**Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis**

**Highlights**
- A comparative structural analysis of disease-associated and neutral mutations
- Assessment of structural regions most vulnerable to disease mutations
- Variations adjacent to protein-protein interfaces are strongly disease associated
- Compositional analysis in comparison with mutations from the 1K Genomes Project

**Authors**
Mu Gao, Hongyi Zhou, Jeffrey Skolnick

**Correspondence**
skolnick@gatech.edu

**In Brief**
Atomic structures and functional analysis of proteins can provide invaluable insights into why certain variations are disease associated. Using experimental structural data and computational predictions of functional effects, Gao et al. provide a comprehensive analysis and structural rationalization of thousands of disease-associated and neutral variations.

Gao et al., 2015, Structure 23, 1362–1369
July 7, 2015 ©2015 Elsevier Ltd All rights reserved
http://dx.doi.org/10.1016/j.str.2015.03.028
Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis

Mu Gao,1 Hongyi Zhou,1 and Jeffrey Skolnick1,*
1Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, 250 14th Street NW, Atlanta, GA 30318, USA
*Correspondence: skolnick@gatech.edu

SUMMARY

Most known disease-associated mutations are missense mutations involving changes of amino acids of proteins encoded by their genes. Given the plethora of genetic studies, sequenced exomes, and new protein structures determined each year, it is appropriate to revisit the role that structure plays in providing insights into the molecular basis of disease-associated mutations. In that regard, a large-scale structural analysis of 6,025 disease-associated mutations as well as 4,536 neutral variations for comparison was performed. While buried amino acids are common among the disease-associated mutations, as reported previously, more are statistically significantly enriched at observed or predicted functional sites. Interesting findings are that ligand-binding sites adjacent to protein-protein interfaces and residues involved in enzymatic function are especially vulnerable to disease-associated mutations. Finally, a compositional analysis of disease-associated mutations in comparison with variants identified in the 1000 Genomes Project provides a structural rationalization of the most disease-associated residue types.

INTRODUCTION

Since the sequencing of the first human genome was completed a decade ago (Collins et al., 2004), tremendous efforts have been made to advance new sequencing methods that allow rapid, massively parallel sequencing at low cost (Metzker, 2010). These next-generation sequencing methods enable large-scale sequencing on thousands of individuals, generating a large amount of data available for comparative genome analysis (Abecasis et al., 2010, 2012). By identifying variations of DNA sequences, in particular, Single Nucleotide Polymorphisms (SNPs), we may begin to decipher the links between phenotypes and genotypes. Of particular interest are genetic mutations that cause various human, especially Mendelian, diseases. Statistical analyses of patients’ and normal people’s sequences often pinpoint mutations strongly associated with patients. Many such mutations are collected in databases such as the Online Database of Mendelian Inheritance in Man (OMIM) (McKusick, 2007) and the Human Gene Mutation Database (HGMD) (Stenson et al., 2008).

Since most cases in these databases are SNPs identified through statistical analysis, it is not clear whether a particular mutation is actually the cause of the implicated disease. These mutations are usually referred to as being disease associated. Most are non-synonymous SNPs (nsSNPs), which occur in the coding regions of genes and result in changes of amino acid type, i.e., missense mutations. It is therefore expected that the change of amino acid impairs the function of the involved protein. However, for the vast majority of cases, it is not clear how the mutation affects the function of the protein. In this regard, studying the impact of mutations may provide a better understanding of the mechanisms of the corresponding diseases and eventually may significantly increase the chance of finding a better treatment for patients with the disease.

Meanwhile, over a comparable period, the three-dimensional structures of many proteins have been determined at high resolution (Berman et al., 2000). Often, these proteins are co-crystallized with other biomolecules that are relevant to their functions, e.g., protein-protein complexes and protein-ligand complexes. Given these structural data, many questions regarding disease-causing mutations can be asked and addressed by a thorough inspection of these structures: can we understand disease-associated mutations in the context of their locations in the structure of the protein? Where are disease-linked missense mutations located in the proteins? What are the functional and structural consequences of these mutations? Is there any location in the protein where variations are more likely to cause disease? What is the structural reason why certain types of mutations are more likely to be disease associated? Answers to these questions not only deepen our understanding of the molecular mechanisms of diseases but also have practical implications for the predictions of disease association by automated computational tools and for personalized medicine.

Over a decade ago, Wang and Moult (2001) performed an early study of 262 disease-associated mutations from 26 proteins and found that about 80% of them destabilize proteins and about 5% involve ligand binding. A subsequent study by Steward et al. (2003) of 1,292 mutations from 63 proteins showed that about 28% are buried and related to protein stability, while 29% are involved in intermolecular interactions such as protein-protein interactions and ligand binding. Similar results were also reported in another study (Sunyaev et al., 2000). However, mainly due to the paucity of solved protein structures, these studies were carried out on very few proteins. Since then, many new
disease-causing mutations have been found, and the number of known protein structures has also grown exponentially, especially for those in complex with functional partners. The previous studies also lacked a comparison with neutral variations. The statistical significance of their discoveries, in particular, the value of structural analysis to the prediction of disease association, is unclear.

Given this background, it is worthwhile revisiting the questions raised above. In this study, we performed a large-scale analysis of 6,025 disease-associated mutations in 642 proteins with experimental structures, as well as 4,536 neutral mutations in 1,743 proteins for comparison. We took advantage of the fact that many of these proteins have multiple structures in complex with other proteins or ligands. Below, we first map these mutations to their locations onto protein structures. We then compare them with neutral variations and identify regions more likely to cause disease. We also analyze the likelihood that different amino acid types are disease associated and provide structural explanations for some of the most deleterious mutations.

RESULTS

We have collected 6,025 disease-associated missense mutations in 642 human genes and 4,536 neutral mutations in 1,743 human genes from the August, 2013 release of the UniProt (Universal Protein Resource) database (see Experimental Procedures) (Wu et al., 2006). All these mutations have at least one experimentally determined protein structure containing the corresponding mutation. The locations of these mutations are subsequently analyzed.

Where Are the Disease-Associated Missense Mutations in Their Protein Structures?

Table 1 shows the statistics of disease-associated mutations. About 22% are buried in protein cores; 12% are found at protein–protein interfaces (PPIs); 36% are located in concave pockets; 12% and 2.9% are found in direct physical contact with small-molecule ligands and metal ions, respectively; and a small fraction, 0.7%, are at either DNA-protein or RNA-protein interfaces. Since the experimental information is most likely incomplete, we further applied two computational tools, EFICAz for enzyme function prediction (Kumar and Skolnick, 2012) and FINDSITEcomb for ligand-binding site predictions, to our datasets (Zhou and Skolnick, 2013). About 7% of disease-associated mutations are classified by EFICAz as being functional discriminating residues (FDRs) for their corresponding enzymatic functions, and an additional 12% of mutations are at FINDSITEcomb predicted ligand-binding sites that did not have bound ligands in their PDB structures. It should be noted that these location classifications are not necessarily mutually exclusive, e.g., about 69% and 45% of observed small-molecule ligand-binding sites are found in geometric pockets or predicted by FINDSITEcomb, respectively. Another interesting observation is that about 15% of disease-associated mutations of observed ligand-binding sites are also identified at the PPIs. This is investigated in detail below.

Figure 1 shows a pie chart obtained by assigning a unique primary location to each mutation, using the following order: DNA/RNA binding, ion binding, small-molecule ligand binding, PPI, buried, EFICAz, FINDSITEcomb, pocket, and other exposed. About 15% of disease-associated mutations are located at known functional sites, i.e., they involve binding to other molecules than proteins; 10% are at PPIs; and 20% of mutations are within protein cores and are likely important for maintaining stability. EFICAz and FINDSITEcomb in addition flagged 3% and 5% of mutations that are potentially important for the functions of the corresponding protein. After these classifications, 17% of mutations are found in pockets, which might engage in unknown interactions with some biomolecules. Overall, this classification scheme could assign a primary functional or structural role for 70% of disease-associated mutations. The remaining 30% of mutations involve exposed surface residues and are located at either a flat surface or small pocket. They are potentially candidate interaction sites for unknown protein–protein interactions or interactions with biomolecules that do not require large concave pockets.

How Useful Is Structural/Functional Information for Predicting Disease Association?

In order to answer this question, we compared the frequency of disease-associated mutations versus neutral mutations at different locations derived from our structural and functional analysis mentioned above. As shown in Figure 2A, in all locations except for unannotated, exposed regions, disease mutations are more likely to appear than neutral mutations. In the protein interior, the frequency of disease mutations is over two times that of neutral mutations. This gives an odds ratio (OR) of 2.66, which is the odds of disease versus neutral among buried mutations over the odds of disease versus neutral among all other mutations in our sets. The result is statistically highly significant, with a p value 4.0 × 10⁻⁶⁷ (Fisher’s exact test, two-tailed). Likewise, ligand-binding sites are about 50% more likely to be observed in the disease set than in the neutral set, which gives a significant OR of 1.53 (p = 6.8 × 10⁻¹²). At an OR of 1.75 (p = 4.4 × 10⁻²¹), the performance of FINDSITEcomb in terms of OR is slightly better than that obtained by counting observed ligand-binding sites in experimental structures. Within a PPI, or pocket region, disease mutations are slightly, but significantly, more frequent than neutral mutations, yielding ORs of 1.23 (p = 1.3 × 10⁻⁹) and 1.09 (p = 0.04), respectively. The latter is less significant, partly due to the fact that not all
residues identified in a predicted, geometric protein pocket are important for protein function or stability; thus, there are more noisy signals in this region. Surprisingly, EFICAz predictions have the highest OR, of about 17 ($p = 1.6 \times 10^{-78}$). Many of these predictions are at highly conserved active sites important for enzymatic function, although they cover only about 7% of disease mutations. On the other hand, among unannotated surface residues, neutral mutations have a high percentage at 45%, compared with about 30% of disease mutations. This is expected, as surface residues without a functional or structural role are less likely to cause disease.

Since neutral mutations are from a much more diverse set of proteins than disease-associated mutations (1,743 versus 642 proteins), there is the concern that the dataset might be biased. In order to eliminate such a concern, we further performed the same analysis on the subset of mutations that appear in the same set of proteins. Figure 2B shows results that qualitatively are in agreement with the analysis on the full set shown above.

Mutations Adjacent to PPIs Are More Likely Associated with Disease Than Mutations within PPIs

Overall, about 58% of disease-associated mutations are found in at least one protein-protein complex structure versus 50% of neutral mutations that are found in a protein-protein complex. We further analyzed the distribution of mutations in protein-protein complexes according to their distance from observed PPIs (Figure 3A). As shown above, about 12.0% of all disease mutations and 10.1% of neutral mutations in our datasets are located at the PPIs, i.e., 0 Å from the interface. This gives a small, but statistically significant, difference of 1.9%, corresponding to an increase of about 20% in OR to 1.23 for the disease-associated mutations. The remaining difference in cumulative fraction comes mainly from the mutations close to but not at the PPIs. Within 3 Å from PPIs, the cumulative difference
rapidly increases to 6.8%; this percentage further increases to 8.3% within 6 Å from PPIs. Thus, it appears that disease mutations are more likely to occur at and/or adjacent to PPIs than neutral mutations. Moreover, the biggest difference is attributed to regions immediately neighboring PPIs, i.e., at a distance from 0 Å to 6 Å, rather than at the PPIs themselves. This difference leads to a significant increase of OR values to as high as 1.75 at 3 Å from PPIs (Figure 3C).

Since disease-causing mutations are more likely to be buried than neutral mutations, as shown in Figure 2, one natural explanation is that buried mutations close to PPIs are more likely disease associated than neutral. The difference in cumulative fraction of buried mutations around PPIs between the disease and neutral sets is about 0.1% at 0 Å, 3.5% within 3 Å, and 5.1% within 6 Å from PPIs. Therefore, mutations of buried residues do make a major contribution to the observed phenomenon. However, this explains only about 61% of the observed difference.

A second main reason for this phenomenon is that adjacent to PPIs, there are functionally relevant pockets involving ligand recognition. It has been shown that protein-protein association naturally creates geometric pockets (Gao and Skolnick, 2012). Some of these pockets could be selected by evolution for ligand binding. As a result, functionally relevant ligand-binding pockets are enriched around PPIs. Mutations at these pockets likely lead to harmful effects. Indeed, if we focus on the subset of mutations that are also ligand binding within protein-protein complexes, we find that disease-associated mutations are much more likely to be found in the neighborhood of PPIs versus neutral mutations. As shown in Figure 3B, the difference in the cumulative fraction of mutations is 0.9% at 0 Å from PPIs, 2.0% within 3 Å from PPIs, and 2.6% within 6 Å from PPIs, which could explain about 31% of the corresponding 8.3% difference in overall cumulative fraction difference mentioned above. This also contributes to an overall bump in the OR shown in Figure 3C. Notice that the risk of being disease associated is about 30% higher in terms of OR when a mutation is observed in a ligand-binding site within 3 Å from a PPI.

Figure 4 shows two examples of disease-associated mutations located at/near PPIs. Medium-chain acyl-coenzyme A (CoA) dehydrogenase (ACADM) is one of four enzymes involved in fatty acid catabolism (Lee et al., 1996). The enzyme is responsible for the α,β-dehydrogenation of fatty acyl-CoA derivatives. The structure of this protein is a dimer of dimers. Figure 4A shows the structure of the basic dimer unit, in which the co-factor FAD (flavin-adenine dinucleotide) is sitting at the dimeric interface, whereas the fatty substrate is located close by. There are five known mutations located in the binding sites of these substrates within 5 Å. They are linked to ACADMD (ACADM medium-chain deficiency), a disease that can cause sudden death of infants (OMIM: 201450).

One important benefit of an interfacial pocket formed adjacent to PPIs is that it could function as a molecular switch controlled by the association/dissociation of the protein complex. This is illustrated in an example involving two proteins, retinitis pigmentosa 2 (RP2) and the small G protein Arf-like 3 (Arl3), which form a complex providing a guanosine triphosphate binding pocket at the PPI (Figure 4B) (Veltel et al., 2008). It is thought that this is a molecular switch for regulating the ciliary process in photoreceptor cells. Mutations in RP2 that disrupt ligand binding at the PPI, such as the arginine finger R118, may cause X chromosome-linked eye disease (OMIM: 312600).

### Post-Translational Modifications and Disease Association

Post-translational modifications play essential roles in many biological processes. We examined how many disease-associated mutations are located at such a site compared with neutral mutations. From the UniProt knowledge base, we searched for experimentally validated modification sites. We found only 27 cases in the disease mutation set and 11 cases in the neutral set, which gives a p value of 0.1 (Fisher’s exact test, two-tailed). The most common modifications are glycosylation: nine in the disease mutations and four in the neutral mutations. If we drop the requirement of having solved structures, we are able to match 77 of 23,846 disease mutations and 68 of 37,687 neutral mutations, which leads to a significant p value of $5.9 \times 10^{-4}$. As one would expect, mutations at post-translational
for the purpose of our analysis. We repeated the same composi-
could argue that these variants are a good neutral background
(Cooper and Youssoufian, 1988). However, because the fre-
CpG dinucleotide, which is relatively vulnerable to mutations
due to the fact that four of the six arginine codons contain the
is the most common mutation in both disease and neutral sets,
the original reference amino acid at the mutation sites. Arginine

First, we ask why mutations from and to a Cys residue are
harmful. Since it is known that Cys often form disulfide bonds, it is
hypothesized that a Cys mutation either disrupts useful disul-
fide bonds in the case of mutation from or introduces an
unwanted disulfide bond in case of mutation to. To test this hy-
pothesis, we examined if there are Cys disulfide bonds in the
structure of the corresponding protein. In the mutation-from
set, about 2/3 of Cys mutations are likely to form a disulfide
bond with another Cys residue nearby within 4.5 Å, whereas
only about 1/3 of Cys mutations in the neutral set have another
Cys nearby. This gives a highly significant p value of 3.2 × 10⁻⁶.
Therefore, disease-associated Cys mutations are more
likely to disrupt an original disulfide bond important for protein
stability or function. On the other hand, in the mutation-to data-
sets, we found that after the mutation to Cys, there are 45 cases
in which there is another Cys nearby in the disease set, versus
only seven cases in the neutral set (p = 0.007). In these cases,
the Cys mutation may introduce an unwanted disulfide bond
that could lead to protein malfunction.

Second, why is a mutation to Pro likely to be harmful? Pro is a
unique amino acid that does not have an hydrogen for backbone
hydrogen bond formation. In addition, it has a more restricted
dihedral angle space than a typical amino acid. This restriction
often creates a kink in helical structures at the position of a
Pro. As a result, mutations to Pro often disrupt helical structures.
We performed an analysis of the secondary structure where a
mutation to Pro occurs, with the result shown in Table 2. The
most common Pro mutation in the disease set lies within a 3₁₀
helix, about 36%, in contrast to 23% cases in the neutral set
(p = 0.006). The second common secondary structural element
where a Pro mutation occurs is a turn: 26% in disease versus
18% in neutral (p = 0.1) sites, where a Pro mutation might disrupt
disulfide bond formation. By contrast, in the coil region, where there is
no ordered secondary structure, a Pro mutation is much less
likely present at 20%, in comparison with 32% in the neutral
set (p = 0.009). The result supports the conclusion that the
disruptive presence of a Pro mutation in the secondary structure
is often disease associated.

Lastly, we investigated mutation from Trp that is often linked to
disease. Trp is the largest amino acid that plays an important role
in protein folding and stability. On average, a Trp residue makes
8.5 side-chain contacts versus 6.0 side-chain contacts of other
types of amino acids. Further calculation estimates that a Trp
mutation in the disease set leads to a mean increase of the
free energy of protein folding ΔΔG of 3.4 kcal/mol (126 cases,
SD = 1.61 kcal/mol), versus 2.9 kcal/mol (27 cases, SD = 1.36
cal/mol) of Trp mutations in the neutral set (p = 0.076, t test, two-tailed).
This analysis suggests that mutations from
Trp likely destabilize the structure of the protein and result in
less fitness. In the vast majority of cases analyzed, this

Figure 4. Examples of Disease-Associated Mutations Observed in
the Neighborhood of PPIs
(A) Dimeric ACADM.
(B) Retinitis pigmentosa 2 and Arl3. In each snapshot, protein structures are
shown in light blue and purple cartoon representations, small molecules are
shown in orange and yellow van der Waals (VdW) representations, and refer-
ence amino acids at the mutation sites are also shown in VdW representations,
where carbon, oxygen, and nitrogen atoms are colored in cyan, red, and blue,
respectively.
destabilization likely has a significant impact on the function of protein, thus leading to various disease conditions.

DISCUSSION

Since the vast majority of disease-associated mutations are assigned based on statistical analysis, it is not clear whether a mutation is actually the cause of the implicated disease. To provide a better understanding of mutations from a structural prospective, we performed a large-scale analysis of 6,025 disease-associated mutations in 642 proteins. We found that about 20% of mutations are buried in protein cores and that these mutations may destabilize protein structures, whereas about 11% of mutations are involved in ligand binding, and another 10% are involved in protein-protein interactions. About 17% are in pockets, 3% involve ion binding, and 1% DNA/RNA binding. These numbers are consistent with previous studies conducted on smaller datasets (Steward et al., 2003; Sunyaev et al., 2000). Using computational approaches, we further predicted that about another 8% of mutations are likely ligand-binding or essential for enzyme function. Buried residues in the disease set have the second highest OR, at 2.66, compared with neutral ones, whereas exposed residues not involving any interactions or pocket-like feature are more likely to be neutral at an OR of 0.50. Functional sites, such as ligand-binding sites and PPI residues, all have significant ORs at 1.53 and 1.23. However, they are not as effective as one naively expects. One reason might be due to purification selections, which removed many deleterious, fatal mutations at these functional sites. It is also worth mentioning that computational predictions achieved good performance. Predictions by FINDSITEcomb and EFICAz yield OR values at 1.75 and 16.9, respectively, which is comparable with or better than results obtained by counting known ligand-binding sites from experimental structures. Most interestingly, it is the predicted loss of enzymatic function that is most strongly disease associated. These are encouraging results, which demonstrate the potential of computational, knowledge-based methods.

One novel observation is that mutations adjacent to PPIs are more likely associated with disease than mutations at the PPIs themselves. At 1.75, the OR is highest about 3 Å from PPIs, which is about 40% higher than the OR at PPIs. There are two contributing factors. One is that the mutations at buried sites within each monomer but close to the PPIs are more likely to be disease causing, presumably by destabilizing the protein complex. The second is due to the presence of interfacial pockets, which are ligand-binding pockets at or adjacent to PPIs (Gao and Skolnick, 2012). Mutations within interfacial pockets could disrupt functionally important ligand-protein interactions. This is supported by subsequent analysis, which shows that mutations within interfacial pockets have more than double the OR of mutations at PPIs.

Although amino acid type change is one main factor for assessing pathogenicity of human mutations in many automatic
associated mutations (de Beer et al., 2013; Vitkup et al., 2003). Obviously, if the mutation involves very different amino acid types, e.g., from hydrophobic to charged residues, it is more likely to cause an issue than mutations among hydrophobic residues. Ideally, one would like to study all 380 pairs of non-synonymous mutations. However, the number of known disease-causing mutations is currently too limited for such a study. We opt instead to focus on each type of amino acid, grouped by mutations from and mutation to a given amino acid type. We found that certain types of mutations, such as mutations involving Cys, Trp, or Pro, are more likely to be disease associated. Through structural analysis, Cys mutations often involve breaking disulfide bridges or forming unwanted disulfide bonds; Trp mutations usually significantly destabilize structures, and mutations to Pro tend to disrupt helical structures. These analyses provide some examples of how structural analysis could provide further insights into the mechanisms of disease.

EXPERIMENTAL PROCEDURES

Dataset
From the August, 2013 release of the UniProt knowledge base (Wu et al., 2006), we collected 23,846 disease-associated variants and 37,782 polymorphisms in human; we assume that the latter are neutral mutations. The classification was manually assigned according to the literature and a previously curated database such as OMIM (McKusick, 2007). These variants are all missense mutations with changes of amino acid type. Unclassified variants were ignored. From this collection, we further selected those with at least one experimentally determined structure deposited in the PDB (Berman et al., 2000). We verified that the protein structure contains the same amino acid at the position indicated in the UniProt mutation data. To avoid possible bias, if the gene contains more than 50 variants, we randomly selected no more than 50 variants in each gene. If multiple variants are found corresponding to the same residue position in a gene, we also randomly selected only one variant. This procedure yielded 6,025 disease-associated mutations in 642 proteins and 4,536 neutral mutations in 1,743 proteins. These two sets are the main dataset used in our study. They share 355 common proteins, with more than 50 variants in each gene. If multiple variants are found corresponding to the same residue position in a gene, we also randomly selected only one variant. This procedure yielded 6,025 disease-associated mutations in 642 proteins and 4,536 neutral mutations in 1,743 proteins. These two sets are the main dataset used in our study. They share 355 common proteins, with more than 50 disease mutations and 1,113 neutral mutations.

Data of the 1000 Genomes Project variants are from de Beer et al. (2013). It should be noted that the counts of mutations from OMIM are mislabeled in that study. The counts of mutations from were mistaken as mutation to, and vice versa. The mistake in de Beer et al. (2013) gives incorrect amino acid mutation frequencies, which are inconsistent with a previous study (Vitkup et al., 2003), whereas our results are in agreement with Vitkup et al. (2003).

Table 2. Analysis of Secondary Structure Location of Mutations to Proline

<table>
<thead>
<tr>
<th>Type</th>
<th>Disease</th>
<th>Neutral</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$3_{10}$ helix</td>
<td>195 (0.36)</td>
<td>29 (0.23)</td>
<td>0.003</td>
</tr>
<tr>
<td>$\alpha$ helix</td>
<td>21 (0.039)</td>
<td>7 (0.056)</td>
<td>–</td>
</tr>
<tr>
<td>Turn</td>
<td>139 (0.26)</td>
<td>23 (0.18)</td>
<td>0.049</td>
</tr>
<tr>
<td>$\beta$ sheet</td>
<td>38 (0.071)</td>
<td>11 (0.088)</td>
<td>–</td>
</tr>
<tr>
<td>Bend</td>
<td>35 (0.065)</td>
<td>15 (0.12)</td>
<td>–</td>
</tr>
<tr>
<td>Coil</td>
<td>110 (0.20)</td>
<td>40 (0.32)</td>
<td>0.005</td>
</tr>
<tr>
<td>Total</td>
<td>538</td>
<td>125</td>
<td></td>
</tr>
</tbody>
</table>

*Insignificant P value is indicated by –.*

Table 3. Contingency Table for the Statistical Test

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>Neutral</th>
<th>Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>$d_{pos}$</td>
<td>$n_{pos}$</td>
<td>$d_{pos}/n_{pos}$</td>
</tr>
<tr>
<td>Negative</td>
<td>$d_{neg}$</td>
<td>$n_{neg}$</td>
<td>$d_{neg}/n_{neg}$</td>
</tr>
<tr>
<td>Total</td>
<td>6,025</td>
<td>4,325</td>
<td>–</td>
</tr>
</tbody>
</table>

Structural and Functional Analysis
Using the experimental structures, we mapped the location of each mutation. For each PDB record, we analyzed the structure given by the basic asymmetric unit, as well as author assigned biounit if it contains protein-protein complexes. The HET code in each PDB header was used to determine the types of ligand, i.e., metal ion, small-molecule ligand, or DNA/RNA. Note that this analysis does not guarantee that a ligand has biologically relevant interaction with its co-crystallized protein. Nevertheless, in many cases, biologically relevant molecules recognize the same binding sites on proteins as other non-biological molecules. Atomic contacts between protein and ligand were determined by the program LPC (Sobolev et al., 1999). If the original amino acid at the mutation site contains at least one heavy atom making physical contact with a ligand, it is assigned as ligand binding. Similarly, PPI residues are defined by a heavy-atom distance of 4.5 Å. The distance between a variant residue and PPI is defined as the shortest heavy-atom distance between the residue and any PPI residue. If a protein has multiple complex structures, we chose the minimal distance among all these complexes. Solvent accessible surface area (SASA) was calculated for each original residue of the mutations using the program NACCESS (Hubbard and Thornton, 1993). If a residue has less than 1% relative SASA percentage, it is defined as a buried residue. Otherwise, the residue is an exposed surface residue. Again, if the residue is observed in multiple structures, we employed the lowest relative SASA value. We also ran FINDSITE* (Zhou and Skolnick, 2013) to annotate computationally ligand-binding sites and EFICAz (Kumar and Skolnick, 2012) to predict FDRs for predicted enzymes. Calculations of the free energy differences, $\Delta G$, of Trp mutations were conducted using the program DMutant, which is based on statistical potentials (Zhou and Zhou, 2002) and ranked among the top performers in an independent assessment (Khan and Vihinen, 2010). Secondary structure analysis was carried out using DSSP (Kabsch and Sander, 1983).

Statistical Analysis
With a few noted exceptions, a Fisher’s exact test was conducted on the contingency table given in Table 3. A positive case is a variant containing a certain feature, such as ligand binding, or a mutation from a Cys residue. A negative case is a variant without such a feature. Outcomes are either disease associated or neutral. The counts of positive/negative cases with disease are denoted as $d_{pos}/n_{pos}$, respectively, and similarly, $d_{neg}/n_{neg}$, for positive/negative cases for neutral ones. Odds are calculated for positive and negative cases separately, and their ratio yields the OR.

Availability
The datasets are available as Supplemental Information and also online at http://csbio.biology.gatech.edu/nsnsp.

SUPPLEMENTAL INFORMATION

Supplemental Information includes a Supplemental Spreadsheet and can be found with this article online at http://dx.doi.org/10.1016/j.str.2015.03.028.

ACKNOWLEDGMENTS

This work is supported by grant no. GM-37408 of the Division of General Medical Sciences of the NIH.

Received: January 26, 2015
Revised: March 3, 2015
Accepted: March 6, 2015
Published: May 28, 2015
REFERENCES


