., Ma, B., Wolfson, H. & Nussinov, R. (2000). Conservation of polar residues as hot spots at protein interfaces. *Proteins: Struct. Funct. Genet.* **39**, 331–342.
27. Chasman, D. & Adams, R. M. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* **307**, 683–706.
28. Liang, J., Edelsbrunner, H. & Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **7**, 1884–1897.
29. Nelson, M. R., Kardia, S. L., Ferrell, R. E. & Sing, C. F. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* **11**, 458–470.
30. Liang, J. & Subramaniam, S. (1997). Computation of molecular electrostatics with boundary element methods. *Biophys. J.* **73**, 1830–1841.
31. Kim, S., Liang, J. & Barry, B. A. (1997). Chemical complementation identifies a proton acceptor for redox-active tyrosine D in photosystem II. *Proc. Natl Acad. Sci. USA*, **94**, 14406–14411.
32. Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V. & Subramaniam, S. (1998). Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. *Proteins: Struct. Funct. Genet.* **33**, 18–29.
33. Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V. & Subramaniam, S. (1998). Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. *Proteins: Struct. Funct. Genet.* **33**, 1–17.
34. Adamian, L. & Liang, J. (2001). Helix–helix packing and interfacial pairwise interactions of residues in membrane proteins. *J. Mol. Biol.* **311**, 891–907.
35. Adamian, L. & Liang, J. (2002). Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. *Proteins: Struct. Funct. Genet.* **47**, 209–218.
36. Liang, J. & Dill, K. A. (2001). Are proteins well-packed? *Biophys. J.* **81**, 751–766.
37. Edelsbrunner, H. & Mücke, E. (1994). Three-dimensional alpha shapes. *ACM Trans. Graph.* **13**, 43–72.
38. Edelsbrunner, H., Facello, M. & Liang, J. (1998). On the definition and the construction of pockets in macromolecules. *Disc. Appl. Math.* **88**, 83–102.
39. Edelsbrunner, H., Facello, M., Fu, P. & Liang, J. (1995). Measuring proteins and voids in proteins. *Proc. 28th Annu. Hawaii Intl. Conf. Syst. Sci.* **5**, 256–264.
40. Facello, M. (1995). Implementation of a randomized algorithm for Delaunay and regular triangulations in three dimensions. *Comput. Aided Genome Des.* **12**, 349–370.
41. Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. (2000). dbSNP: a database of single nucleotide polymorphisms. *Nucl. Acids Res.* **28**, 352–355.
42. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531.
43. Delcher, A. L., Kasif, S., Goldberg, H. R. & Hsu, W. H. (1993). Protein secondary structure modeling with

probabilistic networks. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1**, 109–117.
44. Salzberg, S. L., Searls, D. B. & Kasif, S. (1998). *Computational methods in molecular biology.* New comprehensive biochemistry, vol. 32, Elsevier, Amsterdam.
45. Durbin, R. (1998). *Biological sequence analysis: probalistic models of proteins and nucleic acids*, Cambridge University Press, Cambridge, UK.
46. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 263–266.
47. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, E. L. (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucl. Acids Res.* **27**, 260–262.
48. Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap methods and their application*, Cambridge University Press, Cambridge.
49. Efron, B. & Tibshirani, R. (1993). *Monographs on statistics and applied probability* An Introduction to the Bootstrap, vol. 57, Chapman & Hall, New York.

Edited by B. Honig

(Received 15 November 2002; received in revised form 6 February 2003; accepted 6 February 2003)

Supplementary Material comprising two Tables is available on Science Direct