Why Human Disease-Associated Residues Appear as the Wild-Type in Other Species: Genome-Scale Structural Evidence for the Compensation Hypothesis

Jinrui Xu[1] and Jianzhi Zhang*[,2]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan

[2]Department of Ecology and Evolutionary Biology, University of Michigan

***Corresponding author:** E-mail: jianzhi@umich.edu.

Associate editor: Jeffrey Thorne

Abstract

Many human-disease associated amino acid residues (DARs) appear as the wild-type in other species. This phenomenon is commonly explained by the presence of compensatory residues in these other species that alleviate the deleterious effects of the DARs. The general validity of this hypothesis, however, is unclear, because few compensatory residues have been identified. Here we test the compensation hypothesis by assembling and analyzing 1,077 DARs located in 177 proteins of known crystal structures. Because destabilizing protein structures is a primary reason why DARs are deleterious, we focus on protein stability in this analysis. We discover that, in species where a DAR represents the wild-type, the destabilizing effect of the DAR is generally lessened by the observed amino acid substitutions in the spatial proximity of the DAR. This and other findings provide genome-scale evidence for the compensation hypothesis and have important implications for understanding epistasis in protein evolution and for using animal models of human diseases.

Key words: **disease mutation, epistasis, evolution, intramolecular interaction, protein stability.**

Introduction

It was first reported in 2002 that a number of human disease-associated amino acid residues (DARs) appear as the wild-type in the laboratory mouse and various other species (Kondrashov et al. 2002; Waterston et al. 2002). For example, mutation from Gly to Ser at amino acid position 471 of human androgen receptor causes the complete androgen insensitivity syndrome, characterized by feminization of genetic males, but Ser is the wild-type residue (WTR) in both mouse and rat (Gao and Zhang 2003). Uncovering the cause of this interesting phenomenon can help understand both the molecular basis of human disease and the mechanisms of protein evolution. We previously reported that these special DARs are not enriched in associations with late-onset or mild diseases and that their wild-type status in nonhuman species is not attributable to founder effects as one might hypothesize in the case of the laboratory mouse (Gao and Zhang 2003). Instead, it was proposed from the very beginning (Kondrashov et al. 2002) and is now widely believed (Gao and Zhang 2003; Kulathinal et al. 2004; Ferrer-Costa et al. 2007; Baresic et al. 2010) that human DARs can become WTRs in other species because of the presence in these species of compensatory residues that alleviate the deleterious effects of the DARs. Nevertheless, because potential compensatory residues have been identified in only a few cases (Kondrashov et al. 2002), the general validity of the compensation hypothesis remains unclear. For two reasons, protein structural analysis may provide significant insights. First, a primary mechanism by which DARs cause diseases is reducing protein structural stability (Yue et al. 2005). Second, compensatory residues of a DAR likely reside in the same protein as the DAR and interact with the DAR (Poon et al. 2005; Davis et al. 2009; Baresic et al. 2010), and thus may be detected through structural analysis. Here, we assemble a large set of structurally mapped DARs that appear as the wild-type in at least one nonhuman species and test whether the potential compensatory residues in the spatial neighborhood of the DARs mitigate the destabilizing effects of the DARs in the nonhuman species.

Results

Protein Stability Reduction Caused by DARs

We began with 51,920 DARs from the Human Gene Mutation Database (Stenson et al. 2003) and Universal Protein Resource (UniProt) (UniProt Consortium 2011). Among them, 9,212 DARs were mapped to 579 unique human protein structures from the Protein Data Bank (PDB) (Berman 2008). Of these structurally mapped DARs, 1,077 appear as the wild-type in the one-to-one orthologous proteins of at least one nonhuman species (Altenhoff et al. 2011) and thus are called wt-DARs. Although wt-DARs are often referred to as compensated pathogenic deviations (Kondrashov et al. 2002) in the literature, we avoid the use of this term because it equates a phenomenon (DAR observed as the wild-type in other species) with one of its potential causes (compensation). The remaining 8,135 DARs are referred to as regular DARs, or rg-DARs. We used Rosetta (Kellogg et al. 2011) to predict the change in human protein stability upon mutation from the WTR to the corresponding DAR ($\Delta\Delta G = \Delta G_{DAR} - \Delta G_{WTR}$). The more positive $\Delta\Delta G$ is, the bigger the stability reduction is.

© The Author 2014. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://mbe.oxfordjournals.org/ at University of Michigan on June 25, 2014

Article

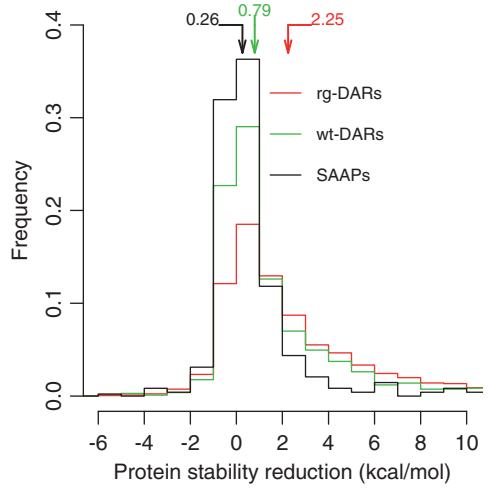

FIG. 1. Frequency distributions of human protein stability reduction ($\Delta\Delta G$) caused by mutations to human single amino acid polymorphisms (SAAPs) with minor allele frequencies (MAFs) > 0.01 (black), disease-associated residues that appear as the wild-type in at least one nonhuman species (wt-DARs) (green), and other disease-associated residues (rg-DARs) (red). The samples include 482 SAAPs, 1,077 wt-DARs, and 8,124 of the 8,135 rg-DARs (11 rg-DARs are not included because Rosetta failed to complete the computations in 72 h). Protein stability reduction is expressed in kcal/mol estimated from REU by linear regression (supplementary fig. S1, Supplementary Material online). Arrows indicate median values of the distributions. The three distributions are all significantly different from one another ($P < 10^{-14}$, Mann–Whitney U test).

Thus, $\Delta\Delta G$ is referred to as the stability reduction upon mutation. The median $\Delta\Delta G$ for mutations to wt-DARs is 1.44 Rosetta energy unit (REU), which is equivalent to ~0.79 kcal/mol according to a linear conversion model (supplementary fig. S1, Supplementary Material online). This amount is significantly smaller than the median $\Delta\Delta G$ (4.09 REU or ~2.25 kcal/mol) for mutations to rg-DARs ($P < 10^{-41}$, Mann–Whitney U test; fig. 1), consistent with an earlier observation that mutations to wt-DARs have on average weaker impacts on structural stabilities than mutations to rg-DARs (Ferrer-Costa et al. 2007).

That wt-DARs impose milder destabilizing effects than rg-DARs has two reasons. First, wt-DARs are more similar to WTRs than are rg-DARs in physicochemical properties (Ferrer-Costa et al. 2007). Second, the structural positions of wt-DARs and rg-DARs may be different such that the same type of mutation has different destabilizing effects when leading to wt-DARs versus rg-DARs. To explore this possibility, we analyzed, among all 380 possible types of amino acid changes, the 128 types that are observed in mutations to both wt-DARs and rg-DARs in our data set (supplementary table S1, Supplementary Material online). Among these 128 types, 13 showed a significantly smaller median $\Delta\Delta G$ for mutations to wt-DARs than mutations to rg-DARs ($P < 0.05$, Mann–Whitney U test; supplementary table S1, Supplementary Material online), whereas none showed the opposite pattern. Thus, for some mutation types, wt-DARs are located at positions with milder stability impacts than rg-DARs. Furthermore, there is a negative correlation between sample size and log(P value) in the above Mann–Whitney U test (supplementary fig. S2, Supplementary Material online), suggesting that more mutation types would show the same significant trend as the 13 mutation types should the samples be larger. Thus, there is indeed evidence that on average wt-DARs are located at positions that have milder stability impacts than are rg-DARs.

The observation that wt-DARs are less destabilizing than rg-DARs suggests that the mechanism mitigating the deleterious effects of DARs in nonhuman species has a limited power. As a comparison, we also computed the average $\Delta\Delta G$ for mutations to known common single amino acid polymorphisms (SAAPs) in humans (i.e., with allele frequencies >0.01) (Sherry et al. 2001), which should be mostly neutral. As expected, this $\Delta\Delta G$ (median = 0.47 REU or ~0.26 kcal/mol) is significantly lower than that for wt-DARs ($P < 10^{-14}$; fig. 1).

Testing the Compensation Hypothesis

Intramolecular compensatory residues may appear anywhere in a protein to mitigate protein stability reduction caused by a wt-DAR, because protein stability is contributed by all residues. However, spatially neighboring residues of the wt-DAR can have strong stabilizing effects via noncovalent bonds. Furthermore, it is currently infeasible to examine the potential compensatory effects of a large number of residues simultaneously, whereas examining these residues one by one requires the information of the order with which these residues emerged in evolution, which is difficult to obtain. Thus, in this study, we focused on only the spatial neighborhood of a wt-DAR when examining potential compensatory residues. For reasons detailed in Materials and Methods, we considered all residues that are within 4 Å from a focal residue to be its neighboring residues, where the distance between two residues is measured by the shortest spatial distance of their nonhydrogen atoms. We found that, in 94.6% of the cases when a DAR is the wild-type in a species, the neighboring residues are not identical between that species and human; these cases were subject to further analysis.

Let us use the example of plasminogen to illustrate our analysis (fig. 2). Plasminogen is the precursor of plasmin, which dissolves the fibrin of blood clots. Normal humans have Arg at amino acid position 532 of plasminogen, but mutation to His at this position causes plasminogen deficiency (Online Mendelian Inheritance in Man or OMIM: 217090), characterized by decreased serum plasminogen activity. Interestingly, His is the wild-type in the giant panda. Four neighboring residues of this DAR differ between wild-type human and giant panda and are candidate compensatory residues. We computed the stability reduction caused by the mutation from Arg to His in the human structure ($\Delta\Delta G_1$; fig. 2A). We also computed the corresponding stability reduction caused by the same mutation in the "pandanized" human structure where all neighboring residues are of the panda version ($\Delta\Delta G_2$; fig. 2B). Consistent with the compensation hypothesis, $\Delta\Delta G_2$ (−4.43 REU or ~−2.43 kcal/mol) is

Downloaded from http://mbe.oxfordjournals.org/ at University of Michigan on June 25, 2014

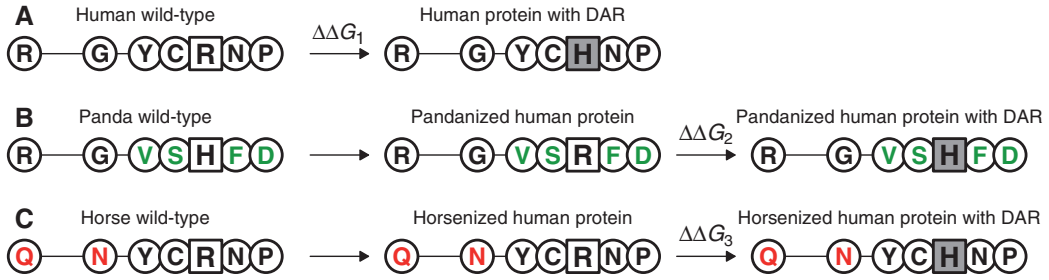

Fig. 2. Testing the compensation hypothesis for the disease-associated residue (DAR) at position 532 of human plasminogen (UniProt accession number: P00747). The DAR site and its orthologous site in nonhuman species are squared, and the DAR is shaded. Spatial neighbors of the DAR site, shown as circles, are identified using the human plasminogen model (2KNF in PDB as the template). (*A*) Wild-type sequence in human (P00747) and the stability reduction ($\Delta\Delta G_1$) of the human plasminogen caused by mutation from the wild-type (R) to the DAR (H). (*B*) Panda wild-type plasminogen (G1MBX3), "pandanized" human plasminogen, and the stability reduction ($\Delta\Delta G_2$) of the pandanized human plasminogen caused by mutation from the human wild-type (R) to the DAR (H). The neighboring residues in panda that differ from those in human are shown in green. (*C*) Horse wild-type plasminogen (F6USP9), "horsenized" human plasminogen, and the stability reduction ($\Delta\Delta G_3$) of the horsenized human plasminogen caused by mutation from the human wild-type (R) to the DAR (H). The neighboring residues in horse that differ from those in human are shown in red. Sequence alignment is provided in supplementary figure S6, Supplementary Material online.

substantially smaller than $\Delta\Delta G_1$ (1.19 REU or ~0.65 kcal/mol), suggesting that one or more of the four neighboring residues in panda that differ from human are compensatory. The negative $\Delta\Delta G_2$ suggests that the replacement of Arg with His increases the panda plasminogen stability and thus may have been beneficial. As a negative control, we considered horse, in which Arg is the wild-type. We computed the stability reduction caused by the mutation from Arg to His in the "horsenized" human structure where all neighboring residues are of the horse version ($\Delta\Delta G_3$; fig. 2*C*). As expected, $\Delta\Delta G_3$ (2.99 REU or ~1.65 kcal/mol) is not smaller than $\Delta\Delta G_1$, indicating that the smaller $\Delta\Delta G_2$, compared with $\Delta\Delta G_1$, is not due to random substitutions. We caution, however, that $\Delta\Delta G$ prediction is notoriously difficult and that Rosetta and other top ranked prediction programs have only moderate accuracies (Khan and Vihinen 2010; Thiltgen and Goldstein 2012). Consequently, $\Delta\Delta G$ comparison for any individual case may not be reliable; only comparisons based on large samples are trustable.

We conducted the same analyses for a large set of wt-DARs. For each wt-DAR, we averaged $\Delta\Delta G_2$ from multiple species if the DAR is found to be the wild-type in multiple species. We then compared the average $\Delta\Delta G_2$ with the corresponding $\Delta\Delta G_1$. Overall, $\Delta\Delta G_2$ (median = 1.23 REU or ~0.68 kcal/mol) is significantly smaller than $\Delta\Delta G_1$ (median = 1.59 REU or ~0.87 kcal/mol) ($P < 10^{-7}$, Wilcoxon signed-rank test; fig. 3). For each wt-DAR, $\Delta\Delta G_1 - \Delta\Delta G_2$ measures the stabilizing effect of the neighboring residues from the species where the DAR is the wild-type. A positive value of ($\Delta\Delta G_1 - \Delta\Delta G_2$) indicates that those neighboring residues are compensatory. In spite of the statistically significant difference between $\Delta\Delta G_1$ and $\Delta\Delta G_2$, the median of ($\Delta\Delta G_1 - \Delta\Delta G_2$) is rather small (0.17 REU or 0.09 kcal/mol). We found that in fact 52.7% of the wt-DARs have $\Delta\Delta G_1 < 1$ kcal/mol, which are not conventionally considered to be destabilizing (Tokuriki and Tawfik 2009). For those wt-DARs considered to be destabilizing ($\Delta\Delta G_1 > 1$ kcal/mol), the median of ($\Delta\Delta G_1 - \Delta\Delta G_2$) is 1.03 REU or

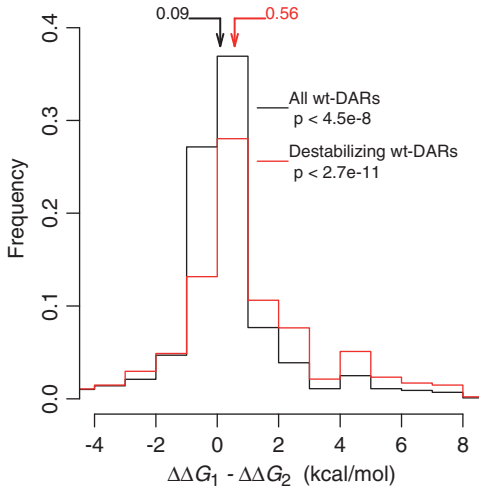

Fig. 3. Frequency distribution of the difference in protein stability reduction upon mutation from a human WTR to a DAR in the absence ($\Delta\Delta G_1$) and presence ($\Delta\Delta G_2$) of neighboring residues from a species where the DAR is the wild-type. The larger the difference, the greater the compensation effect. Destabilizing wt-DARs have $\Delta\Delta G_1 > 1$ kcal/mol. Arrows indicate median values of the corresponding distributions. For both distributions, $\Delta\Delta G_1 - \Delta\Delta G_2$ is significantly biased toward positive values, as indicated by the *P* values from the Wilcoxon signed-rank test.

~0.56 kcal/mol ($P < 10^{-10}$, fig. 3). Because some proteins harbor many more wt-DARs than do other proteins, we also respectively averaged $\Delta\Delta G_1$ and $\Delta\Delta G_2$ values from different wt-DARs in the same protein before comparison, but the results were similar ($P < 0.003$; $P < 0.007$ for destabilizing wt-DARs; supplementary fig. S3, Supplementary Material online).

To compare $\Delta\Delta G_3$ and $\Delta\Delta G_2$, we focused on destabilizing wt-DARs. For each wt-DAR, we need a pair of species whose WTRs are the same as the human DAR and the human WTR, respectively. We chose those species pairs that have the same numbers of neighboring residue

Downloaded from http://mbe.oxfordjournals.org/ at University of Michigan on June 25, 2014

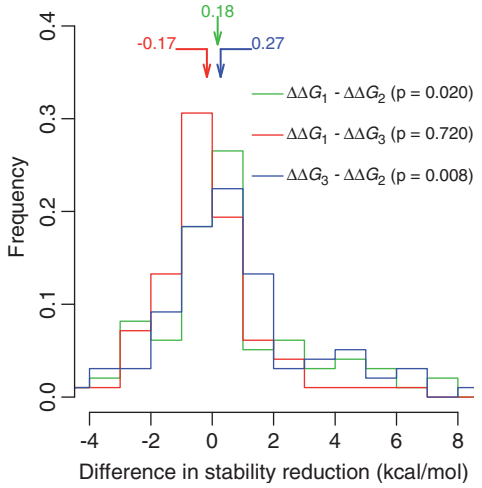

Fig. 4. Frequency distribution of the difference in protein stability reduction upon mutation from a human WTR to a DAR among various genetic backgrounds. $\Delta\Delta G_1$, in the human background (see fig. 2*A*); $\Delta\Delta G_2$, in the presence of neighboring residues from a species where the DAR is the wild-type (see fig. 2*B*); $\Delta\Delta G_3$, in the presence of neighboring residues from a nonhuman species where the human WTR is the wild-type (see fig. 2*C*). The *P* values are from one-tail Wilcoxon signed-rank test. A total of 314 pairs of WTRs and destabilizing DARs are examined.

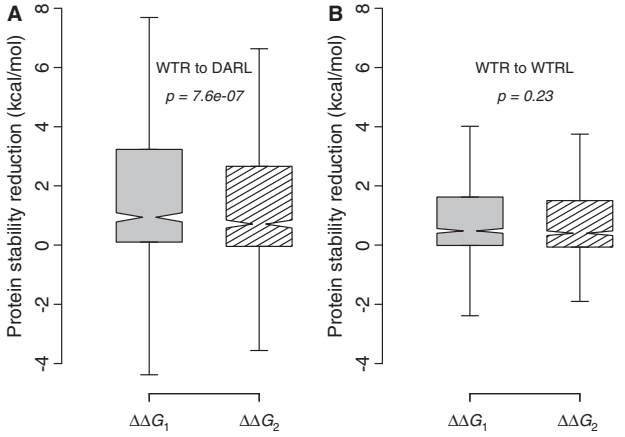

Fig. 5. Protein stability reduction upon mutation. (*A*) Distribution of protein stability reduction upon mutation from a human WTR to a residue that is physicochemically similar to a DAR in the absence ($\Delta\Delta G_1$, gray bar) and presence ($\Delta\Delta G_2$, striped bar) of neighboring residues from a species where the DAR is the wild-type. (*B*) Distribution of protein stability reduction upon mutation from a human WTR to a residue that is physicochemically similar to the WTR (WTRL) in the absence ($\Delta\Delta G_1$, gray bar) and presence ($\Delta\Delta G_2$, striped bar) of neighboring residues from a species where the DAR is the wild-type. The *P* values are from Wilcoxon signed-rank test. A total of 590 pairs of WTRs and DARs are examined in each panel.

differences from the human protein. This requirement reduced our sample size substantially but allowed a fair comparison between $\Delta\Delta G_3$ and $\Delta\Delta G_2$. We found that $\Delta\Delta G_2$ remains significantly smaller than $\Delta\Delta G_1$ ($P = 0.02$; fig. 4), whereas $\Delta\Delta G_3$ is not significantly different from $\Delta\Delta G_1$ ($P > 0.5$; fig. 4). Furthermore, $\Delta\Delta G_2$ is significantly smaller than $\Delta\Delta G_3$ ($P < 0.01$; fig. 4). Thus, as predicted by the compensation hypothesis, the compensatory effects are bestowed by the neighboring residues in species where the human DARs are the wild-type, but not by the neighboring residues in species where the human WTRs are the wild-type.

Compensatory Effects Extend to Amino Acids Similar to DARs

If the above detected compensatory effects of neighboring residues are due to physical interactions between the neighboring residues and the DARs, the compensatory effects may also exert on amino acids that are physicochemically similar to the DARs. Because the greater the physicochemical similarity between two amino acids, the higher the substitution rate between them in evolution (Miyata et al. 1979; Zhang 2000), we used the PAM250 substitution matrix (Dayhoff et al. 1978) to gauge physicochemical similarities between amino acids. For each DAR, we identified the non-WTR amino acid(s) that the DAR will most likely be replaced with in evolution according to PAM250 and referred to it as DAR-like (DARL). There may be more than one DARL if several amino acids are equally likely to replace the DAR. Similarly, for each WTR, we identified the non-DAR amino acid(s) that the WTR will most likely be replaced with in evolution (WTRL). If the WTRL set and DARL set identified for a WTR and its DAR overlap, we do not consider the case further. We then examined the stability reduction caused by mutation from WTR to DARL in the human protein ($\Delta\Delta G_1$) and the corresponding stability reduction in the presence of the neighboring residues from a species in which the DAR is the wild-type ($\Delta\Delta G_2$). As predicted, the compensatory effects of the neighboring residues also exert on DARLs ($P < 10^{-6}$; fig. 5*A*). By contrast, no such effect for WTRLs is detectable ($P > 0.2$; fig. 5*B*).

Discussion

Taken together, our results provide genome-scale evidence that, in species where DARs appear as the wild-type, residues at the spatial proximities of the DARs mitigate their deleterious effects in destabilizing the protein structures. Because reducing protein stability is a primary mechanism by which DARs cause diseases, our findings support the hypothesis that compensatory residues render the otherwise unacceptable DARs acceptable in evolution.

A few biologically or medically important protein families have been intensively crystallized, whereas most other protein families have few members with solved structures. To examine whether our results have been influenced by this imbalanced data, we focused on a subset of protein structures with pairwise sequence identity <60%. We found that our results in figure 3 can be repeated by this subset of data (supplementary fig. S4, Supplementary Material online), suggesting that the compensation hypothesis is supported robustly by many protein families rather than a few. It is worth pointing out that Rosetta predictions of $\Delta\Delta G$ are not always accurate (Kellogg et al. 2011), which limits the statistical power of our

Downloaded from http://mbe.oxfordjournals.org/ at University of Michigan on June 25, 2014

analysis, but also means that our conclusions are likely to be conservative.

Despite the detection of statistically significant compensatory effects, the median difference between $\Delta\Delta G_1$ and $\Delta\Delta G_2$ is quite small even for destabilizing wt-DARs (0.56 kcal/mol), indicating that the overall compensatory effect detected is small. Although the actual compensation may be larger if some compensatory residues are outside the 4 Å neighborhood examined, even the small compensatory effect detected could have appreciable impacts. Because wild-type proteins are only marginally stable (folding energy = −3 to −10 kcal/mol) (Tokuriki and Tawfik 2009) and mutations to destabilizing wt-DARs have a median $\Delta\Delta G$ of 3.54 kcal/mol, proteins with wt-DARs could become marginally unstable ($\Delta G > 0$ kcal/mol). When $\Delta G \sim 0$, a small change in ΔG could result in a substantial change in the fraction of folded protein molecules. For example, a wild-type protein with $\Delta G = -3$ kcal/mol has >99% of molecules folded under 37 °C (see Materials and Methods). Upon mutation to an average destabilizing wt-DAR ($\Delta\Delta G = 3.54$ kcal/mol), folded protein molecules drop to 30% ($\Delta G = 0.54$ kcal/mol). With the help of the detected median compensatory effect ($\Delta\Delta G = -0.56$ kcal/mol), the fraction of folded molecules rises to 51% ($\Delta G = -0.02$ kcal/mol). Because most diseases are recessive, heterozygotes with one wild-type allele and one null allele (i.e., having 50% functional molecules as in the wild-type) are often phenotypically normal. Hence, a homozygote with the median destabilizing wt-DAR and median compensatory effect, producing 51% of folded molecules, likely has a normal phenotype. In other words, the compensation detected, although small in terms of $\Delta\Delta G$, may be sufficient in restoring the normal phenotype. The substantial reduction of the fraction of unfolded molecules, which are often cytotoxic, may render the compensation even more important.

That a large fraction of wt-DARs are explainable, at the genomic scale, by the presence of spatially neighboring compensatory residues supports the importance of (intramolecular) epistasis in protein evolution (Breen et al. 2012). The compensatory residues of the DARs identified through our evolutionary analysis may help understand the molecular basis of the involved diseases. Nevertheless, rampart epistasis in protein evolution also means that findings from animal models of human diseases need to be interpreted with care (Liao and Zhang 2008). It is noteworthy that in 5.4% of the cases when a DAR is the wild-type in a species, that species has identical neighboring residues as human. In these cases, whether compensatory residues reside outside the neighborhood defined or other mechanisms are at work remains to be explored.

Materials and Methods

Neighboring Residues

For each residue in a protein, we calculated the number of residues whose spatial distance from this focal residue is between 0 and 0.1 Å, between 0.1 and 0.2 Å, and so on. We then computed the residue density, defined as the number of residues per $Å^3$, for each range of radial distance. We averaged the density across all residues of all nonredundant protein structures from the protein structure database CATH (Sillitoe et al. 2013). The density peaks at 1.4 and 3.3 Å (supplementary fig. S5, Supplementary Material online), representing residue pairs in contact via N–O and hydrogen bonds, respectively. The density drops drastically and appears uniformly distributed at spatial distances above 4 Å. Because the density is contributed by residues that are in contact and residues that are not in contact, the uniformly low density suggests that residues with distances beyond 4 Å tend not to be in contact. Further, proteins are primarily stabilized by electrostatic bonds, hydrogen bonds, and van der Waals interactions, which have distances of ~3.0, 2.6–3.5, and averaging 3.6 Å between two nonhydrogen atoms, respectively. Therefore, we identify potential compensatory residues within the 4 Å radius.

Protein Structures

Human protein structures were downloaded from PDB (Berman 2008), whereas the SIFTS database (Velankar et al. 2013) was used to map the structures with corresponding proteins in UniProt (UniProt Consortium 2011). Based on the alignments of the structures and their corresponding wild-type sequences, we removed the structures that have point mutations or insertions/deletions (indels) totaling >10% of amino acids in the structures. For the remaining structures that contain point mutations or indels totaling ≤10%, we used them as templates to predict structure models of their corresponding wild-type proteins for the aligned regions, by MODELLER (Eswar et al. 2008). Because the templates and queries have sequence identities ≥90%, the predicted structure models are likely to be highly accurate. These models and native structures formed the structure pool for testing the compensation hypothesis.

We mapped DARs onto the protein structures. When one DAR is mapped to multiple structures, we used the structure containing the highest number of DARs, which reduces structure redundancy in the sample and saves computational time. One-to-one orthologs were obtained from the orthologous matrix database (Altenhoff et al. 2011). Only structure–ortholog alignments with deletion sites <10% of the amino acid residues in the structures were used. From these alignments, we found that 1,077 human DARs appear as the wild-type in at least one nonhuman species. In an alignment between a human protein and one of its orthologs where a DAR appears as the wild-type, if none of the neighboring residues of the DAR site in the human protein corresponds to a gap site in the ortholog and at least one neighboring residue differs between the human protein and the ortholog, the corresponding neighboring residues in the ortholog are considered to be potential compensatory residues for the DAR. A total of 1,008 wt-DARs have at least one set of potential compensatory residues.

Human SAAPs were acquired from UniProt. SAAPs were cross-linked to their single nucleotide polymorphisms (SNPs) in dbSNP where the minor allele frequencies (MAFs) in

Downloaded from http://mbe.oxfordjournals.org/ at University of Michigan on June 25, 2014

humans were obtained. Only SAAPs with MAFs $\geq$ 0.01 were used.

Prediction of $\Delta\Delta G$

Program "ddg_min" in Rosetta with default parameters was used for energy minimizations of protein structures. Then, "ddg_monomer" was used to predict protein stability reductions upon point mutations. Low Resolution Protocol was set for the prediction using default parameters except for the following changes. We repacked the residues with C_α in 7 Å rather than 8 Å to the site of the point mutation. The 7 Å in C_α distance was chosen because we found it corresponds to 4 Å in heavy atom distance from the structures used in the "neighboring residues" section. The iteration parameter was set to 30 instead of 50 to save computational time. FoldX (Guerois et al. 2002) was used to optimize the neighboring residue side chain orientation in a protein structure upon the replacement of neighboring residues.

Relationship between Fraction of Protein Molecules Folded and Protein Stability

Under the assumption of thermodynamic equilibrium, the fraction of protein molecules folded is given by $\frac{1}{1+e^{\Delta G/(kT)}}$, where ΔG is protein stability, k is Boltzmann constant (1.986 cal/mol/K), and T is absolute temperature (Pakula and Sauer 1989).

Supplementary Material

Supplementary figures S1–S6 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/) is valid.

Acknowledgments

The authors thank Wei-Chin Ho, Jian-Rong Yang, and three anonymous reviewers for valuable comments. This work was supported in part by research grant R01GM103232 from the U.S. National Institutes of Health to J.Z. All data used in this study can be obtained at http://www.umich.edu/~zhanglab/download/Jinrui_MBE_Suppl/index.htm.

References

Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. 2011. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.* 39:D289–D294.

Baresic A, Hopcroft LE, Rogers HH, Hurst JM, Martin AC. 2010. Compensated pathogenic deviations: analysis of structural effects. *J Mol Biol.* 396:19–30.

Berman HM. 2008. The Protein Data Bank: a historical perspective. *Acta Crystallogr A.* 64:88–95.

Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2012. Epistasis as the primary factor in molecular evolution. *Nature* 490: 535–538.

Davis BH, Poon AF, Whitlock MC. 2009. Compensatory mutations are repeatable and clustered within proteins. *Proc Biol Sci.* 276: 1823–1827.

Dayhoff MO, Schwartz R, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dathoff MO, editor. Atlas of protein sequence and structure. Silver Spring (MD): National Biomedical Research Foundation. p. 345–352.

Eswar N, Eramian D, Webb B, Shen MY, Sali A. 2008. Protein structure modeling with MODELLER. *Methods Mol Biol.* 426:145–159.

Ferrer-Costa C, Orozco M, de la Cruz X. 2007. Characterization of compensated mutations in terms of structural and physico-chemical properties. *J Mol Biol.* 365:249–256.

Gao L, Zhang J. 2003. Why are some human disease-associated mutations fixed in mice? *Trends Genet.* 19:678–681.

Guerois R, Nielsen JE, Serrano L. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol.* 320:369–387.

Kellogg EH, Leaver-Fay A, Baker D. 2011. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* 79:830–838.

Khan S, Vihinen M. 2010. Performance of protein stability predictors. *Hum Mutat.* 31:675–684.

Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A.* 99: 14878–14883.

Kulathinal RJ, Bettencourt BR, Hartl DL. 2004. Compensated deleterious mutations in insect genomes. *Science* 306:1553–1554.

Liao BY, Zhang J. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A.* 105:6987–6992.

Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. *J Mol Evol.* 12:219–236.

Pakula AA, Sauer RT. 1989. Genetic analysis of protein stability and function. *Annu Rev Genet.* 23:289–310.

Poon A, Davis BH, Chao L. 2005. The coupon collector and the suppressor mutation: estimating the number of compensatory mutations by maximum likelihood. *Genetics* 170:1323–1332.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29:308–311.

Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, et al. 2013. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.* 41:D490–D498.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat.* 21:577–581.

Thiltgen G, Goldstein RA. 2012. Assessing predictors of changes in protein stability upon mutation using self-consistency. *PLoS One* 7: e46084.

Tokuriki N, Tawfik DS. 2009. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol.* 19:596–604.

UniProt Consortium. 2011. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* 39:D214–D219.

Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, Oldfield TJ, O'Donovan C, Martin MJ, Kleywegt GJ. 2013. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* 41:D483–D489.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.

Yue P, Li Z, Moult J. 2005. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol.* 353:459–473.

Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J Mol Evol.* 50: 56–68.

Downloaded from http://mbe.oxfordjournals.org/ at University of Michigan on June 25, 2014