


Protein Destabilization as a Common Factor in Diverse Inherited Disorders

Rachel L. Redler¹ · Jhuma Das¹ · Juan R. Diaz¹ · Nikolay V. Dokholyan¹ 

Received: 15 October 2015 / Accepted: 4 November 2015 / Published online: 19 November 2015
© Springer Science+Business Media New York 2015

Abstract Protein destabilization by amino acid substitutions is proposed to play a prominent role in widespread inherited human disorders, not just those known to involve protein misfolding and aggregation. To test this hypothesis, we computationally evaluate the effects on protein stability of all possible amino acid substitutions in 20 disease-associated proteins with multiple identified pathogenic missense mutations. For 18 of the 20 proteins studied, substitutions at known positions of pathogenic mutations are significantly more likely to destabilize the native protein fold (as indicated by more positive values of $\Delta\Delta G$). Thus, positions identified as sites of disease-associated mutations, as opposed to non-disease-associated sites, are predicted to be more vulnerable to protein destabilization upon amino acid substitution. This finding supports the notion that destabilization of native protein structure underlies the pathogenicity of broad set of missense mutations, even in cases where reduced protein stability and/or aggregation are not characteristic of the disease state.

Keywords Destabilization · Inherited disorder · Pathogenic mutation · Stability · Aggregation

Introduction

The need for proteins to adopt a stable folded structure constrains protein evolvability, as the destabilizing effects of many nascent amino acid substitutions negate any functional improvements they might confer (Drummond and Wilke 2008; Zeldovich et al. 2007; Bloom et al. 2006). Amino acid substitutions encoded by disease-associated single nucleotide polymorphisms (SNPs) may owe their pathogenicity to disruption of one or more crucial features of the native protein, such as overall folding stability, ligand binding, allosteric coupling, catalytic activity, and post-translational maturation (Yue et al. 2005; Wang and Moulton 2001; Xu and Zhang 2014). Several computational tools, including FoldX (Guerois et al. 2002), PolyPhen (Ramensky et al. 2002), PANTHER (Thomas et al. 2003), SIFT (Ng and Henikoff 2003), nsSNPAnalyzer (Bao et al. 2005), PhD-SNP (Capriotti et al. 2006), Eris (Yin et al. 2007a, b), SNAP (Bromberg and Rost 2007), MutPred (Li et al. 2009), SNPs&GO (Calabrese et al. 2009), and PolyPhen2 (Adzhubei et al. 2010) have been developed to (i) determine the impact of pathogenic and benign mutations on the structure and function of a human protein, and (ii) identify mutations that can eliminate the deleterious effects introduced by pathogenic ones. However, it is still unclear whether disease-associated missense mutations are significantly enriched in regions of protein sequence with highest vulnerability to destabilization. To evaluate this possibility and the hypothesis that protein destabilization plays a role in the pathologies of diverse monogenic disorders, we perform an exhaustive survey of amino acid substitutions in 20 proteins with missense mutations linked to various human diseases. For each substitution, we calculate the resulting change in folding free energy ($\Delta\Delta G$) using the Eris software (Yin et al. 2007a, b; Ding and

Rachel L. Redler and Jhuma Das have contributed equally to this work.

✉ Nikolay V. Dokholyan
dokh@unc.edu

¹ Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Dokholyan 2006), utilizing available structures in the PDB as starting structures (Methods section). Substitutions involving cysteine residues (substitutions either to or from cysteine) are excluded, since these $\Delta\Delta G$ calculations cannot account for the contribution of disulfide bonds to protein stability. Proteins surveyed in this study range in number of subunits from 1 (monomeric) to 10 (decameric) and are involved in human diseases involving disparate tissues and pathogenic mechanisms (Table 1).

Despite a lack of evidence for involvement of misfolding and/or aggregation of disease-linked mutant proteins in many cases, the majority (18/20) of proteins assayed exhibit a statistically significant shift toward more positive

$\Delta\Delta G$ values (more destabilizing) for substitutions at sites of disease-associated missense mutations, as compared to those for mutations at non-disease-associated (“neutral”) sites (Table 1; Fig. 1). While a majority of proteins analyzed show a significantly increased propensity for global protein destabilization for substitutions at disease-associated positions (with p -values on the order of 10^{-13} to 10^{-16}), prion and beta-glucuronidase proteins exhibit no significant differences in their distributions of calculated $\Delta\Delta G$ values (when comparing substitutions at disease-linked and neutral amino acid positions). It is possible that pathogenic substitutions in prion and beta-glucuronidase proteins do not influence the overall stability of the

Table 1 Disease-linked proteins analyzed in this study

Gene	Protein product	Prominent-associated disorder(s)	PDBID	# of subunits in biological assembly	Disease-associated/neutral AA positions ^a	p value (neutral vs. pathological $\Delta\Delta G$) ^b
<i>SOD1</i>	Cu/Zn superoxide dismutase	Amyotrophic lateral sclerosis	1SPD	2	63/86	$<2.20 \times 10^{-16}$
<i>TTR</i>	Transthyretin	Amyloidosis	1TTA	4	53/73	$<2.20 \times 10^{-16}$
<i>HBA1</i>	Hemoglobin, alpha subunit	Alpha-thalassemia	2HHB	4 ($2\alpha + 2\beta$)	24/117	1.11×10^{-11}
<i>HBB</i>	Hemoglobin, beta subunit	Beta-thalassemia	2HHB	4 ($2\alpha + 2\beta$)	32/114	1.10×10^{-10}
<i>HPRT1</i>	Hypoxanthine phosphoribosyltransferase 1	Lesch–Nyhan syndrome	1BZY	4	80/134	$<2.20 \times 10^{-16}$
<i>GLA</i>	Alpha-galactosidase	Fabry disease	1R46	2	177/213	$<2.20 \times 10^{-16}$
<i>PKLR</i>	Human erythrocyte pyruvate kinase	Hemolytic anemia	2VGB	4	108/409	5.33×10^{-7}
<i>PAH</i>	Phenylalanine hydroxylase	Phenylketonuria	2PAH	4	185/144	$<2.20 \times 10^{-16}$
<i>ARSB</i>	Arylsulfatase B	Mucopolysaccharidosis VI	1FSU	1	70/404	$<2.20 \times 10^{-16}$
<i>OTC</i>	Ornithine carbamoyltransferase	Hyperammonemia	10TH	3	148/173	$<2.20 \times 10^{-16}$
<i>CD40LG</i>	CD40 ligand	Hyper-IgM syndrome	1ALY	3	35/111	3.00×10^{-15}
<i>UROD</i>	Uroporphyrinogen decarboxylase	Porphyria	1URO	2	51/306	1.84×10^{-13}
<i>PRNP</i>	Prion protein	Various prion diseases	1I4M	2	25/83	0.474
<i>GCH1</i>	GTP cyclohydrolase I	Dopa-responsive dystonia	1FB1	10	57/139	9.31×10^{-13}
<i>CYB5R3</i>	Cytochrome b5 reductase 3	Methemoglobinemia	1UMK	1	26/245	6.06×10^{-7}
<i>OAT</i>	Ornithine aminotransferase	Gyrate atrophy	10AT	2	30/374	1.29×10^{-7}
<i>GUSB</i>	Beta-glucuronidase	Mucopolysaccharidosis VII	1BHG	4	34/577	0.0790
<i>PDHA1</i>	Pyruvate dehydrogenase (alpha subunit)	Lactic acidosis	1NI4	4	46/315	8.10×10^{-6}
<i>PAX6</i>	Paired box 6	Aniridia	6PAX	1	38/95	1.54×10^{-6}
<i>LMNA</i>	Lamin A/C (globular domain)	Muscular dystrophy; cardiomyopathy	1IFR	1	25/88	1.39×10^{-6}

^a Includes only positions at which amino acid substitutions were introduced for $\Delta\Delta G$ calculations in this study (excludes cysteines and any missing residues in crystal structures)

^b As determined by a two-sample Kolmogorov–Smirnov test comparing the complete sets of calculated $\Delta\Delta G$ values for substitutions at positions of known disease mutations (“pathological”) to substitutions at sites without identified pathological mutations (“neutral”)

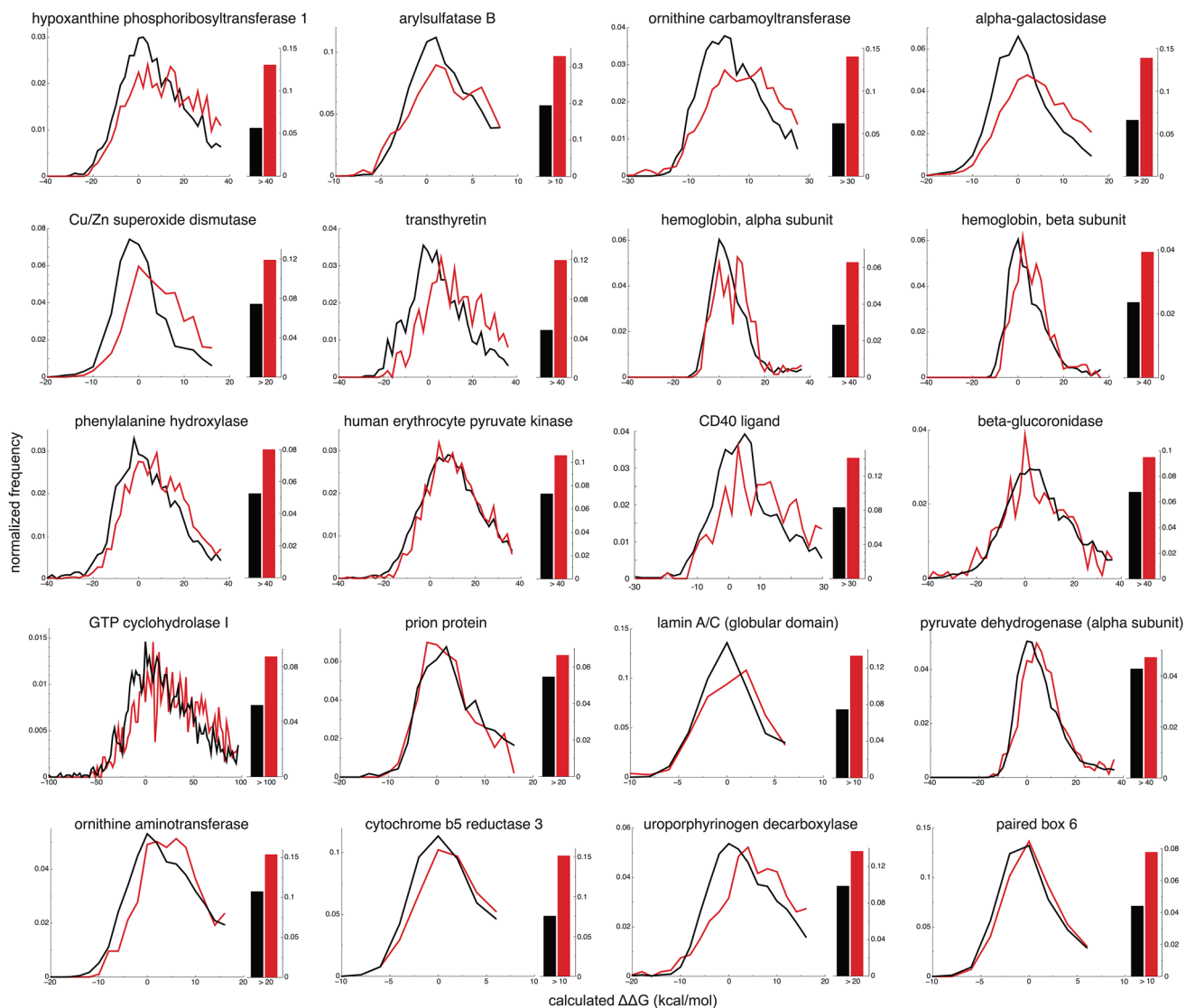


Fig. 1 Amino acid substitutions at sites of known disease-linked missense mutations are more destabilizing than those at neutral positions. Histograms show normalized counts of $\Delta\Delta G$ values calculated for all possible substitutions (excluding those involving cysteine) at sites of known disease-linked substitutions (red curves)

and all other “neutral” amino acid positions (black curves). Bar graphs show the normalized frequencies of $\Delta\Delta G$ values exceeding $n \times 10$ kcal/mol, where n is the number of subunits in the biological oligomeric assembly of each protein (Color figure online)

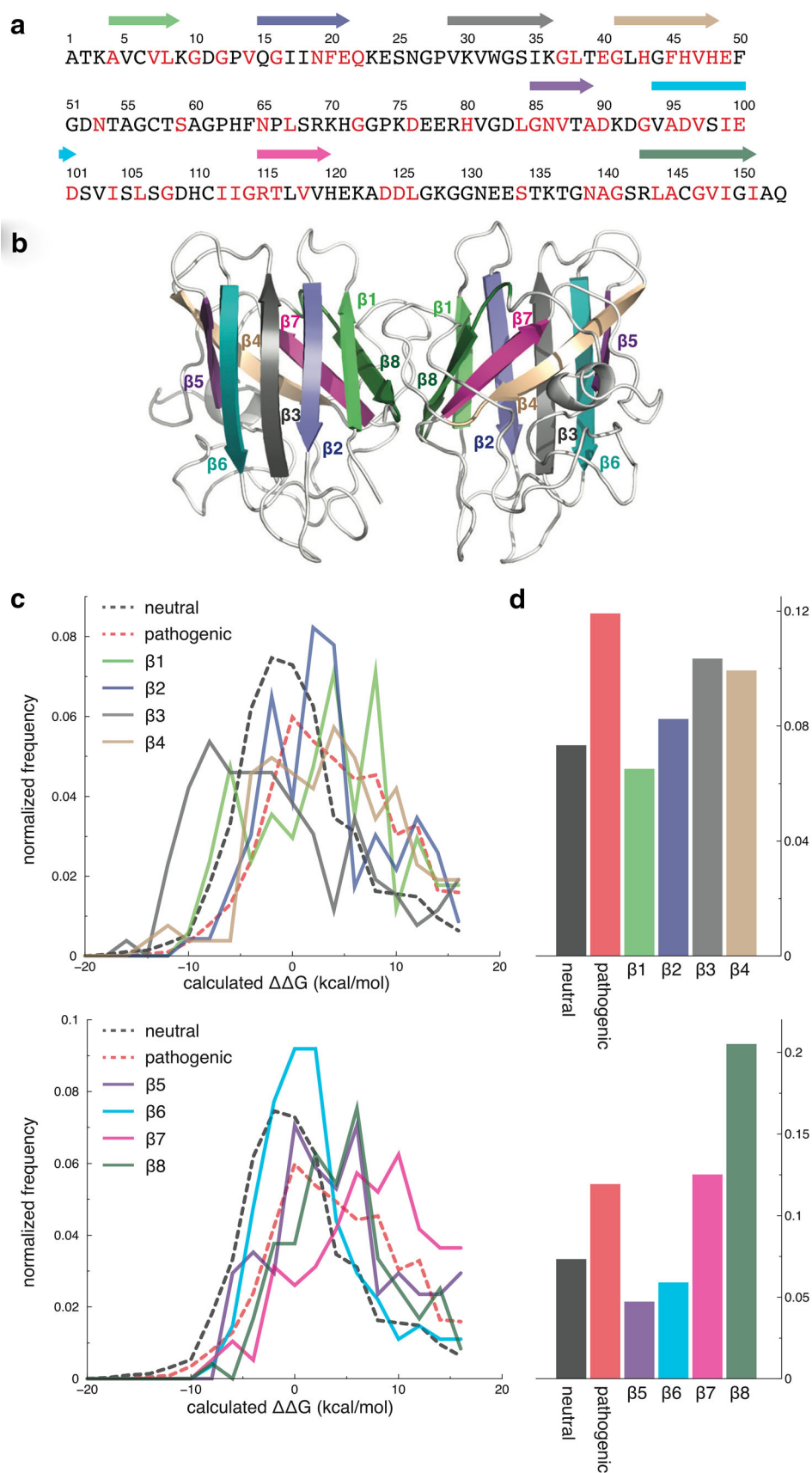
proteins, but disrupt physiological functions by perturbing protein dynamics or interactions with other macromolecules (Sahni et al. 2015).

Using the $\Delta\Delta G$ values calculated as described above, we next evaluated in more detail the vulnerability of individual secondary structure elements of Cu/Zn superoxide dismutase (SOD1) to destabilizing amino acid substitutions (Khare et al. 2006). Sites of ALS-linked mutations in the gene encoding SOD1 are distributed throughout the protein sequence (Redler and Dokholyan 2012), with the exception of a stretch of residues encompassing the third β -strand (Fig. 2a). We hypothesize that the striking absence of pathogenic substitutions identified in this region could be

explained by an increased tolerance to amino acid substitutions. If substitutions in $\beta 3$ are more likely to be neutral or protective (having a distribution shifted toward more negative $\Delta\Delta G$ values), then individuals with these polymorphisms would not be expected to present with ALS. Alternatively, if $\beta 3$ substitutions tend to substantially destabilize SOD1 (more positive $\Delta\Delta G$ values) to the point that expression of the mutant protein is potentially toxic, such substitutions may have never been identified in individuals due to embryonic or early postnatal lethality. To explore this possibility, we compare the $\Delta\Delta G$ distributions of SOD1’s individual β -strands to each other and to the distributions for “pathogenic” and “neutral” amino acid

Fig. 2 Lower vulnerability to destabilization by amino acid substitution may explain the lack of ALS-linked missense mutations in $\beta 3$ of SOD1.

a Amino acid sequence of SOD1. Sites of ALS-causative missense mutations are shown in red and arrows above the sequence mark β -strand regions. **b** Secondary structure of SOD1 (PDBID: 1SPD). **c** Histograms representing calculated $\Delta\Delta G$ values of mutations in neutral and pathogenic sites (as shown in Fig. 1) and in positions comprising each β -strand. **d** Bar graphs showing the normalized frequencies of $\Delta\Delta G$ values exceeding 20 kcal/mol for neutral, pathogenic, and β -strand amino acid positions (Color figure online)



positions in SOD1. We find that the distribution of calculated $\Delta\Delta G$ values is substantially shifted toward more negative values (more stabilizing) for $\beta 3$ (gray curve, Fig. 2b) compared to all other individual β -strands, as well as to the sets of pathogenic and neutral positions. Our results suggest that the absence of ALS-associated mutations in $\beta 3$ is due to this region's exceptional tolerance to amino acid substitution.

Based on this exhaustive survey, we conclude that missense mutations linked to genetic disorders are significantly more likely to occur at positions that perturb the structural stability of native proteins, even when their associated phenotypes do not include observable protein destabilization or aggregation. Prior work probing the link between stability change and pathogenicity of missense mutations has focused on characterization of relatively small sets of disease-linked amino acid substitutions (rather than evaluating all possible substitutions), or comprehensive *in silico* mutagenesis for a single protein (Steffl et al. 2013; Yin et al. 2007a, b). For example, Sahni et al. (2015) recently evaluated the effect of disease-associated missense mutations on overall protein stability and the robustness of interactions with native binding partners, concluding that a minority of disease-linked substitutions lead to a significant overall decrease in protein stability. Rather, they report that pathogenic substitutions are more likely to disrupt a protein's interactions with its native binding partners. In assessing protein products of missense mutations more comprehensively, we find evidence for a widespread vulnerability of disease-associated amino acid positions to destabilizing substitutions.

Most proteins are marginally stable (Dokholyan and Shakhnovich 2001; Dokholyan 2008; Williams et al. 2006) in their functional forms and mutations linked to inherited human disorders exert their toxic effects through a variety of mechanisms, including disruption of the native folding behavior of the affected gene's protein product and concomitant loss of function or novel toxic properties. Our findings, in concordance with previous work, are consistent with the idea that even small reductions in protein stability can lead to dysfunctional proteins associated with human disease, and that disease-linked missense mutations are enriched in regions of protein sequence with highest vulnerability to destabilization. On the other hand, studies of protein evolution have shown that proteins can tolerate many amino acid substitutions, including substitutions at highly conserved regions, by introducing compensatory mutations to counterbalance the effects of deleterious mutations; this phenomenon may explain the fixation of human disease-associated amino acids as wild type residues in orthologous proteins of other species (Xu and Zhang 2014).

All these findings point to the fact that the mutational landscape of proteins is exceedingly complex. Understanding of the biophysical underpinnings of selection for stable, correctly folded, and functional gene products is still lacking (Depristo et al. 2005). Our work reveals a widespread vulnerability of sites of disease-associated mutations to destabilizing substitutions, even when reduced protein stability and/or aggregation are not characteristic of the disease state. Furthermore, the methodology employed provides a platform for the rational control of protein stability through mutagenesis, which could be useful in refining protein evolution models and improving prediction algorithms.

Methods

We employ the Eris suite (Yin et al. 2007a, b; Ding and Dokholyan 2006) (a protein stability evaluation software) to probe the effects of all possible amino acid substitutions in 20 proteins with multiple identified pathogenic missense mutations. Upon substitution, Eris algorithms re-pack the side chains of the residues surrounding the mutated residue using a Monte Carlo simulated annealing procedure. The change of protein stability induced by the mutations to the wild type protein is calculated in terms of $\Delta\Delta G$ ($\Delta\Delta G = \Delta G^{\text{mutant}} - \Delta G^{\text{wild type}}$), utilizing the Medusa force field. We calculate $\Delta\Delta G$ for all possible substitutions at all amino acids included in the crystal structure, excluding substitutions to and from cysteine (i.e., for each amino acid in a given structure, $\Delta\Delta G$ is calculated for 18 non-native variants). We classify amino acid positions as disease-associated if these sites contain at least one pathogenic missense mutation, as documented in the Human Gene Mutation Database (www.hgmd.cf.ac.uk). The differences between sets of all calculated $\Delta\Delta G$ values for disease-associated and neutral amino acid positions are evaluated for significance using the Kolmogorov–Smirnov test (Press et al. 2007).

Acknowledgments This work was supported by the National Institutes of Health grant R01GM080742 to N.V.D. R.L.R. was supported by the National Institutes of Health Predoctoral Fellowship F31NS073435 from the National Institute of Neurological Disorders and Stroke. We thank Michael Caplow and Feng Ding for helpful discussions regarding study design.

Compliance with Ethical Standards

Conflict of interest The author declares that there is no conflict of interests.

Research Involving Human and Animal Rights This study does not involve research with humans and/or animals and it follows all the ethical standards.

References

- Adzhubei IA et al (2010) A method and server for predicting damaging missense mutations. *Nat Meth* 7:248–249
- Bao L, Zhou M, Cui Y (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucl Acids Res* 33:W480–W482
- Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci USA* 103:5869–5874
- Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. *Nucl Acids Res* 35:3823–3835
- Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat* 30:1237–1244
- Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 22:2729–2734
- Depristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* 6:678–687
- Ding F, Dokholyan NV (2006) Emergence of protein fold families through rational design. *PLoS Comput Biol* 2:e85
- Dokholyan NV (2008) Protein designability and engineering. *Structural Bioinformatics*, 2nd edn. Wiley, Hoboken, pp 961–982
- Dokholyan NV, Shakhnovich EI (2001) Understanding hierarchical protein evolution from first principles. *J Mol Biol* 312:289–307
- Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352
- 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
- Khare SD, Caplow M, Dokholyan NV (2006) FALS mutations in Cu, Zn superoxide dismutase destabilize the dimer and increase dimer dissociation propensity: a large-scale thermodynamic analysis. *Amyloid* 13(4):226–235
- Li B et al (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 25:2744–2750
- Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. *Nucl Acids Res* 31:3812–3814
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) *Numerical recipes: the art of scientific computing*, 3rd edn. Cambridge University Press, New York
- Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 30:3894–3900
- Redler R, Dokholyan NV (2012) The complex molecular biology of amyotrophic lateral sclerosis (ALS). *Progr in Molec Biol and Transl Sci* 107:215–262
- Sahni N et al (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161:647–660
- Steff S, Nishi H, Petukh M, Panchenko AR, Alexov E (2013) Molecular mechanisms of disease-causing missense mutations. *J Mol Biol* 425:3919–3936
- Thomas PD et al (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 13:2129–2141
- Wang Z, Moult J (2001) SNPs, protein structure, and disease. *Hum Mutat* 17:263–270
- Williams PD, Pollock DD, Goldstein RA (2006) Functionality and the evolution of marginal stability in proteins: inferences from lattice simulations. *Evol Bioinform Online* 2:91–101
- Xu J, Zhang J (2014) Why human disease-associated residues appear as the wild-type in other species: genome-scale structural evidence for the compensation hypothesis. *Mol Biol Evol* 31:1787–1792
- Yin S, Ding F, Dokholyan NV (2007a) Eris: an automated estimator of protein stability. *Nat Meth* 4:466–467
- Yin S, Ding F, Dokholyan NV (2007b) Modeling backbone flexibility improves protein stability estimation. *Structure* 15:1567–1576
- 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
- Zeldovich KB, Chen P, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci USA* 104:16152–16157