33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

A Structure-Based Computational Mutagenesis Elucidates the Spectrum of Stability-Activity Relationships in Proteins

Majid Masso, *Member, IEEE*, and Iosif I. Vaisman

Abstract— **Protein engineering experiments involving single amino acid substitutions are routinely implemented for the analysis of protein structure, stability, and function. The resulting change in just one of these characteristics relative to the native protein constitutes the focus of any single study, as is the case with predictive computational models developed for the same purpose. Other than investigations into stability-activity trade-offs specifically resulting from active site residue replacements in a few enzymes, a literature survey fails to reveal a comprehensive analysis of stability-activity relationships in proteins upon mutation. Here, we employ a computational mutagenesis for quantifying overall protein structural change upon mutation, which is applied to a dataset of 938 single residue replacements distributed at positions throughout twenty diverse proteins. These mutants are selected based on the availability of both experimental stability and activity change data, and their structural change data are used to characterize the full range of stability-activity relationships.**

I. Introduction

Motivated by recent work in developing a suite of models (http://proteins.gmu.edu/automute) to predict stability and activity changes in proteins upon single residue replacements [1-3], the aim of this study is to further understand stability-activity relationships by identifying mutants for which both properties have been experimentally investigated. In our models, mutants are represented as feature vectors of data generated by a computational mutagenesis methodology that we developed; in particular the procedure yields a scalar, termed a mutant *residual score*, which empirically quantifies overall change in sequence-structure compatibility relative to the native protein upon mutation. A literature search reveals the existence of wide support for a "stability-function hypothesis" stating that active site mutations in enzymes increase stability at the cost of activity, a conjecture which is founded on intuitive physicochemical considerations as well as convincing experimental data from studies on specific proteins [4]. The inverse trade-off, decreased stability in favor of increased activity, remains more controversial due to a lack of consensus in the interpretation of the available data. Additionally, mutants for which stability and activity both either increase or decrease are generally considered anomalous, especially within the context of active sites, and they have even been excluded from further study [4, 5]. Here we investigate mutations distributed throughout enzymes as well as residue substitutions in other types of proteins (e.g., nucleic acid or receptor binding), in order to universally model how stability and activity co-vary upon mutation.

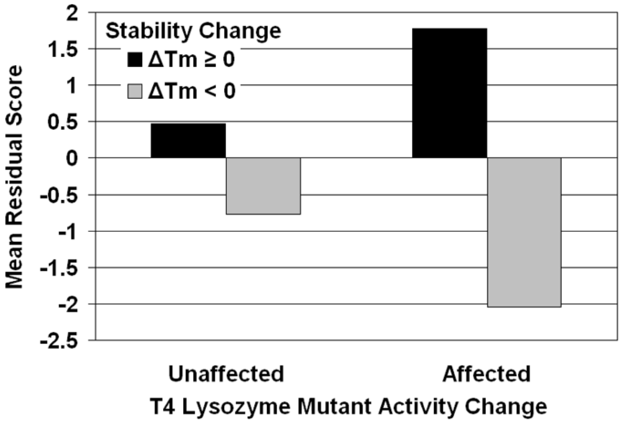


Fig. 1. Elucidation of a distinctive structure (mean residual score) – stability – activity relationship based on a subset of 121 experimental single residue replacements in the T4 lysozyme protein.

To this end, we begin by considering single residue replacements in the lysozyme protein of bacteriophage T4, an enzyme with a polypeptide sequence consisting of 164 amino acids. Rennell *et al.* [6] had published qualitative experimental activity change data for 2015 (65%) of these T4 lysozyme mutants, enabling mutants to be identified as either unaffected (U) or affected (A). We utilized the Rennell *et al.* data in a prior study by computing mutant residual scores and mean residual score (MRS) by mutant category (MRS(U) = –0.76, MRS(A) = –1.40), in order to report on the capability of MRS to elucidate a statistically significant structure (MRS) – function (U/A categories) relationship in T4 lysozyme (*t*-test, $p < 0.001$) [7]. Next, Saraboji *et al.* [8] had retrieved and analyzed a set of 171 mutants of T4 lysozyme with experimental thermal stability change (ΔTm) values that were previously reported in the literature, 121 of which overlap with those of the Rennell *et al.* study. By classifying stability change as increasing (inc, $\Delta Tm \geq 0$) or decreasing (dec, $\Delta Tm < 0$), this common subset of 121 mutants populate all four stability-activity categories, whose mean residual scores exhibit a distinctive trend

Manuscript received March 25, 2011.

Majid Masso is with the Laboratory for Structural Bioinformatics, School of Systems Biology, George Mason University, Manassas, VA 20110 USA (phone: 703-257-5756; fax: 703-993-8976; e-mail: mmasso@gmu.edu).

Iosif I. Vaisman is with the Laboratory for Structural Bioinformatics, School of Systems Biology, George Mason University, Manassas, VA 20110 USA (e-mail: ivaisman@gmu.edu).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

warranting further investigation (Fig. 1). The fact that the two inc (dec) stability change mutant categories display positive (negative) MRS is a testament to the influence of sequence-structure compatibility on overall protein stability (MRS(inc) = 0.64, MRS(dec) = –0.87), revealing a statistically significant structure (MRS) – stability (inc/dec categories) relationship (*t*-test, $p < 0.0005$). Though a trend also exists between MRS and the U/A categories (MRS(U) = –0.48, MRS(A) = –0.66), it does not represent a statistically significant structure–function relationship as was the case with the Rennell *et al.* data, since 121 (4%) T4 lysozyme mutants are too few to adequately represent the mutational landscape.

Saraboji *et al.* retrieved their dataset of T4 lysozyme mutants from the ProTherm Database, a repository of published thermodynamic data for proteins and mutants [9]. This motivated our own search of the database to obtain a larger collection of single residue mutants from a variety of proteins, for which experimental stability and activity change data are available. Notably, ProTherm collects only quantitative mutant activity change data, reported as a percentage of wild type (WT) protein activity. Mutant stability change data include values of the free energy of unfolding based on thermal ($\Delta\Delta G$, 59 mutants) or denaturant ($\Delta\Delta G^{H2O}$, 221 mutants) denaturations (Fig. 2(a)), each measured in kcal/mol, and the difference in midpoint temperatures of thermal unfolding (ΔTm, 251 mutants), reported as °C (Fig. 2(b)). Taking into account bias (e.g., active site mutations are a legitimate focus for many studies) and noise (e.g., multiple instances of the same mutant studied under a variety of temperature and/or pH conditions, with measurements that can vary widely) inherent in the data, the emerging trends suggest that stability changes of small magnitude often leave mutant activity unaffected (or enhanced), whereas larger increases or decreases in mutant stability generally coincide with affected activity (Fig. 2(c)). With the aid of residual scores as empirical measures of structural change upon mutation, the aim of this study is to provide a plausible rationale for this assertion.

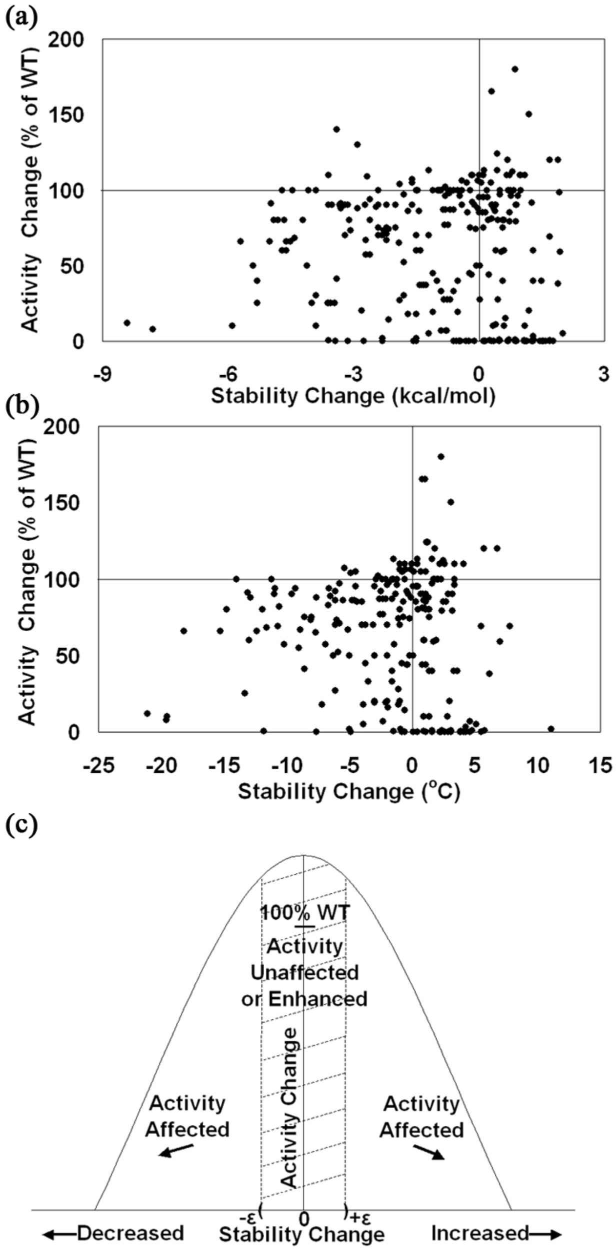


Fig. 2. Scatterplots of single residue replacements retrieved from the ProTherm Database for which published experimental data is available with respect to activity change as well as either (a) $\Delta\Delta G$ or $\Delta\Delta G^{H2O}$ stability change (kcal/mol) or (b) ΔTm stability change (°C). Allowing for dataset bias and noise, the plots in (a) and (b) lead to the hypothesized generalization of stability-activity relationships shown in (c).

II. Materials and Methods

A. Mutant Residual Scores

The residual score of a mutant is calculated using a structure-dependent and knowledge-based computational mutagenesis methodology whose details we previously reported elsewhere [2, 3]. We begin by representing every protein structure discretely as a collection of points in three-dimensional space, with each point corresponding to the center-of-mass of the side chain atoms for a constituent amino acid residue. These points are used as vertices by the Delaunay tessellation algorithm, a computational geometry technique, for generating a tetrahedral tiling of the space occupied by the protein structure. Each tetrahedron objectively identifies at its four vertices a quadruplet of structurally nearest-neighbor residues in the protein; however, since the tessellation is space-filling and tiles pack against one another, each vertex is generally shared by multiple tetrahedra, meaning each residue may participate in a number of such nearest-neighbor quadruplets. To ensure biochemically feasible quadruplet interactions, every protein structure tessellation is initially modified by the removal of edges longer than 12 angstroms.

Without regard to order (i.e., excluding permutations), there are 8855 possible quadruplets that can be generated from a 20-letter protein alphabet, and an average sized protein yields only a few hundred quadruplets (i.e., tetrahedra). Thus, we tessellated a large, diverse set of protein structures in order to calculate for each quadruplet the observed relative frequency of occurrence; a rate expected by chance was then obtained from a multinomial reference distribution. A knowledge-based four-body statistical potential was generated by taking, for every quadruplet, the logarithm of the ratio of observed to expected rates (i.e., a log-likelihood score). Using this potential, every tetrahedron in the tessellation of a protein structure can be assigned a score based on the residue quadruplet represented at the four vertices. Since a vertex is generally shared by multiple tetrahedra, the sum of their scores is referred to as an *environment score* for the residue represented by the vertex. Changing the residue identity at this shared vertex alters the scores of those tetrahedra and leads to a new environment score at the vertex, and subtracting the old residue environment score from the new one yields the residual score for this single residue mutant.

B. Mutant Datasets

A search of the ProTherm Database yields a collection of mutants (531 total including repeats, 227 distinct), each with quantitative experimental stability and activity data, representing single residue substitutions in ten proteins that also have structural coordinate files in the Protein Data Bank (PDB) [10]: barnase (PDB ID: 1bniA), lambda repressor (1lrpA), ribonuclease H1 (2rn2A), chicken lysozyme (4lyzA), ribonuclease T1 (1rn1C), staphylococcal nuclease (1stnA), adenylate kinase (2akyA), citrate synthase (1ctsA), ribonuclease A (1rtbA), and phage T4 lysozyme (3lzmA).

Our own literature search for published experimental data identified 711 additional mutants, representing distinct single residue substitutions in ten more proteins with PDB structural coordinate files, to produce a combined dataset of 938 distinct mutants from 20 diverse proteins: ribonuclease H2 (1io2A) [11], gene V protein (1vqbA) [12], ribonuclease Sa (1rggA) [13], P-30 protein (Onconase) (1oncA) [14], HIV-1 reverse transcriptase (1rtjA) [15], HIV-1 protease (3phvA) [16], tryptophan synthase (1wbjA) [17], Fyn tyrosine kinase (1shfA) [18], AmpC beta-lactamase (1ke4B) [19], and interleukin-3 (1jliA) [20]. The dataset is available at http://proteins.gmu.edu/automute/stability-activity.txt.

III. Results and Discussion

The initial dataset retrieved from the ProTherm Database, upon which the scatterplots of Figs. 2(a) and (b) are based, consists of 227 distinct mutants representing single residue replacements in ten proteins. We classify activity change values as either unaffected (U, ≥ 20% of WT) or affected (A, < 20% of WT), and stability change values (for all three experimental techniques) as either increased (inc, ≥ 0) or decreased (dec, < 0). A plot of the mean residual score (MRS) for each of these four mutant categories (Fig. 3) displays the same distinctive trend as that obtained for the 121 T4 lysozyme mutants. A trend with respect to stability change (MRS(inc) = 0.93, MRS(dec) = –0.68) reflects a statistically significant structure (MRS) – stability (inc/dec categories) relationship (*t*-test, $p < 0.0001$); however, as in the case of the 121 T4 lysozyme mutants, a statistically significant structure (MRS) – function (U/A categories) relationship is not evident owing to the small dataset size.

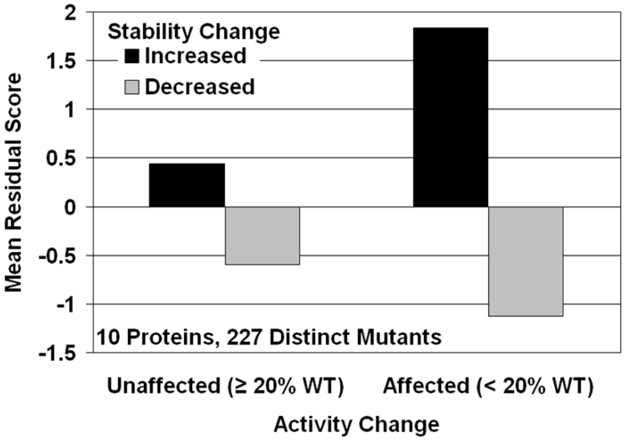
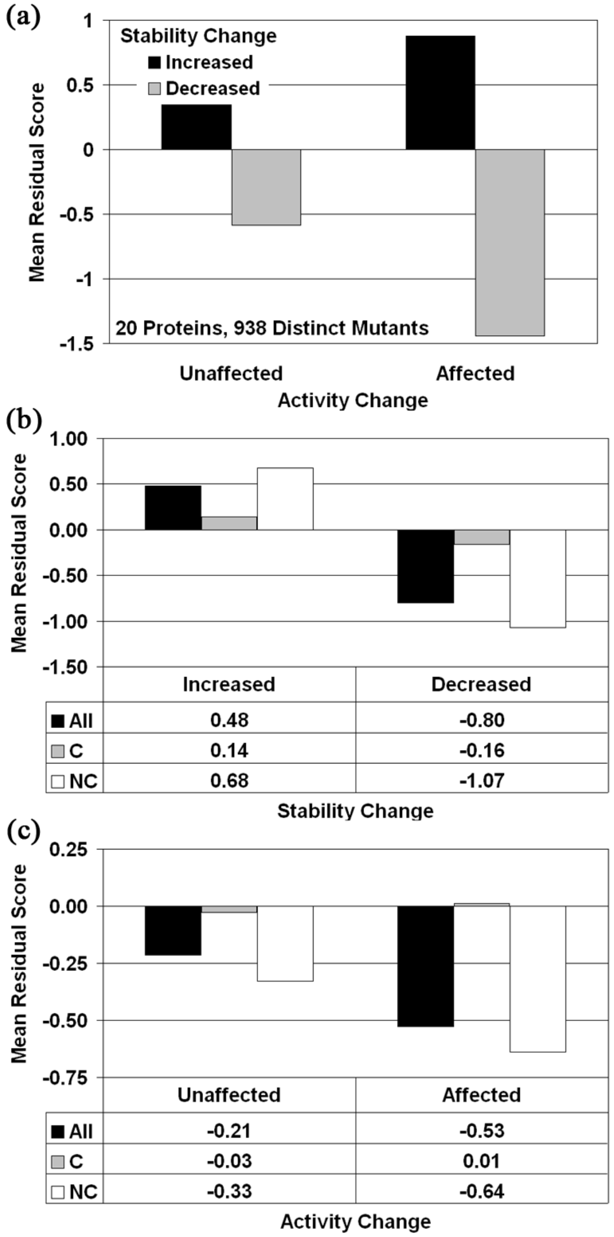


Fig. 3. The same distinctive structure – stability – activity relationship as that of Fig. 1, based on 227 distinct single residue replacements in ten proteins with experimental data retrieved from the ProTherm Database.

Next, the 711 distinct mutants from ten additional proteins that we retrieved from the literature are similarly classified and combined with the 227 ProTherm mutants, leading to a combined dataset of 938 distinct mutants with the following distribution: 279 U\inc, 421 U\dec, 94 A\inc, and 144 A\dec. Again, a plot of MRS calculated for each of the four mutant categories (Fig. 4(a)) mirrors those obtained earlier in Figs. 1 and 3. Importantly, for each of the four pairs of categories in Fig. 4(a), a statistically significant difference exists between the pairwise MRS values (*t*-test, $p < 0.01$ for each pair). Furthermore, trends evident when independently considering stability change (Fig. 4(b)) and activity change (Fig. 4(c)) reflect a statistically significant structure (MRS) – stability (inc/dec categories) relationship (*t*-test, $p < 0.0001$) as well as a statistically significant structure (MRS) – function (U/A categories) relationship (*t*-test, $p < 0.05$), respectively.

Third, the mutants comprising each category in Figs. 4(b) and (c) are further clustered based on whether they represent conservative or non-conservative replacements of the respective native residues [21], and MRS values are calculated for each of these mutant subgroups.. Note that the subset of non-conservative mutants drive the overall trends in Figs. 4(b) and (c), whereas conservative substitutions by definition minimally impact protein structure regardless of their effects on activity or stability, a trait effectively represented in the figures through MRS values of small magnitude for the conservative mutant subgroups.

Finally, we hypothesize that the distinctive plots in Figs. 1, 3, and 4(a) are a consequence of the following general



	Increased	Decreased
■ All	0.48	-0.80
▨ C	0.14	-0.16
□ NC	0.68	-1.07

	Unaffected	Affected
■ All	-0.21	-0.53
▨ C	-0.03	0.01
□ NC	-0.33	-0.64

Fig. 4. (a) Recurrence of the structure – stability – activity relationship based on a diverse set of 938 distinct single residue replacements in twenty proteins. These mutants elucidate statistically significant (b) structure – stability and (c) structure – function relationships (C / NC = conservative / non-conservative residue replacements).

principles. Residue substitutions that have a minimal impact on protein sequence-structure compatibility (MRS of small magnitude) concomitantly have a minimal effect on stability (either inc or dec) and leave activity unaffected. On the other hand, two scenarios exist as to why activity would be detrimentally affected: mutant sequence-structure compatibility (MRS) either significantly increases or decreases relative to the native protein, corresponding to mutants that are highly stable (i.e., too rigid to accommodate substrates or catalyze reactions) or highly unstable (too flexible due to lack of a sufficient noncovalent bonding network to maintain the proper fold), respectively.

References

[1] M. Masso and Vaisman, II, "AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements," Protein Eng Des Sel, vol. 23, 2010, pp. 683-687.

[2] M. Masso and Vaisman, II, "Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis," Bioinformatics, vol. 24, 2008, pp. 2002-2009.

[3] M. Masso and I. I. Vaisman, "Accurate prediction of enzyme mutant activity based on a multibody statistical potential," Bioinformatics, vol. 23, 2007, pp. 3155-3161.

[4] V. L. Thomas, A. C. McReynolds, and B. K. Shoichet, "Structural bases for stability-function tradeoffs in antibiotic resistance," J Mol Biol, vol. 396, 2010, pp. 47-59.

[5] C. Deutsch and B. Krishnamoorthy, "Four-body scoring function for mutagenesis," Bioinformatics, vol. 23, 2007, pp. 3009-3015.

[6] D. Rennell, S. E. Bouvier, L. W. Hardy, and A. R. Poteete, "Systematic mutation of bacteriophage T4 lysozyme," J Mol Biol, vol. 222, 1991, pp. 67-88.

[7] M. Masso, Z. Lu, and I. I. Vaisman, "Computational mutagenesis studies of protein structure-function correlations," Proteins, vol. 64, 2006, pp. 234-245.

[8] K. Saraboji, M. M. Gromiha, and M. N. Ponnuswamy, "Relative importance of secondary structure and solvent accessibility to the stability of protein mutants. A case study with amino acid properties and energetics on T4 and human lysozymes," Comput Biol Chem, vol. 29, 2005, pp. 25-35.

[9] K. A. Bava, M. M. Gromiha, H. Uedaira, K. Kitajima, and A. Sarai, "ProTherm, version 4.0: thermodynamic database for proteins and mutants," Nucleic Acids Res, vol. 32, 2004, pp. D120-D121.

[10] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, et al., "The Protein Data Bank," Nucleic Acids Res, vol. 28, 2000, pp. 235-242.

[11] A. Mukaiyama, M. Haruki, M. Ota, Y. Koga, K. Takano, et al., "A hyperthermophilic protein acquires function at the cost of stability," Biochemistry, vol. 45, 2006, pp. 12673-12679.

[12] T. C. Terwilliger, H. B. Zabin, M. P. Horvath, W. S. Sandberg, and P. M. Schlunk, "In vivo characterization of mutants of the bacteriophage f1 gene V protein isolated by saturation mutagenesis," J Mol Biol, vol. 236, 1994, pp. 556-571.

[13] G. I. Yakovlev, V. A. Mitkevich, K. L. Shaw, S. Trevino, et al., "Contribution of active site residues to activity and thermal stability of ribonuclease Sa," Protein Sci, vol. 12, 2003, pp. 2367-2373.

[14] U. Arnold, C. Schulenburg, D. Schmidt, and R. Ulbrich-Hofmann, "Contribution of structural peculiarities of onconase to high stability and folding kinetics," Biochemistry, vol. 45, 2006, pp. 3580-3587.

[15] J. A. Wrobel, S. F. Chao, M. J. Conrad, J. D. Merker, R. Swanstrom, et al., "A genetic approach for identifying critical residues in the fingers and palm subdomains of HIV-1 reverse transcriptase," Proc Natl Acad Sci U S A, vol. 95, 1998, pp. 638-645.

[16] B. Mahalingam, J. M. Louis, C. C. Reed, J. M. Adomat, J. Krouse, et al., "Structural and kinetic analysis of drug resistant mutants of HIV-1 protease," Eur J Biochem, vol. 263, 1999, pp. 238-245.

[17] K. Yutani, K. Ogasahara, A. Kimura, and Y. Sugino, "Effect of single amino acid substitutions at the same position on stability of a two-domain protein," J Mol Biol, vol. 160, 1982, pp. 387-390.

[18] A. A. Di Nardo, S. M. Larson, and A. R. Davidson, "The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core," J Mol Biol, vol. 333, 2003, pp. 641-655.

[19] B. M. Beadle and B. K. Shoichet, "Structural bases of stability-function tradeoffs in enzymes," J Mol Biol, vol. 321, 2002, pp. 285-296.

[20] C. J. Bagley, J. Phillips, B. Cambareri, M. A. Vadas, and A. F. Lopez, "A discontinuous eight-amino acid epitope in human interleukin-3 binds the alpha-chain of its receptor," J Biol Chem, vol. 271, 1996, pp. 31922-31928.

[21] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcut, "A model for evolutionary change in proteins," in Atlas of protein sequence and structure, vol. 5, M. O. Dayhoff, Ed. Washington D.C.: National Biomedical Research Foundation, 1978, pp. 345-352.