MOLECULAR EVOLUTIONARY STUDIES USING STRUCTURAL GENOMICS AND PROTEOMICS

by

Jinrui Xu

A dissertation submitted in partial fulfillment
of the requirements of the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2015

Doctoral Committee:

Professor Jianzhi Zhang, Chair
Assistant Professor Barry Grant
Professor Timothy D. Johnson
Associate Professor Jun Li
Associate Professor Patricia Wittkopp

In memory of my grandmother, Zheng Guilan (1936-2012)

ACKNOWLEDGEMENTS

First and foremost, my thanks go to my advisor Dr. Jianzhi Zhang. I appreciate all his contributions of time and ideas to my Ph.D. work. His guidance helped me in all the time of research and writing of this thesis. Without the guidance, I could not achieve any chapter of this thesis. Moreover, he is amazingly knowledgeable and enthusiastic in research. These properties have been contagious and motivational. He set an excellent example of scientists for me. Besides my advisor, I would like to thank the rest of my thesis committee: Drs. Barry Grant, Timothy Johnson, Jun Li and Patricia Wittkopp for their encouragement and insightful comments on my studies.

My thanks also go to all members of the Zhang lab. They contributed immensely to my professional and personal life. I would like to thank them for various helps on my research. I am thankful to Wei-Chin Ho, Bryan Moyers, Jian-Rong Yang and Zhengting Zou for comments on my papers. Particularly, I am grateful to Dr. Jian-Rong Yang for intense discussions on my projects. I am also thankful for casual discussions with many other members: Chuan Li, Brian Metzger, Calum Maclean and Xinzhu Wei. The discussions helped me better understand concepts in evolution and biology. More importantly, I appreciate friendships with the lab members. It is the companionship, consolation and encouragement from friends that made my life both inside and outside the lab enjoyable.

Last but certainly not the least, I would like to thank my wife and parents. They have been a constant source of strength and inspiration. Without their unconditional love and supports,

I would not have completed my graduate study. Especially, I am grateful to my wife Xin Xin for her devotion to the family and companionship since we first met more than 10 years ago. No words can convey my appreciation for her being my wife and the best friend.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

The field of molecular evolution has progressed with the accumulation of various molecular data. It started with the analysis of protein sequence data, followed by that of gene sequence dada and genome sequence data. In recent years, two rapidly developing areas, structural genomics and proteomics, have offered new types of data for addressing molecular evolution questions. Structural genomics refers to genome-wide collection and analysis of protein structures, whereas proteomics is the study of all proteins in a cell or organism. In this thesis, I conducted molecular evolutionary projects using data provided by structural genomics and proteomics. First, I used protein structure information to explain why some human-disease associated amino acid residues (DARs) appear as the wild-type in other species. Because destabilizing protein structures is a primary reason why DARs are deleterious, I focused on protein stability in this analysis and discovered 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 finding of compensatory amino acid substitutions in evolution has important implications for understanding epistasis in protein evolution and for using animal models of human diseases. Second, the recently published human proteomes include peptides encoded by annotated pseudogenes, which are relics of formerly functional genes. These translated pseudogenes may actually be functional and subject to purifying selection. Alternatively, their translations may be accidental and do not indicate functionality. My evolutionary analysis strongly suggests that a sizable fraction of the translated pseudogenes are subject to purifying selection acting at the protein level. Third, for the purpose of understanding protein evolution and structure-function relationships, protein structures are commonly classified according to their structure similarities. A fold encompasses protein structures with similar core topologies. Current fold classifications implicitly assume that folds are discrete islands in the protein structure space, whereas increasing evidence suggests significant similarities among folds and supports a continuous fold space. I developed a likelihood method to classify proteins into existing folds by considering the continuity in fold space. My results using this method demonstrated the growing importance of considering this continuity in fold classification. Together, my work illustrated the utility of structural genomics and proteomics in answering evolutionary questions and provided better understanding of gene and protein evolution.

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

The development of molecular evolution has been closely synchronized with the emergence of new molecular techniques and data. The advent of protein sequencing led to the invention of molecular phylogenetics (Fitch and Margoliash 1967) and the discovery of molecular clocks (Zuckerkandl and Pauling 1965). Gel electrophoresis allowed probing protein polymorphisms in populations (Lewontin and Hubby 1966). Comparing DNA sequences led to developments of statistical tests for detecting natural selection at the molecular level. Genome sequencing and comparison led to the discovery of rampant horizontal gene transfers in prokaryotes. In recent years, structural genomics and proteomics have emerged and progressed rapidly. The abundant data produced by the two fields provide enormous opportunities for the study of molecular evolution.

Structural genomics combines high-throughput experimental and modeling approaches to describe 3D structures of proteins encoded by a genome (Kim 1998; Montelione and Anderson 1999; Skolnick et al. 2000). All proteins are first organized into homology families. From each family without known structures, a representative protein is subject to experiments such as X-ray crystallography or NMR spectroscopy for solving its structure. This structure is taken as a template to model the structures of its homologous proteins. With this strategy, structural genomics produces the solved structures that cover the complete protein space, whereas traditional structural biology prefers to solve the structures with known important functions. Since 2000, a coordinated international project, protein structure initiative (PSI), has been carried out for structural genomics (Gaasterland 1998). The PSI has generated more than 5,000 protein structures so far. These structures are annotated by automatic pipelines and expert knowledge (Ellrott et al. 2011). The semi-completely described structure space renders genome-wide studies of structure-function relationship and structure evolution possible.

In molecular evolution, phylogenomics has been used to date the relative ages of protein structures (Caetano-Anolles et al. 2011). Information on ancient structures and their functions shed lights on how primordial functions evolved and interacted to give rise to expanded functional repertoires (Caetano-Anolles et al. 2012). Moreover, the structures can be used to date geological events because the appearances of structures with particular functions are found correlated with the times of corresponding geological events (Kim et al. 2012). For example, structures for aerobic respiration appeared around the Great Oxidation Event (GOE) (Kim et al. 2012), which is congruent with other evidence (Stolper et al. 2010; David and Alm 2011). Besides, structure properties such as solvent accessibility influence protein evolutionary rate substantially (Bloom et al. 2006; Franzosa and Xia 2009). Therefore, the abundant structures are useful for improving models of protein evolution.

The goal of proteomics is to achieve a quantitative description of all protein expressions and modifications in a cell, tissue, or organism, via primarily mass spectrometry (James 1997; Anderson and Anderson 1998; Blackstock and Weir 1999). Proteomic data can be used to quantify protein expressions and sequence proteins. These applications are useful for addressing some molecular evolutionary questions. For example, human proteomic data have been used to test Ohno's hypothesis, which asserts that the expression levels of X-linked genes are doubled in males to compensate the degeneration of their Y homologs (Ohno 1967). Using human proteomic data from 22 tissues, Chen and Zhang found no X-upregulation at the protein level, refuting Ohno's hypothesis (Chen and Zhang 2015). Moreover, mass spectrometry is used to sequence collagen proteins from the remains of the extinct species: *Toxondon* sp. and *Macrauchenia* sp. of South American native ungulates (SANUs). The alignment of the fossil proteins and available collagens from extant mammals resolve the evolutionary history of SANUs (Welker et al. 2015), whereas phylogenies based on morphology and ancient DNA have been proved unconvincing (Welker et al. 2015). Despite a few examples, proteomics has not been widely used in molecular evolution. However, its importance is gradually being recognized (Diz et al. 2012).

In this thesis, I used structural genomic and proteomic data to address four questions in molecular evolution. In Chapter 2, I used protein structure information to understand the enigmatic phenomenon that some human-disease associated amino acid residues (DARs) appear as wild-type residues 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 (Kondrashov et al. 2002). However, the general validity of the hypothesis remains unclear because only a few compensatory residues have been identified and tested. I mapped DARs that appear as wild-type residues in non-human species onto their protein structures, and took the residues that are spatially close to the DAR as potential compensatory residues. This is reasonable because neighboring residues can interact with, and thus have strong effects on the DARs via non-covalent interactions. I demonstrated that the potential compensatory residues mitigate the deleterious destabilizing effects of DARs, providing evidence for the compensation hypothesis at the genomic scale.

The work in Chapter 3 was inspired by a surprising observation that the human proteomes published very recently include peptides encoded by 322 annotated pseudogenes (Kim et al. 2014; Wilhelm et al. 2014). An interesting question is whether these translated pseudogenes are functional as proteins. Alternatively, the pseudogenes may be transcribed and translated by chance, indicating no functionality. The functions studied here include all biochemical and physiological functions that have been under purifying selection. That is, I used the action of purifying selection as an indicator for function. I found that a sizeable fraction of the translated pseudogenes are subject to purifying selection. This and other lines of evidence indicate that some translated pseudogenes are functional.

In Chapter 4, I developed a method to classify protein structures into existing folds, where a fold contains protein structures with similar secondary structure compositions, orientations, and connection orders (Andreeva et al. 2008; Cuff et al. 2011). Current fold classifications assume that folds are discrete islands in the structure space. However, increasing evidence suggests significant similarities among folds and supports a continuous fold space (Harrison et al. 2002; Kolodny et al. 2006). My method considers the fold space continuity in classifying a query structure into the existing folds. My classifications differ from the current classifications for 4-12% of all domains and up to 5%-20% of recently solved domains. These differences confirm the continuous nature of the fold space and demonstrate the importance of considering this continuity in fold classification.

In the addition to the main chapters, I addressed, in Appendix 1, a fundamental question in protein structure comparison: how to interpret structure similarity scores. Many structure similarity scores have been developed to gauge the similarity between two protein structures (Kabsch 1978; Holm and Sander 1995; Zemla 2003). Nevertheless, none of the scores can by itself provide information on (1) how significant the structure similarity is and (2) how likely the two structures in comparison are from the same fold. Some have tried to use Z-scores to answer the first question by measuring the deviation of a structure similarity score from similarity scores of randomly paired structures (Hasegawa and Holm 2009). However, structure similarity scores do not follow a Gaussian distribution but follow an extreme value distribution (EVD) (Levitt and Gerstein 1998). Therefore, using Z-scores of structure similarity is inappropriate. I answered the two questions for the TM-score, which is a widely used structure similarity score (Zhang and Skolnick 2004; Zhang and Skolnick 2005). For the first question, I fitted TM-scores of random structure pairs using EVD, and calculated *p*-value from the distribution for any focal TM-score. The *p*-value is the probability that the random structure pairs have TM-scores equal to or higher than the focal TM-score, indicating the significance of the focal TM-score. For the second question, I derived a posterior probability that two structures share a fold given their TM-score. The *p*-value and posterior probability make TM-scores easy to interpret and use for structure classifications.

1.2 REFERENCES

Anderson NL, Anderson NG. 1998. Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis* **19**(11): 1853-1861.

Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. 2008. Data growth and its impact on the SCOP database: new developments. *Nucleic acids research* **36**(Database issue): D419-425.

Blackstock WP, Weir MP. 1999. Proteomics: quantitative and physical mapping of cellular proteins. *Trends in biotechnology* **17**(3): 121-127.

Bloom JD, Drummond DA, Arnold FH, Wilke CO. 2006. Structural determinants of the rate of protein evolution in yeast. *Molecular biology and evolution* **23**(9): 1751-1761.

Caetano-Anolles D, Kim KM, Mittenthal JE, Caetano-Anolles G. 2011. Proteome evolution and the metabolic origins of translation and cellular life. *Journal of molecular evolution* **72**(1): 14-33.

Caetano-Anolles G, Kim KM, Caetano-Anolles D. 2012. The phylogenomic roots of modern biochemistry: origins of proteins, cofactors and protein biosynthesis. *Journal of molecular evolution* **74**(1-2): 1-34.

Chen X, Zhang J. 2015. No X-Chromosome Dosage Compensation in Human Proteomes. *Molecular biology and evolution*.

Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, Pellegrini-Calace M, Jones D, Thornton J, Orengo CA. 2011. Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic acids research* **39**(Database issue): D420-426.

David LA, Alm EJ. 2011. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469**(7328): 93-96.

Diz AP, Martinez-Fernandez M, Rolan-Alvarez E. 2012. Proteomics in evolutionary ecology: linking the genotype with the phenotype. *Molecular ecology* **21**(5): 1060-1080.

Ellrott K, Zmasek CM, Weekes D, Sri Krishna S, Bakolitsa C, Godzik A, Wooley J. 2011. TOPSAN: a dynamic web database for structural genomics. *Nucleic acids research* **39**(Database issue): D494-496.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. *Science* **155**(3760): 279-284.

Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. *Molecular biology and evolution* **26**(10): 2387-2395.

Gaasterland T. 1998. Structural genomics taking shape. *Trends in genetics : TIG* **14**(4): 135.

Harrison A, Pearl F, Mott R, Thornton J, Orengo C. 2002. Quantifying the similarities within fold space. *Journal of molecular biology* **323**(5): 909-926.

Hasegawa H, Holm L. 2009. Advances and pitfalls of protein structural alignment. *Current opinion in structural biology* **19**(3): 341-348.

Holm L, Sander C. 1995. Dali: a network tool for protein structure comparison. *Trends in biochemical sciences* **20**(11): 478-480.

James P. 1997. Protein identification in the post-genome era: the rapid rise of proteomics. *Quarterly reviews of biophysics* **30**(4): 279-331.

Kabsch W. 1978. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **34**(5): 827-828.

Kim KM, Qin T, Jiang YY, Chen LL, Xiong M, Caetano-Anolles D, Zhang HY, Caetano-Anolles G. 2012. Protein domain structure uncovers the origin of aerobic metabolism and the rise of planetary oxygen. *Structure* **20**(1): 67-76.

Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S et al. 2014. A draft map of the human proteome. *Nature* **509**(7502): 575-581.

Kim SH. 1998. Shining a light on structural genomics. *Nature structural biology* **5 Suppl**: 643-645.

Kolodny R, Petrey D, Honig B. 2006. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Current opinion in structural biology* **16**(3): 393-398.

Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. *Proceedings of the National Academy of Sciences of the United States of America* **99**(23): 14878-14883.

Levitt M, Gerstein M. 1998. A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of Sciences of the United States of America* **95**(11): 5913-5920.

Lewontin RC, Hubby JL. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics* **54**(2): 595-609.

Montelione GT, Anderson S. 1999. Structural genomics: keystone for a Human Proteome Project. *Nature structural biology* **6**(1): 11-12.

Ohno S. 1967. *Sex chromosomes and sex-linked genes*. Springer Science & Business Media.

Skolnick J, Fetrow JS, Kolinski A. 2000. Structural genomics and its importance for gene function analysis. *Nature biotechnology* **18**(3): 283-287.

Stolper DA, Revsbech NP, Canfield DE. 2010. Aerobic growth at nanomolar oxygen concentrations. *Proceedings of the National Academy of Sciences of the United States of America* **107**(44): 18755-18760.

Welker F, Collins MJ, Thomas JA, Wadsley M, Brace S, Cappellini E, Turvey ST, Reguero M, Gelfo JN, Kramarz A et al. 2015. Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. *Nature*.

Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* **509**(7502): 582-587.

Zemla A. 2003. LGA: A method for finding 3D similarities in protein structures. *Nucleic acids research* **31**(13): 3370-3374.

Zhang Y, Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**(4): 702-710.

-. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**(7): 2302-2309.

Zuckerkandl E, Pauling L. 1965. Molecules as documents of evolutionary history. *Journal of theoretical biology* **8**(2): 357-366.

CHAPTER 2

WHY HUMAN DISEASE-ASSOCIATED RESIDUES APPEAR AS THE WILD-TYPE IN OTHER SPECIES: GENOME-WIDE STRUCTURAL EVIDENCE FOR THE COMPENSATION HYPOTHESIS

2.1 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 1077 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.

2.2 INTRUDUCTION

It was first reported 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 non-human 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 non-human species and test whether the potential compensatory residues in the spatial neighborhood of the DARs mitigate the destabilizing effects of the DARs in the non-human species.

2.3 RESULTS

2.3.1 Protein stability reduction caused by DARs

We began with 51,920 DARs from the Human Gene Mutation Database (HGMD) (Stenson et al. 2003) and Universal Protein Resource (UniProt) (The_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, 1077 appear as the wild-type in the one-to-one orthologous proteins of at least one non-human species (Altenhoff et al. 2011) and thus are called wt-DARs. Although wt-DARs are often referred to as compensated pathogenic deviations (CPDs) (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 8135 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 ΔΔG is, the bigger the stability reduction is. Thus, ΔΔG is referred to as the stability reduction upon mutation. The median ΔΔG 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 **(Fig. A.2.1.1**). This amount is significantly smaller than the median ΔΔG (4.09 REU or ~2.25 kcal/mol) for mutations to rg-DARs ($p < 10^{-41}$, Mann-Whitney U test; **Fig. 2.1**), 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 vs. 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 dataset (**Table A.2.1.1**). Among these 128 types, 13 showed a significantly smaller median ΔΔG for mutations to wt-DARs than mutations to rg-DARs ($p < 0.05$, Mann-Whitney U test; **Table A.2.1.1**), while 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 (**Fig. A.2.1.2**), 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 non-human species has a limited power. As a comparison, we also computed the average ΔΔG 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 ΔΔG (median = 0.47 REU or ~0.26 kcal/mol) is significantly lower than that for wt-DARs ($p < 10^{-14}$; **Fig. 2.1**).

2.3.2 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 non-covalent bonds. Furthermore, it is currently infeasible to examine the potential compensatory effects of a large number of residues simultaneously, while 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 4Å from a focal residue to be its neighboring residues, where the distance between two residues is measured by the shortest spatial distance of their non-hydrogen 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. 2.2**). 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 (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. 2.2A**). 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. 2.2B**). Consistent with the compensation hypothesis, $\Delta\Delta G_2$ (-4.43 REU or ~-2.43 kcal/mol) is 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. 2.2C**). 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. 2.3**). 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 ~0.56 kcal/mol ($p < 10^{-10}$, **Fig. 2.3**). 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; **Fig. A.2.1.3**).

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 wild-type residues 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 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. 2.4**), whereas $\Delta\Delta G_3$ is not significantly different from $\Delta\Delta G_1$ ($P > 0.5$; **Fig. 2.4**). Furthermore, $\Delta\Delta G_2$ is significantly smaller than $\Delta\Delta G_3$ ($P < 0.01$; **Fig. 2.4**). 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.

2.3.3 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. 2.5A**). By contrast, no such effect for WTRLs is detectable ($p > 0.2$; **Fig. 2.5B**).

2.4 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 crystalized, while 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 Fig. 2.3 can be repeated by this subset of data (**Fig. A.2.1.4**), 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 ΔΔG are not always accurate (Kellogg et al. 2011), which limits the statistical power of our 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. While the actual compensation may be larger if some compensatory residues are outside the 4Å 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 ΔΔG 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 ΔG = -3 kcal/mol has >99% of molecules folded under 37°C (see Materials and Methods). Upon mutation to an average destabilizing wt-DAR (ΔΔG = 3.54 kcal/mol), folded protein molecules drop to 30% (ΔG = 0.54 kcal/mol). With the help of the detected median compensatory effect (ΔΔG = -0.56 kcal/mol), the fraction of folded molecules rises to 51% (ΔG = -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 ΔΔG, 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.

2.5 MATERIALS AND METHODS

2.5.1 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 0.1Å, between 0.1 and 0.2Å, 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 non-redundant protein structures from the protein structure database CATH (Sillitoe et al. 2013). The density peaks at 1.4 and 3.3 Å (**Fig. A.2.1.5**), representing residue pairs in contact via N-O and hydrogen bonds, respectively. The density drops drastically and appears uniformly distributed at spatial distances above 4Å. 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 4Å 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 ~3.0Å, 2.6-3.5Å, and averaging 3.6Å between two non-hydrogen atoms, respectively. Therefore, we identify potential compensatory residues within the 4Å radius.

2.5.2 Protein structures

Human protein structures were downloaded from PDB (Berman 2008), while the SIFTS database (Velankar et al. 2013) was used to map the structures with corresponding proteins in UniProt (The_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 (OMA) 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 1077 human DARs appear as the wild-type in at least one non-human 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 correspond 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 1008 wt-DARs have at least one set of potential compensatory residues.

Human single amino acid polymorphisms (SAAPs) were acquired from UniProt. SAAPs were cross-linked to their single nucleotide polymorphisms (SNPs) in dbSNP where the minor allele frequencies (MAFs) in humans were obtained. Only SAAPs with MAFs ≥ 0.01 were used.

2.5.3 Prediction of ΔΔG

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 7Å rather than 8Å to the site of the point mutation. The 7Å in C_α distance was chosen because we found it corresponds to 4Å 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.

2.5.4 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).

2.5.5 Data availability

All data used in this study can be obtained at http://www.umich.edu/~zhanglab/download/Jinrui_MBE_Suppl/index.htm.

2.6 ACKNOWLEDGEMENTS

We thank Wei-Chin Ho, Jian-Rong Yang, and three anonymous reviewers for valuable comments. This work was supported in part by research grant GM103232 from the U.S. National Institutes of Health to J.Z.

Figure 2.1 Frequency distribution of human protein stability reduction upon mutation. The stability reductions are caused by mutations to human single amino acid polymorphisms (SAAPs) with minor allele frequencies (MAF) > 0.01 (black), disease-associated residues that appear as the wild-type in at least one non-human species (wt-DARs) (green), and other disease-associated residues (rg-DARs) (red). The samples include 482 SAAPs, 1077 wt-DARs, and 8124 of the 8135 rg-DARs (11 rg-DARs are not included because Rosetta failed to complete the computations in 72 hours), respectively. Protein stability reduction is expressed in kcal/mol estimated from Rosetta Energy Unit (REU) by linear regression (Fig. A.2.1.1). Arrows indicate median values of the distributions. The three distributions are all significantly different from one another ($p < 10^{-14}$, Mann-Whitney U Test).

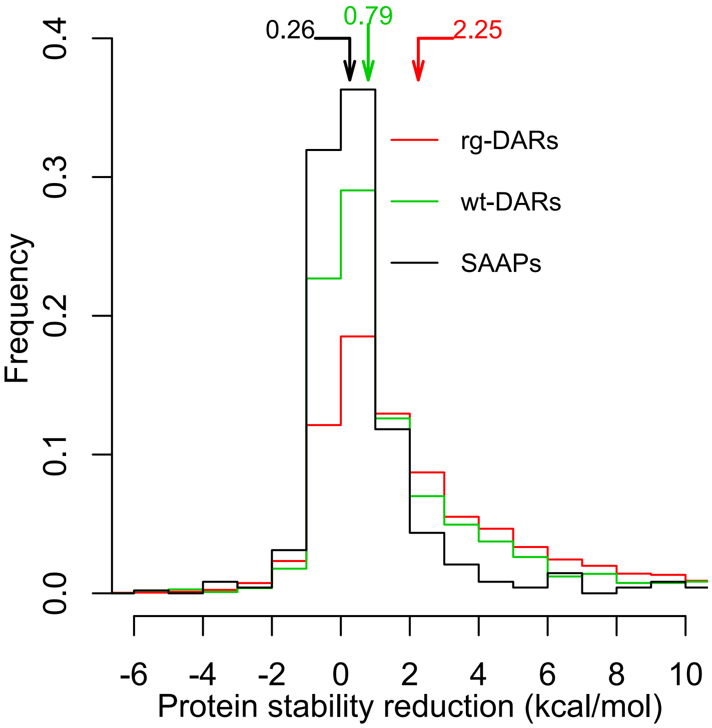


Figure 2.2 Testing the compensation hypothesis for the disease-associated residue (DAR) at position 532 of human plasminogen (UniProt accession number: P00747). The DAR site and its orthologous site in non-human species are squared, and the DAR is shaded. Spatial neighbors of the DAR site, shown as circles, 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 Fig. A.2.1.6.

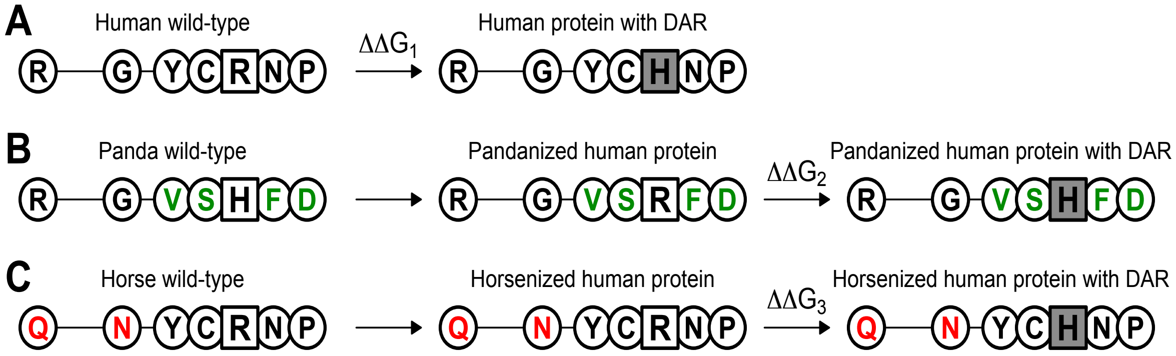


Figure 2.3 Frequency distribution of the difference in protein stability reduction upon mutation from a human wild-type residue (WTR) to a disease-associated residue (DAR) in the absence ($\Delta\Delta G_1$) and presence ($\Delta\Delta G_2$) of neighboring residues from a species where the DAR is the 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.

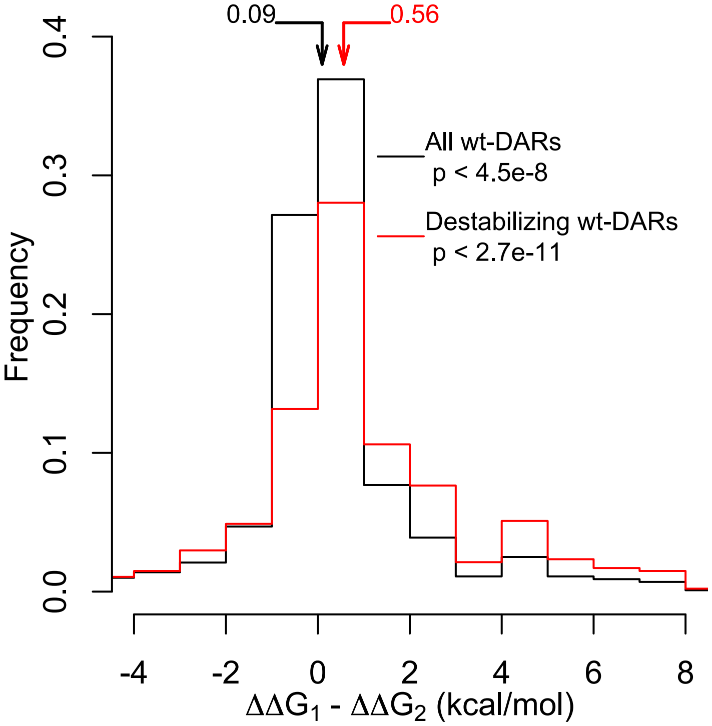


Figure 2.4 Frequency distribution of the difference in protein stability reduction upon mutation from a human wild-type residue (WTR) to a destabilizing disease-associated residue (DAR) among various genetic backgrounds. $\Delta\Delta G_1$, in the human background (see Fig. 2.2A); $\Delta\Delta G_2$, in the presence of neighboring residues from a species where the DAR is the wild-type (see Fig. 2.2B); $\Delta\Delta G_3$, in the presence of neighboring residues from a non-human species where the human WTR is the wild-type (see Fig. 2.2C). The *p*-values are from one-tail Wilcoxon signed-rank test. A total of 314 pairs of WTRs and destabilizing DARs are examined.

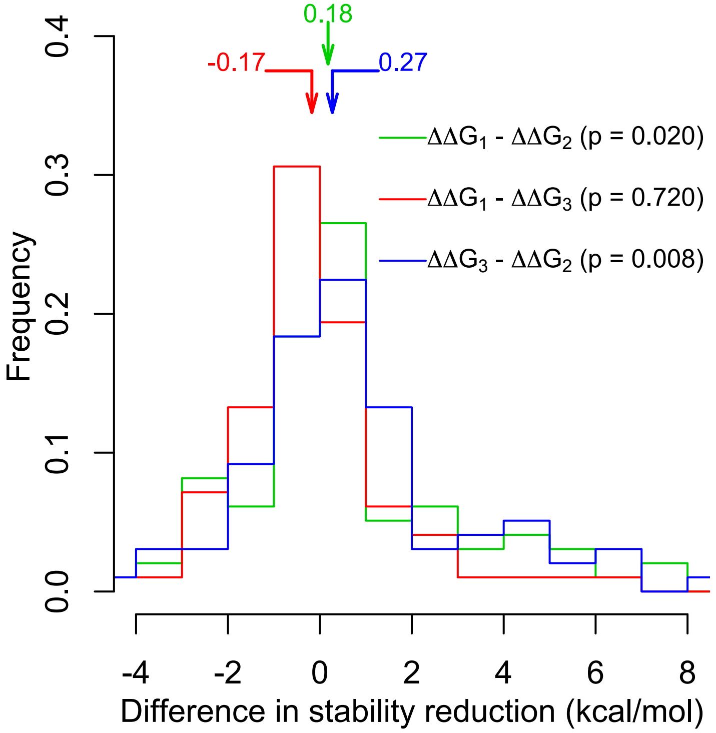
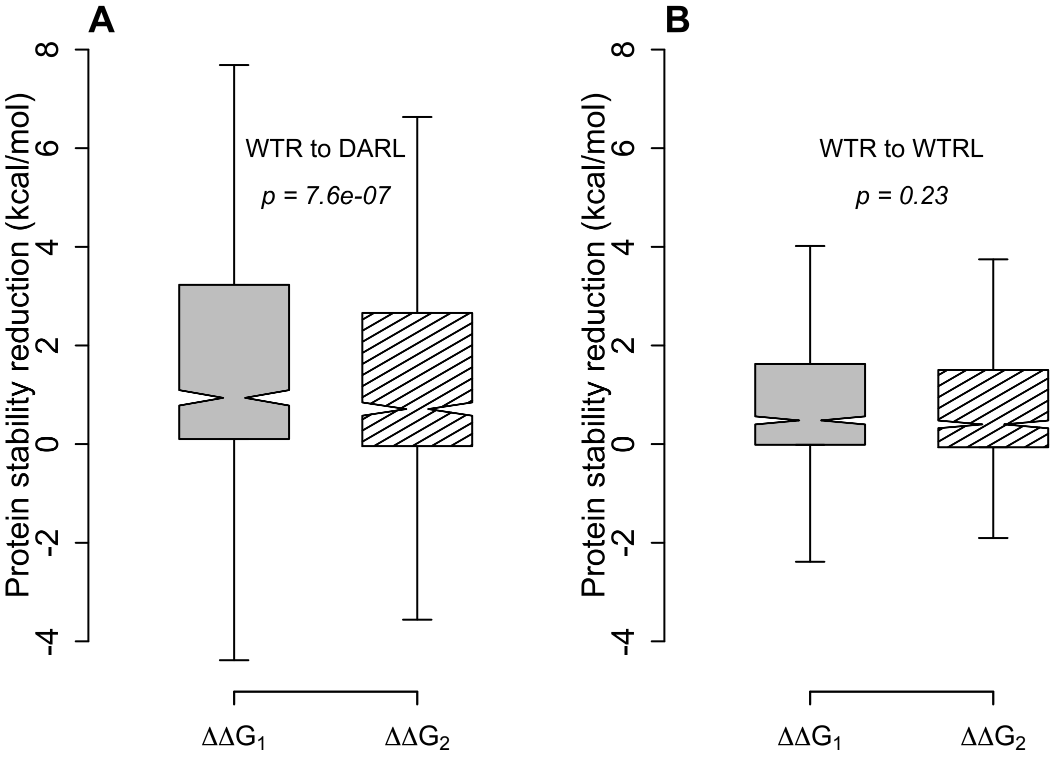


Figure 2.5 Protein stability reduction upon mutation to a residue that is physicochemically similar to DAR or WTR. (**A**) Distribution of protein stability reduction upon mutation from a human wild-type residue (WTR) to a residue that is physicochemically similar to a disease-associated residue (DARL) in the absence ($\Delta\Delta G_1$, grey bar) and presence ($\Delta\Delta G_2$, striped bar) of neighboring residues from a species where the disease-associated residue (DAR) is the wild-type. (**B**) Distribution of protein stability reduction upon mutation from a human wild-type residue (WTR) to a residue that is physicochemically similar to the WTR (WTRL) in the absence ($\Delta\Delta G_1$, grey 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.



2.7 REFERENCES

Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C 2011. OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res 39: D289-294.

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 Sping, 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: D490-498.

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.

The_UniProt_Consortium 2011. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res 39: D214-219.

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.

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: D483-489.

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.

CHAPTER 3

ARE HUMAN TRANSLATED PSEUDOGENES FUNCTIONAL?

3.1 ABSTRACT

Pseudogenes are relics of former genes that no longer possess biological functions. Operationally, they are identified based on disruptions of open reading frames (ORFs) or presumed losses of promoters. Intriguingly, two recent human proteomic studies reported peptides encoded by 322 pseudogenes. These peptides may play previously unrecognized physiological functions. Alternatively, they may have resulted from accidental translations of pseudogene transcripts and possess no function. Comparing between human and macaque orthologs, we show that the nonsynonymous to synonymous substitution rate ratio (ω) is significantly smaller for translated pseudogenes than other pseudogenes. In particular, 15% of translated pseudogenes have ω values significantly lower than 1, indicative of the action of purifying selection. This and other findings provide unambiguous evidence that some but not all translated pseudogenes have selected functions at the protein level. Hence, neither ORF disruption nor evidence for translation disproves or proves gene functionality.

3.2 INTRODUCTION

The term “pseudogene” was coined by Jacq and colleagues to describe a DNA sequence that resembles a gene coding for the frog 5S ribosomal RNA but contains mutations rendering its product nonfunctional (Jacq et al. 1977). Since then, “pseudogene” has been used to denote gene relics that no longer encode functional products. Pseudogenes have been identified abundantly in human and many other genomes (Karro et al. 2007).

Most pseudogenes originate from duplicate copies of functional genes. They are referred to as unprocessed or processed pseudogenes, depending on whether the duplication is DNA mediated or RNA mediated (Podlaha and Zhang 2010). A functional gene may also become a pseudogene without duplication, if its function no longer confers a fitness advantage to the organism due to a change in the environment or genetic background. Such pseudogenes are called unitary pseudogenes (Zhang et al. 2010). Because it is difficult to prove the lack of biological function for a segment of DNA, a pseudogene is operationally defined by its homology to a functional gene yet the presence of signs of non-functionality. The most obvious sign of non-functionality is a disruption of the canonical open reading frame (ORF) that exists in a homologous functional gene. Because RNA-mediated gene duplication only copies the transcribed region of a gene, the duplicate lacks the original promoter and is most likely "dead-on-arrival" (Podlaha and Zhang 2010). Thus, RNA-mediated duplicates, which typically lack introns that exist in their parental genes, are commonly considered pseudogenes. Based on these operational criteria, numerous pseudogenes have been annotated in sequenced genomes (Karro et al. 2007; Podlaha and Zhang 2010).

Because the operational definition of pseudogene does not require a definitive proof of non-functionality, claims of functionality have been made a number of times for operationally defined pseudogenes especially when they are transcribed. For example, human *PTENP1*, a highly transcribed pseudogene originating from a copy of the tumor suppressor gene *PTEN*, competes with *PTEN* for the microRNAs that normally suppress *PTEN* expression and *PTENP1* tends to be lost in cancer patients compared with healthy controls (Poliseno et al. 2010). But because biochemical activities may have no fitness benefit, proof of a true biological function requires the demonstration that the activity or the pseudogene is under natural selection. No such proof has been given in the case of *PTENP1*. In another example, mouse pseudogene *Makorin1-p1* was shown to regulate its parental gene (Hirotsune et al. 2003) and be under purifying selection (Podlaha and Zhang 2004). But subsequent studies questioned the validities of both the functional data (Gray et al. 2006) and evolutionary data (Kaneko et al. 2006). More recently, an evolutionary genomic analysis of human transcribed pseudogenes that have macaque orthologs found a small yet significant decrease in human-macaque sequence divergence in transcribed pseudogene regions, compared with corresponding flanking regions, suggesting that some transcribed pseudogenes are under purifying selection (Khachane and Harrison 2009). But it is unknown how many transcribed pseudogenes have selected functions.

Very recently, two human proteomic studies reported peptides encoded by 322 human pseudogenes (Kim et al. 2014; Wilhelm et al. 2014). These peptides may signal pseudogene function at the protein level, a rarely considered possibility. Alternatively, they may have resulted from spurious translations and indicate no protein function. We here distinguish between these two hypotheses by comparing the nonsynonymous/synonymous substitution rate ratio (ω) between translated pseudogenes and other pseudogenes based on human-macaque orthologs.

3.3 RESULTS

3.3.1 Detecting purifying selections of translated pseudogenes

We subjected 15,343 human pseudogenes annotated in Ensembl (version 78) to a bioinformatics pipeline to acquire a set of 78 human-macaque orthologous pseudogenes that encode peptides on the basis of human proteomic data (see Materials and Methods). For comparison, we acquired a set of 644 human-macaque orthologous pseudogenes that are transcribed (but have no proteomic hit) in humans and a set of 1455 human-macaque orthologous pseudogenes that are not transcribed (and have no proteomic hits) in humans (see Materials and Methods).

We estimated ω for the ORF region of each human-macaque orthologous pseudogene alignment (see Materials and Methods). The median ω of the translated pseudogenes is 0.70, significantly lower than that (0.90) of the transcribed pseudogenes ($P = 0.04$, Mann-Whitney U test; **Fig. 3.1**) and that (0.88) of non-transcribed pseudogenes ($P = 0.01$; **Fig. 3.1**), whereas the latter two groups have similar ω ($P = 0.21$; **Fig. 3.1**).

Some annotated pseudogenes exist in the genome regions of other genes referred to as host genes. Such pseudogenes are transcribed with the host genes, and thus may be translated as part of the protein product of the host genes. We removed the 27 translated pseudogenes that overlap with coding genes, and found the median ω of the remaining translated pseudogenes to be still 0.70 and significantly lower than that of transcribed pseudogenes ($P = 0.05$; **Fig. 3.1**), indicating that the relatively low ω of translated pseudogenes is not due to hitchhiking pseudogenes.

We found the median ORF length of the translated pseudogenes to be 369 nucleotides, significantly greater than that (333) of the transcribed pseudogenes ($P = 0.006$), supporting the notion that the coding capacity is selectively maintained in some translated pseudogenes.

Perhaps not surprisingly, the median ω of the translated pseudogenes (0.70) is substantially larger than that (0.11) of their parental genes ($P < 7\times10^{-13}$). This high median ω may be because the translated pseudogenes are subject to weaker purifying selection, or because only a subset of them is subject to purifying selection. We found that only 15% of the 78 translated pseudogenes have ω values significantly lower than 1 (nominal $P < 0.05$, likelihood ratio test). These translated pseudogenes have a median ω of 0.27. The other 85% of translated pseudogenes have a median ω of 0.85, which is not significantly lower than that of transcribed pseudogenes ($P = 0.38$; **Fig. 3.1**), suggesting that pseudogene translation does not indicate functionality in most cases.

In the 12 translated pseudogenes with ω significantly lower than 1, we found that, when the original ORF is disrupted by a premature stop codon, the pseudogene can exploit another in-frame start codon to circumvent the premature stop codon. The resultant protein is shortened but contains at least one complete or partial protein domain. For example, *FUNDC2P2* is a pseudogene of a duplicate of *FUNDC2* (FUN14 domain containing 2). In the pseudogene transcript, a premature stop codon appears downstream of the original start codon, which would result in a truncated peptide of 24 residues (**Fig. 3.2**). Interestingly, a peptide identified in the proteomic data is uniquely mapped to the transcript sequence after the premature stop codon. An alternative ORF that starts with an in-frame ATG closely following the premature stop codon could code for a protein that contains the identified peptide (**Fig. 3.2**). Thus, this in-frame ATG is likely the alternative start codon for the transcript. The protein encoded by the alternative ORF is 81% the length of the parental protein and contains the complete FUN14 domain of the parental protein, suggesting that it carries a similar molecular function. Furthermore, the human *FUNDC2P2* has orthologs from all species surveyed: chimpanzee, gorilla, orangutan, macaque, and mouse, suggesting that it has been maintained by natural selection in genome evolution.

3.3.2 Preferred types of pseudogenes to transcribe and translate

The 1,455 non-transcribed pseudogenes encompass more processed pseudogenes than unprocessed pseudogene, with the ratio equal to 5.7. This ratio is significantly higher than that (4.3) in the 644 transcribed pseudogenes ($P = 0.02$, Fisher exact test). This is expected because the processed pseudogenes lost their original promoters, and thus unlikely to transcribe. Interestingly, the processed to unprocessed pseudogene ratio (12) in the 78 translated pseudogenes is significantly higher than those in the transcribed ($P = 0.01$, Fisher exact test) and non-transcribed ($P = 0.05$) pseudogenes. Moreover, 11 of the 12 translated pseudogenes with ω significantly lower than 1 are processed pseudogenes. We apply this analysis on all shared pseudogenes, and observe similar results. Given transcription, a processed pseudogene is more likely than an unprocessed pseudogene to be translated, probably because the translational product of the former is more likely to be beneficial or less likely to be deleterious than that of the latter, due to the interference of potentially mis-spliced exons/introns in the latter.

3.3.3 Tissue specific expression of translated pseudogenes

We found that each of the 78 translated pseudogenes have peptides identified from on average two tissues (including cell lines) out of 157 tissues surveyed in the human proteomic data. The corresponding number (118) is much larger for their parental genes. Furthermore, the protein expression tissues of each translated pseudogene are a subset of those of its parental gene. The translated pseudogenes appear in 140 tissues in total (a tissue is counted as many times as the number of pseudogenes found translated in the tissue), including 13 times in testis and 127 times in other tissues. This ratio of 13/127 = 0.1 is significantly greater than the corresponding ratio (0.02) for their parental genes ($P < 10^{-4}$, Fisher's exact test). A similar ratio of 0.1 was found among the translated pseudogenes with ω significantly smaller than 1. Given the enrichment of processed genes in translated pseudogenes, the preferred translation of pseudogenes in testis can be explained by the hyper-transcription hypothesis, which states that in haploid germ cells of the testis, abundant RNA polymerase II complexes and an overall permissive chromatin promote widespread gene expression (Schmidt 1996; Soumillon et al. 2013).

3.4 DISCUSSION

In summary, our evolutionary analysis showed that human translated pseudogenes have significantly lower ω values than transcribed or non-transcribed pseudogenes. About 15% of translated pseudogenes have ω values significantly smaller than 1, suggesting that they possess selected functions at the protein level. But the rest of them have ω values similar to transcribed or non-transcribed pseudogenes, suggesting that most if not all of them likely possess no selected function at the protein level. Therefore, while a small fraction of translated pseudogenes have selected functions, translation per se is not a guarantee of functionality.

In the translated pseudogenes, processed pseudogenes are more enriched than unprocessed pseudogenes. A potential reason may be that translation of unprocessed pseudogens is more deleterious than that of processed pseudogenes, and thus purged out quickly. Transcripts of unprocessed pseudogenes tend to be truncated and retain introns, and thus their protein products are usually unfolded, which is non-functional or even toxic. Moreover, newly born unprocessed pseudogenes tend to carry their parental promoters, and thus are often highly expressed together with their parental genes. The highly expressed gene copy is very likely to be deleterious and need to be purged out.

We find that the translated pseudogenes are enriched in testis. Because the majority of translated pseudogenes are processed, the phenomenon may be explained by the hyper-transcription hypothesis (Schmidt 1996; Soumillon et al. 2013). Furthermore, according to the out-of-testes hypothesis (Kaessmann 2010), normally non-transcribed genes tend to express first there. And then with beneficial products, the testis-expressed genes are preserved and evolve efficient promoters to be expressed in other tissues and function there. Therefore, the translated pseudogenes expressed in the testis with no significant purifying selection detected may be in the out-of-testes process and stand a chance to be new functional genes.

3.5 MATERIALS AND METHODS

3.5.1 Genome, transcriptome, and proteome data

Human (hg38) and macaque (rhsMac3) genome sequences and exon coordinates were obtained from the UCSC genome browser (Rosenbloom et al. 2015). RNA-Seq data of all human genes in 16 tissues were downloaded from the human body map (Petryszak et al. 2014). Human pseudogenes and their peptides identified by mass spectrometry were collected from two human proteome drafts (Kim et al. 2014; Wilhelm et al. 2014).

3.5.2 Orthologous pseudogene identification, sequence alignment, and ω estimation

Human pseudogenes were obtained from Ensembl version 78 (Cunningham et al. 2015), including gene coordinates and pseudogene transcripts, which were annotated but not necessarily transcribed. From 15,343 annotated human pseudogenes, we removed 69 polymorphic and 226 immunity-related pseudogenes. The polymorphic pseudogenes have intact alleles in some human individuals and therefore were excluded. We removed immunity-related (i.e., immunoglobulin or T-cell receptor) pseudogenes because they may be subject to positive rather than negative selection when functional. For each human pseudogene, its syntenic region in macaque was identified in the LiftOver Browser (Kent et al. 2003). In parallel, the human pseudogene transcript was searched against the macaque genome using BLASTN (Altschul et al. 1990). The resulting high-scoring segment pairs (HSPs) that overlap the macaque syntenic region but not any coding exon were considered as orthologous exons of the human pseudogenes. These macaque exons were tilted up to the human transcript following the BLAST alignment. A total of 8,070 human pseudogenes were found to have macaque orthologs.

We first aligned human and macaque orthologous pseudogene transcripts using ClustalW (Larkin et al. 2007). If the human transcript had peptide hits in the proteomic data, the longest ORF that codes for the peptide was identified as the coding ORF. If there was no peptide hit, the longest ORF was chosen as the potential coding ORF. In the coding ORF alignment, stop codons and codons with gaps were considered interruptive codons. The aligned codons between the human start codon and the first interruptive codon in the alignment were considered as the coding region for the pseudogene. The likelihood-based CODEML program (Yang 2007) was used to calculate ω for this region. As for the parental gene of a pseudogene, its CODEML-derived estimate of ω based on human and macaque orthologs was obtained from Ensembl. The parental gene was defined as the human functional gene with the lowest E-value to the human pseudogene by BLAST.

3.5.3 Datasets of translated, transcribed, and non-transcribed pseudogenes

Kim et al. identified peptides encoded by 107 pseudogenes annotated in Ensembl version 78 (Kim et al. 2014) and Wilhelm et al. identified peptides encoded by 241 pseudogenes

(Wilhelm et al. 2014). However, the two data sets have only 26 pseudogenes in common. This small overlap may be due to a high false negative rate caused by low concentrations of pseudogene-encoded peptides. This is particularly likely in the present case because the two proteomic datasets were generated using different tissue samples, pipelines, and protocols. Additionally, the small overlap may also signal a high false positive rate. For instance, we found that, for 122 cases, no ORF in any pseudogene matched the identified peptides, and thus excluded them from further analyses. We combined the remaining pseudogenes in the two datasets and regarded the 200 unique pseudogenes as the translated pseudogenes. Only 135 of the 200 human pseudogenes have macaque orthologs. The alignments with 100% sequence identity or with fewer than 30 codons were removed because ω cannot be estimated reliably. Occasionally, a pseudogene may have multiple transcripts and thus multiple alignments. The longest alignment was chosen for analysis. The procedure above resulted in 78 translated pseudogenes with qualified alignments for further analyses.

Because transcriptions of pseudogenes may be tissue-specific, we used the human body map data to identify transcribed pseudogenes. The human body map provided mRNA-Seq profiles of all genes in Ensembl across 16 human tissues. We followed the literature to use FPKM ≥ 1 as a criterion for expression (Blazie et al. 2015). We found that 1,164 of the 8,070 shared pseudogenes have FPKM ≥ 1 in at least one of the 16 tissues, but without peptide hits in the two human proteomic datasets. These pseudogenes are referred to as transcribed pseudogenes. We assumed the longest ORF as the conceptually coding ORF in each pseudogene transcript, and then applied the same procedure used for translated pseudogenes to generate codon alignments for these genes between human and macaque. This ended up with 644 transcribed pseudogenes with codon alignments.

To generate non-transcribed pseudogenes, we identified 2,576 shared pseudogenes that have 0 FPKM in each of the 16 tissues and no peptide hit reported in the two human proteomes. From these pseudogenes, 1,455 of them have qualified codon alignments and are subject to further analyses.

3.6 ACKNOWLEDGEMENTS

This work was supported in part by U.S. National Institutes of Health grant R01GM103232 to J.Z.

Figure 3.1 Comparison of nonsynonymous to synonymous rate ratio among translated, transcribed and non-transcribed pseudogenes.

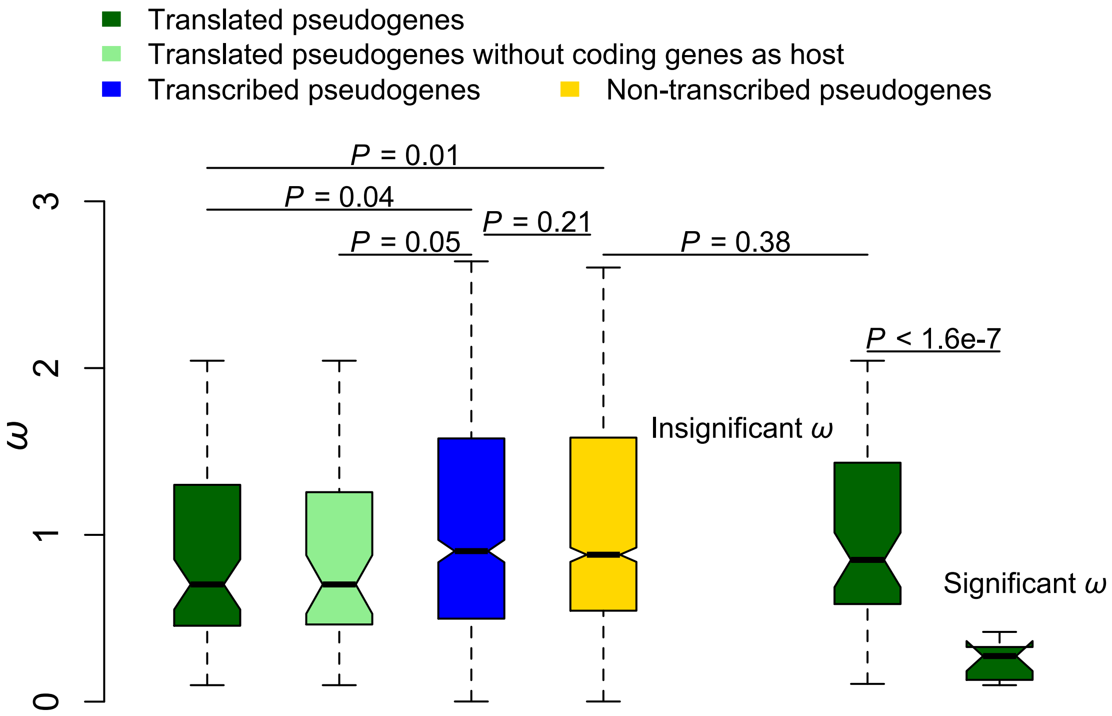
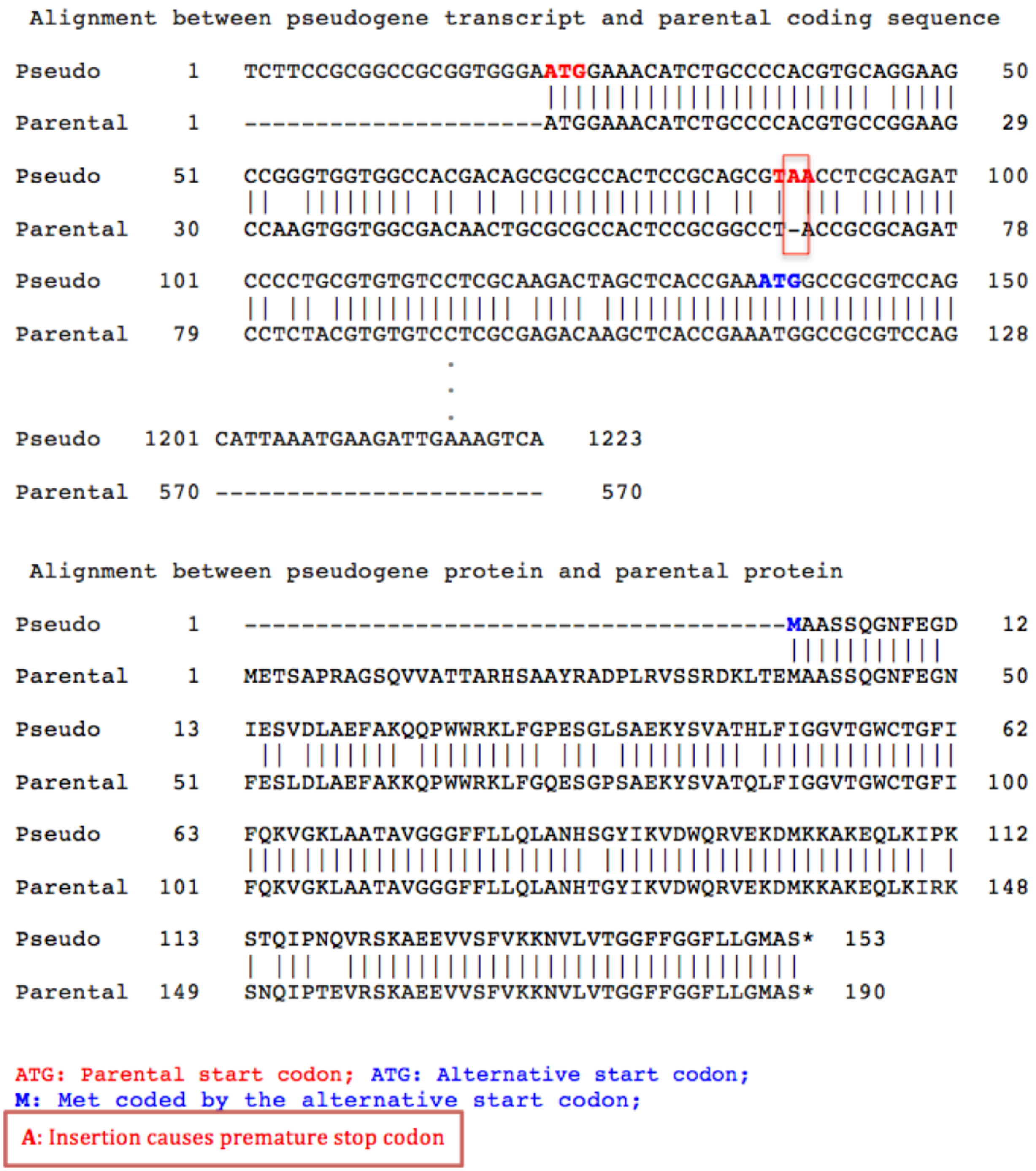


Figure 3.2 An example of translated pseudogene using alternative start codon to circumvent premature stop codon. *FUNDC2* (FUN14 domain containing 2) is the parental gene. *FUNDC2P2* is the pseudogene.



```
 Alignment between pseudogene transcript and parental coding sequence

Pseudo      1    TCTTCCGCGGCCGCGGTGGGAATGGAAACATCTGCCCCACGTGCAGGAAG    50
                                      ||||||||||||||||||||||| |||||
Parental    1    ---------------------ATGGAAACATCTGCCCCACGTGCCGGAAG    29

Pseudo     51    CCGGGTGGTGGCCACGACAGCGCGCCACTCCGCAGCGTAACCTCGCAGAT   100
                 ||  |||||||| || || |||||||||||||| || | ||| |||||||
Parental   30    CCAAGTGGTGGCGACAACTGCGCGCCACTCCGCGGCCT-ACCGCGCAGAT    78

Pseudo    101    CCCCTGCGTGTGTCCTCGCAAGACTAGCTCACCGAAATGGCCGCGTCCAG   150
                 || || ||||||||||||| |||| |||||||||||||||||||||||||
Parental   79    CCTCTACGTGTGTCCTCGCGAGACAAGCTCACCGAAATGGCCGCGTCCAG   128
                                   .
                                   .
                                   .
Pseudo   1201 CATTAAATGAAGATTGAAAGTCA   1223

Parental  570 -----------------------    570


 Alignment between pseudogene protein and parental protein

Pseudo      1    --------------------------------------MAASSQGNFEGD    12
                                                       |||||||||||
Parental    1    METSAPRAGSQVVATTARHSAAYRADPLRVSSRDKLTEMAASSQGNFEGN    50

Pseudo     13    IESVDLAEFAKQQPWWRKLFGPESGLSAEKYSVATHLFIGGVTGWCTGFI    62
                  || ||||||| ||||||||| ||| ||||||||| ||||||||||||||
Parental   51    FESLDLAEFAKKQPWWRKLFGQESGPSAEKYSVATQLFIGGVTGWCTGFI   100

Pseudo     63    FQKVGKLAATAVGGGFFLLQLANHSGYIKVDWQRVEKDMKKAKEQLKIPK   112
                 |||||||||||||||||||||||| ||||||||||||||||||||||| |
Parental  101    FQKVGKLAATAVGGGFFLLQLANHTGYIKVDWQRVEKDMKKAKEQLKIRK   148

Pseudo    113    STQIPNQVRSKAEEVVSFVKKNVLVTGGFFGGFLLGMAS*  153
                 | |||  ||||||||||||||||||||||||||||||||
Parental  149    SNQIPTEVRSKAEEVVSFVKKNVLVTGGFFGGFLLGMAS*  190


ATG: Parental start codon; ATG: Alternative start codon;
M: Met coded by the alternative start codon;
A: Insertion causes premature stop codon
```

3.7 REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* **215**(3): 403-410.

Blazie SM, Babb C, Wilky H, Rawls A, Park JG, Mangone M. 2015. Comparative RNA-Seq analysis reveals pervasive tissue-specific alternative polyadenylation in Caenorhabditis elegans intestine and muscles. *BMC biology* **13**(1): 4.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S et al. 2015. Ensembl 2015. *Nucleic acids research* **43**(Database issue): D662-669.

Gray TA, Wilson A, Fortin PJ, Nicholls RD. 2006. The putatively functional Mkrn1-p1 pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in trans. *Proc Natl Acad Sci U S A* **103**(32): 12039-12044.

Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-Boris A, Yoshiki A. 2003. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* **423**(6935): 91-96.

Jacq C, Miller JR, Brownlee GG. 1977. A pseudogene structure in 5S DNA of Xenopus laevis. *Cell* **12**(1): 109-120.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome research* **20**(10): 1313-1326.

Kaneko S, Aki I, Tsuda K, Mekada K, Moriwaki K, Takahata N, Satta Y. 2006. Origin and evolution of processed pseudogenes that stabilize functional Makorin1 mRNAs in mice, primates and other mammals. *Genetics* **172**(4): 2421-2429.

Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, Cayting P, Harrrison P, Gerstein M. 2007. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic acids research* **35**(Database issue): D55-60.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America* **100**(20): 11484-11489.

Khachane AN, Harrison PM. 2009. Assessing the genomic evidence for conserved transcribed pseudogenes under selection. *BMC genomics* **10**: 435.

Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S et al. 2014. A draft map of the human proteome. *Nature* **509**(7502): 575-581.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**(21): 2947-2948.

Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych N et al. 2014. Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic acids research* **42**(Database issue): D926-932.

Podlaha O, Zhang J. 2004. Nonneutral evolution of the transcribed pseudogene Makorin1-p1 in mice. *Mol Biol Evol* **21**(12): 2202-2209.

-. 2010. Pseudogenes and their evolution. In *Encyclopedia of Life Sciences*, pp. 1-8. John Wiley & Sons, Chichester, UK.

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**(7301): 1033-1038.

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M et al. 2015. The UCSC Genome Browser database: 2015 update. *Nucleic acids research* **43**(Database issue): D670-681.

Schmidt EE. 1996. Transcriptional promiscuity in testes. *Current biology : CB* **6**(7): 768-769.

Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthes P, Kokkinaki M, Nef S, Gnirke A et al. 2013. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell reports* **3**(6): 2179-2190.

Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* **509**(7502): 582-587.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**(8): 1586-1591.

Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. 2010. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome biology* **11**(3): R26.

CHAPTER 4

PROTEIN FOLD CLASSIFICATION IN A CONTINUOUS STRUCTURE SPACE

4.1 ABSTRACT

Protein domain structure classification is important for understanding the structure-function relationship and protein evolution. Classification at the fold level is of special interest because it is the lowest level of classification that does not depend on protein sequence similarity. However, the current fold classifications such as those in SCOP and CATH are controversial because they implicitly assume that folds are discrete islands in the structure space, whereas increasing evidence suggests significant similarities among folds and supports a continuous fold space. Ignoring this continuity compromises protein structure classification and hinders the understanding of structure-function relationship and protein evolution. Here we develop a likelihood method to classify a query domain into the existing folds of CATH or SCOP by considering both the structural similarity between the query and various folds and within-fold structural heterogeneities. Depending on the structural similarity score used and the original classification scheme, the new classification differs from the original classification for 4-12% of all domains and up to 20% of domains with recently solved structures. These results confirm the continuous nature of the fold space and demonstrate the growing importance of considering this continuity in fold classification. Our method is in principle applicable to classifications at all levels.

4.2 INTRODUCTION

Since the 1970s, classification of protein domain structures has gained wide popularity because of its utility in predicting protein function and studying protein evolution. Many hierarchical classifications of domain structures have been developed (Swindells et al. 1998). Among them, SCOP (Andreeva et al. 2008) and CATH (Orengo et al. 1997; Cuff et al. 2011; Sillitoe et al. 2013) databases are commonly regarded as the gold standards because of their substantial manual inspections. The hierarchical levels of SCOP from bottom to top are family, superfamily, fold, and class. Families and superfamilies consist of domains that are homologous or structurally very similar. Folds comprise superfamilies of domains with similar secondary structure compositions, orientations, and connection orders. Classes, as the top level, include folds with similar secondary structure compositions. In CATH, the hierarchies are homology superfamily (H), topology (T), architecture (A), and class (C). The H, T, and C levels in CATH are respectively equivalent to the superfamily, fold, and class levels in SCOP. Fold in SCOP or T in CATH is of special interest to structural biologists because members of a fold are structurally similar yet have no detectable protein sequence similarity (Orengo et al. 1999; Grishin 2001; Caetano-Anolles and Caetano-Anolles 2003; Wang and Caetano-Anolles 2009). Thus, fold classification can provide significant insights into protein function and evolution that are beyond the realm of sequence analysis.

The current fold classification in SCOP and CATH implicitly assumes that different folds represent isolated islands in the structure space. This assumption was based on early visual observations from a small number of folds that are structurally highly dissimilar. With the explosion of the number of solved domain structures and the use of structure similarity metrics, increasing evidence supports the concept of a continuous fold space where domains from different folds have significant structural similarities (Shindyalov and Bourne 2000; Harrison et al. 2002; Kolodny et al. 2006; Pascual-Garcia et al. 2009). This discovery prompted multiple authors to question the current fold hierarchy (Kolodny et al. 2006) and propose alternative representations such as structure similarity networks (Nepomnyachiy et al. 2014) and maps (Choi and Kim 2006; Osadchy and Kolodny 2011). In a network, domains are connected if their structure similarity exceeds an arbitrary threshold, whereas in a map, domains are points in a plane or space reduced from a pairwise structure similarity matrix of all domains. However, none of these new representations are intuitive due to the lack of obvious fold boundaries. As a result, the conventional fold representation still dominates the literature in the study of protein structure-function relationship and protein evolution. Thus, we use "fold space" interchangeably with "structure space" in this work.

Various automatic pipelines have been developed to classify domain structures into folds, and they can be generally divided into two types. The first type (Taylor and Orengo 1989; Pearl et al. 2001; Getz et al. 2002; Harrison et al. 2003; Rogen and Fain 2003; Cheek et al. 2004; Camoglu et al. 2005; Fox et al. 2014) directly classifies domains according to their structure and/or sequence similarities with existing folds. The second type (Cheng and Baldi 2006; Kim and Patel 2006; Yan et al. 2009; Jo and Cheng 2014) uses a machine learning approach. It first collects positive samples from domain pairs in the same folds and then train classifiers on these domain pairs. These classifiers are then used to predict whether a query domain is in the same fold as another domain. To our knowledge, none of the current classification methods explicitly consider fold space continuity. As a result, it is unclear to what degree the fold space continuity affects protein structure classification and whether it is legitimate to ignore this continuity in classification.

To answer these questions, we here propose and implement a new method to classify domain structures to existing folds by considering fold space continuity. Briefly, we calculate the likelihood that a structure belongs to a fold by considering the similarity between the structure and the fold as well as the similarities among the structures already classified into the fold. By comparing our new classification with the current CATH and SCOP classifications, we assess the importance of considering the fold space continuity in fold classification.

4.3 RESULTS

4.3.1 Fold classification without considering within-fold structure heterogeneity

To classify domain structures, we need an objective quantity to measure structure similarities between two domains. TM-score (Zhang and Skolnick 2004; Xu and Zhang 2010), calculated by the software TM-align (Zhang and Skolnick 2005), is chosen for this purpose. High TM-score indicates short average spatial distance between aligned residues in a structure alignment (see Materials and Methods). Unlike many other similarity scores (Kabsch 1978; Holm and Sander 1995; Siew et al. 2000; Zemla 2003), TM-scores of different domain pairs are directly comparable (Zhang and Skolnick 2004; Zhang and Skolnick 2005; Xu and Zhang 2010) due to the normalization using either the average sequence length of the two domains under comparison or the length of the shorter domain. The former normalization penalizes the length difference between the two domains, which is appropriate when both domains are complete and comparable (i.e., one is not a subunit of the other). This normalization emphasizes the global structure similarity between domains, and the obtained TM-score is referred to as the global TM-score. By contrast, the latter normalization is appropriate when one domain corresponds to a subunit of the other or when one or both domains are incomplete. We refer to such normalized

TM-scores as local TM-scores. After normalization, the range of TM-scores is between 0 and 1. Larger TM-scores indicate higher structural similarities. Both types of TM-scores are used in our analyses.

We focus primarily on the CATH database in this study because it is updated regularly and contains more recently solved domain structures than other databases. We refer to the T level in CATH as fold, because it is equivalent to the fold hierarchy in SCOP. We collected from CATH (version 3.5.0) 21,309 representative domains whose mutual sequence identities are ≤ 60% and sequence lengths are ≥ 40 residues. These domains are from 1,158 folds in the CATH classification. Of these folds, 141 comprise at least 25 representative domains each. We used these large folds in subsequent analysis, because smaller folds provide insufficient information for statistical analysis. In spite of the low fraction of folds analyzed here, for two reasons, these large folds are highly likely to cover most continuous regions of the fold space. First, these large folds include 17,043 or 82% of all representative domains. Second, the large folds are closer to one another than they are to the 1017 small folds ($P < 1.5\text{e-}18$; Wilcoxon rank test), where the closeness between two folds is measured by the highest TM-score of all domain pairs across the two folds.

We randomly choose 10% of domains from each of the large folds as our query domains, whereas the rest of the domains stay in their originally classified folds. To classify a query, TM-scores are calculated between the query and all domains in a fold. The maximum TM-score observed represents the query-fold similarity, and is referred to as query-fold TM_{max}-score. The query is assigned to the fold with the highest query-fold TM_{max}-score. We repeated this entire process 30 times to estimate the frequency of inconsistency between the TM_{max}-based classification and the CATH classification.

Our global TM_{max}-score-based classification is inconsistent with the current CATH fold classification for an average of 1.1% of queries (**Fig. 4.1**). This value increases to 2.9% under the local TM_{max}-score-based classification (**Fig. 4.1**). We also tried using either the mean or median TM-score instead of TM_{max}-score to define domain-fold similarity, but the frequency of inconsistency rises to 17-30% (**Fig. 4.1**). These results indicate that the CATH fold classification is primarily based on the information contained in TM_{max}-scores, especially in terms of the global structural similarity. Thus, global TM_{max}-score-based fold classification, which can be fully automated, may be used as a proxy for CATH classification.

4.3.2 Within-fold structure heterogeneity varies among folds

Different folds in the current CATH classification may have different levels of structure heterogeneity. To measure structure heterogeneity within a fold, we first calculated the (global or local) TM_{max}-score for each domain in the fold, which is defined by the highest TM-score between the focal domain and all other domains in the fold. We then calculated the mean and standard deviation of TM_{max}-scores of all domains in the fold. The higher the mean within-fold TM_{max}-score, the lower the structure heterogeneity within the fold. Our analysis reveals that some folds are highly homogenous with the mean within-fold TM_{max}-score approaching 1, whereas some other folds are highly heterogeneous with the mean within-fold TM_{max}-score as low as 0.6-0.7 (**Fig. 4.2A**). Furthermore, the standard deviation of within-fold TM_{max}-scores also varies greatly among folds and a very strong negative correlation exists between the mean and standard deviation of within-fold TM_{max}-scores (**Fig. 4.2B**). This latter observation indicates that, when a fold has a low mean TM_{max}-score, it is typically because some of the within-fold TM_{max}-scores are very low rather than all within-fold TM_{max}-scores are low.

4.3.3 A likelihood method for fold classification considering within-fold structure heterogeneity

How well a query fits a fold should not only be determined by the query-fold TM_{max}-score, but also the distribution of within-fold TM_{max}-scores; folds with wider distributions of within-fold TM_{max}-scores are more accommodating to a query than those with narrower distributions. The likelihood that a query belongs to a particular fold can be measured by the probability that the fraction of within-fold TM_{max}-scores equal to or smaller than the query-fold TM_{max}-score. We refer to this probability as the cumulative empirical probability (CEP). Note that CEP is a measure of the fit of a query-fold TM_{max}-score to the TM_{max}-scores of all members already classified to the fold. CEP is not the posterior probability that a query belongs to a fold, and the sum of CEPs for all folds is not necessarily 1. Fig. 4.3 shows a hypothetical example where CEP classifies a query into fold2 despite that the query-fold2 TM_{max}-score is lower than the query-fold1 TM_{max}-score (**Fig. 4.3A**). This occurs because the fraction of within-fold TM_{max}-scores that are equal to or smaller than the corresponding query-fold TM_{max}-score is smaller for fold1 (**Fig. 4.3B**) than for fold2 (**Fig. 4.3C**). Note, however, that classifications by CEP and TM_{max}-score would always be consistent if the fold space is completely discrete, because then the TM_{max}-scores of a query with fold1 and fold2 would be extremely different.

Estimating CEP requires the information on the empirical distribution of within-fold TM_{max}-scores. When the number of domains in a fold is not very large, CEP estimates may be inaccurate. For example, when the query-fold TM_{max}-score is lower than all observed within-fold TM_{max}-scores, one assigns CEP = 0, although the true CEP must be > 0. To minimize this problem, we can fit the observed within-fold TM_{max}-scores ($\boldsymbol{x}$) by a Gaussian mixture model

(GMM) and then estimate CEP using the fitted continuous distribution (see Materials and Methods). The use of GMM is inspired by the fact that (i) the distribution of within-fold $\mathrm{TM}_{\mathrm{max}}$-scores usually has multiple modes presumably due to the existence of multiple superfamilies in the fold and (ii) that GMM is highly flexible and fits almost any distribution. The parameters of the GMM are inferred under the Bayesian framework with model settings proposed by Richardson and Green (Richardson and Green 1997). With the posterior distributions of the parameters, the posterior predictive distribution of $\mathrm{TM}_{\mathrm{max}}$-scores $f(\tilde{x}|\boldsymbol{x})$ is estimated using a Monte Carlo method, where $f(\tilde{x}|\boldsymbol{x})$ denotes the probability density of a potentially observed $\mathrm{TM}_{\mathrm{max}}$-score ($\tilde{x}$) given the observed $\mathrm{TM}_{\mathrm{max}}$-scores ($\boldsymbol{x}$). CEP is then determined using $f(\tilde{x}|\boldsymbol{x})$ as if the potentially observed $\mathrm{TM}_{\mathrm{max}}$-scores are actually observed. We refer to this CEP estimate as the **c**umulative **p**osterior **p**redictive **p**robability (C3P).

4.3.4 Domain classification using CEP and C3P with global $\mathrm{TM}_{\mathrm{max}}$-scores

Let us first use global $\mathrm{TM}_{\mathrm{max}}$-scores in CEP and C3P classifications. This way of TM-score normalization emphasizes the global similarity between domains. For the same 30 random sets of queries previously used, the CEP classification differs from the CATH classification in 3.4% of cases on average (**Fig. 4.1**). We refer to the query domains that have different classifications by CEP and CATH as reclassified domains. The majority of these domains are attracted to a small number of folds in CEP classification (**Fig. 4.4A**). These folds tend to have large structure heterogeneities (i.e., with low averages and high standard deviations of within-fold $\mathrm{TM}_{\mathrm{max}}$-scores). In fact, the structure heterogeneity of a fold and the number of reclassified domains attracted to the fold are significantly correlated (**Fig. 4.5A, B**). By contrast, there is no significant correlation between the number of reclassified domains attracted to a fold and the fold size (**Fig. 4.5C**).

On average, C3P classification differs from the CATH classification in 4.3% of cases (**Fig. 4.1**), and the reclassified queries by C3P are also attracted to a small number of folds (**Fig. 4.4A**). The general patterns of C3P reclassifications are similar to what was observed in CEP reclassifications (**Fig. A.2.2.1A-C**). Averaged over the 30 query sets, 97% of the queries are classified consistently by CEP and C3P (**Fig. 4.4B**). Moreover, 60% of the reclassifications by CEP are reclassified the same way by C3P, and 49% of the reclassifications by C3P are reclassified the same way by CEP (**Fig. 4.4B**).

Below we provide an example of reclassification by global TM_{max}-score-based C3P (**Fig. 4.6**). Domain 1ny8A00 (CATH Id) has a TM_{max}-score of 0.55 with fold 3.30.300 (**Fig. 4.6A**) and a TM_{max}-score of 0.54 with fold 3.30.460 (**Fig. 4.6B**), and was classified into fold 3.30.300 by CATH. However, the mean within-fold TM_{max}-score is quite high (0.81) for fold 3.30.300 (**Fig. 4.6C**). Consequently, the probability for a within-fold TM_{max}-score to be ≤ 0.55 is small (C3P = 0.06). By contrast, fold 3.30.460 has a substantial structure heterogeneity (mean within-fold TM_{max}-score = 0.71; **Fig. 4.6D**), rendering it quite likely that a within-fold TM_{max}-score is ≤0.54 (C3P = 0.15). As a result, 1ny8A00 is reclassified to fold 3.30.460 by C3P.

4.3.5 Domain classification using CEP and C3P with local TM_{max}-scores

In this section, we use CEP and C3P with local TM_{max}-scores for classification. This treatment is consistent with the focus on substructure similarity between domains in the study of fold space continuity. For the 30 sets of queries, CEP and C3P classifications both differ from CATH classification for 12% of cases (**Fig. 4.1**), suggesting that the impact of fold space continuity on fold classification is larger if local structure similarity is considered. Similar to what was observed in the previous section, the reclassified queries are also attracted to a small number of folds (**Fig. 4.4C**). And the number of domains reclassified into a fold correlates with measures of the fold's structure heterogeneity (**Fig. 4.5D, E; Fig. A.2.2.1D, E**), but is uncorrelated with the number of domains in the fold (**Fig. 4.5F; Fig. A.2.2.1F**). CEP and C3P classifications are consistent with each other for 97% of cases (**Fig. 4.4D**). Seventy-seven percent of the reclassifications by CEP are reclassified the same way by C3P, while 81% of the reclassifications by C3P are reclassified the same way by CEP (**Fig. 4.4D**).

4.3.6 Classification of newly solved domain structures in CATH by CEP and C3P

The query domains used in previous sections were randomly chosen from the 17,043 representative domains in CATH v3.5.0. These queries are unbiased samples and their reclassification results by CEP and C3P represent the overall impact of structure space continuity on fold classification. However, if we need to classify a newly solved domain structure into the current CATH fold hierarchy, how big of an impact would the use of CEP or C3P have? To address this question, we took the 17,043 representative domains from the 141 large folds in CATH v3.5.0 (available from Sept., 2011) as the initial classification. In CATH v4.0.0 (available from March 2013), these large folds contain 8280 representative domains that did not exist in CATH v3.5.0. We now use these 8280 newly added domains as queries. When the global TM_{max}-score is used, CEP (or C3P) classifications differ from CATH classifications for 4.0% (or 4.8%) of these 8280 domains. When the local TM-score is used, CEP (or C3P) classifications differ from CATH classifications for 20.8% (or 20.5%) of these domains. These values are higher than the corresponding numbers for the 30 sets of randomly picked domains, suggesting that the current research biases towards domains whose classifications are affected more by the structure space continuity, relative to the past research.

4.3.7 Classification of domains in the SCOP database

We next examined the fold classification in SCOP, another widely used protein classification system. Using the same criteria as used for CATH, we generated 30 sets of 606 representative queries from 89 large folds in SCOP version 1.73. Global $\mathrm{TM}_{\mathrm{max}}$-score-based fold classification is largely consistent with the SCOP classification, with only 0.9% of inconsistent cases (**Fig. 4.7**). This number increases to 2.4% under local $\mathrm{TM}_{\mathrm{max}}$-score-based classification. The frequency of inconsistent classification is much greater when the query-fold similarity is measured by either the mean or median TM-score instead of $\mathrm{TM}_{\mathrm{max}}$-score (**Fig. 4.7**). These results indicate that, similar to CATH, SCOP fold classification can be automated using query-fold global $\mathrm{TM}_{\mathrm{max}}$-score.

For the same 30 random sets of queries, the global $\mathrm{TM}_{\mathrm{max}}$-score-based and local $\mathrm{TM}_{\mathrm{max}}$-score-based CEP classifications differ from the SCOP classification for an average of 5.9% and 7.6% of queries, respectively (**Fig. 4.7**). These numbers become 7.8% and 8.6%, respectively, for global and local $\mathrm{TM}_{\mathrm{max}}$-score-based C3P classifications, respectively (**Fig. 4.7**).

By comparing SCOP versions 1.73 (available from Nov. 2007) and 1.75 (available from June 2009 and the most updated version), we found that 801 representative domains were added into the 89 large folds in version 1.75 since version 1.73. These most recent additions to SCOP were subject to CEP and C3P classifications. The global and local $\mathrm{TM}_{\mathrm{max}}$-score-based CEP classifications of these domains are inconsistent with the SCOP classification for 5.9% and 7.5% of the cases, respectively. These numbers become 7.2% and 9.1%, respectively, under C3P.

Reclassifications are rarer for SCOP than for CATH except under global TM_{max}-score-based CEP and C3P (**Fig. 4.1; Fig. 4.7**). The SCOP data used here comprise 89 large folds and 65% of the total 9,964 representative domains in v1.73, whereas the CATH data consist of 141 large folds and 82% of the total 21,309 representative domains in v3.5.0. The sparser SCOP data than CATH data may render the classification more straightforward for the former than the latter. Intriguingly, however, global TM_{max}-score-based CEP and C3P classifications are less consistent with SCOP than CATH classifications. To identify the underlying reason, we focus on the CEP classifications of the 6,134 non-redundant queries in the 30 SCOP sets. Each query has an original fold assigned by SCOP. The domain used to calculate query-original fold TM_{max}-score is referred to as the partner domain. The relative length difference between the query and the partner domain is defined by the absolute value of their length difference divided by the shorter length. We found the relative length difference significantly greater for SCOP than CATH queries (**Fig. 4.8**). Because length difference reduces global TM_{max}-scores, query-original fold global TM_{max}-scores are reduced more drastically for SCOP than CATH queries, resulting in more reclassifications for the former than the latter. Indeed, reclassified SCOP queries tend to have larger relative length differences with their partner domains than average SCOP queries (**Fig. 4.8**).

4.4 DISCUSSION

In this work, we first showed that the fold classification in CATH is highly similar to the classification by the query-fold TM_{max}-score, especially the global TM_{max}-score. Considering fold space continuity, we developed the CEP and C3P methods to classify domain structures into existing folds using both query-fold TM_{max}-scores and within-fold TM_{max}-scores. We found that, using global TM_{max}-scores, considering space continuity leads to reclassifications of ~4% of domains. However, when local TM_{max}-scores are used, considering space continuity leads to reclassification of ~12% of domains. This increased reclassification rate under local TM_{max}-scores is potentially due to prevalent substructure similarities across folds (Orengo et al. 1997; Harrison et al. 2002; Krishna and Grishin 2005) known as the Russian doll effect (Orengo et al. 1997). With this effect, a query is similar to multiple folds when local TM_{max}-scores are used. Consequently, the reclassification rate is increased.

Under the global TM_{max}-score, considering fold space continuity leads to the reclassification of 6~8% SCOP domains, compared to 3~4% of CATH domains. This increased reclassification rate for SCOP domains is potentially because the Russian doll effect is stronger for SCOP folds than CATH folds (**Fig. 4.8**), which renders the global query-original fold TM_{max}-score lower for SCOP queries than CATH queries, prompting more reclassifications for the former than the latter. These considerations suggest that fold space continuity is affected by within-fold domain length heterogeneity.

Our results show that fold space continuity requires a sizable number of domain reclassifications. We may have underestimated the number of required reclassifications because the datasets used are sparser than the complete fold space due to the removal of small folds. With more domain structures solved, the fold space will become more continuous. Furthermore, we observed a rise in reclassification rates for newly solved structures in CATH. Given the potential future surge of new domain structures predicted from the substantial increase of domains in CATH in recent years, considering fold space continuity is urgently needed for fold classifications.

As mentioned, some automatic fold classifiers (Cheng and Baldi 2006; Kim and Patel 2006; Yan et al. 2009; Jo and Cheng 2014) use machine learning approaches and are trained with within-fold domain pairs from multiple folds. However, pooling domain pairs from many folds for training ignores the among-fold variation in within-fold structure heterogeneity. As a result, the impact of fold space continuity is not adequately considered in classification. One might think that such classifiers can be improved by training with individual folds, but this is infeasible because of small sample sizes of most folds that cause overfitting of the classifiers with large numbers of parameters. Thus, CEP and C3P are unique in that they explicitly consider structure space continuity in fold classification. Furthermore, CEP and C3P are probabilities and are thus easy to interpret. In addition, C3P is derived using Bayesian hierarchical models, which alleviate overfitting and allow experts to set various priors according to their beliefs on fold structural variations.

We found that the classification using global query-fold TM_{max}-scores is inconsistent with CATH classification for only 1% of cases, confirming that CATH classifies a new domain based on its best match to existing members of various folds. It is clear that this way of classification will result in the problem that some domains of the same fold are less similar to one another than to domains from other folds, which is observed in CATH (Xu and Zhang 2010). For example, domains A and B with low similarity to each other may be classified into the same fold because of their respective high similarities to some existing members in the fold. The non-transitive domain pairs such as A and B were observed previously (Orengo et al. 1997; Pascual-Garcia et al. 2009), but its prevalence and impact on classification in CATH were unclear. We found that when a query is classified based on the mean instead of maximal query-fold TM-score, the classification differs from CATH for ~20% of cases. This substantial rise in inconsistency suggests that non-transitive domain pairs are quite common in CATH.

Due to numerous non-transitive domain pairs in a fold, mean within-fold $\mathrm{TM_{max}}$-scores are much larger than mean within-fold $\mathrm{TM_{mean}}$-scores. Because CEP and C3P are designed to improve current fold classifications such as CATH and SCOP, which are similar to $\mathrm{TM_{max}}$-score-based classification, it is reasonable to use $\mathrm{TM_{max}}$-scores in CEP and C3P. Classification methods based on machine learning approaches are also designed to facilitate current fold classifications. However, their practice is inappropriate due to the use of all pairs of domains within a fold rather than the most similar pairs as training sets. This may explain high reclassification rates (~20%) of machine learning methods even though fold space continuity is not considered.

In summary, we have developed CEP and C3P to classify domains into folds by considering fold space continuity. The inconsistencies between CEP/C3P and current hierarchical classifications in CATH and SCOP demonstrate a substantial impact of structure continuity on fold classification, requiring considering structure continuality in future classifications of domain structures. Structure continuity also calls for model-based clustering of domains where the number of folds and memberships in each fold are both probabilistic.

4.5 MATERIALS AND METHODS

4.5.1 Protein structure similarity score

TM-score defined below is used to assess the structural similarity between two protein structures. $\text{TM-score} = \frac{1}{L}\left[\sum_{i=1}^{L_{ali}} \frac{1}{1+d_i^2/d_0^2}\right]_{\max}$, where L is the length of the shorter protein or mean length of the two proteins being compared, L_{ali} is the number of equivalent residues in the two proteins, d_i is the distance of the ith pair of the equivalent residues between the two superposed structures, $d_0 = 1.24\sqrt[3]{L - 15} - 1.8$ is used to normalize the TM-score so that the average magnitude of the TM-score for random protein pairs is independent of the size of the proteins, and "max" indicates the highest value among all possible superposition. TM-score ranges in (0, 1] with a higher value indicating a higher similarity. TM-scores between two domains are calculated using the TM-align software (Zhang and Skolnick 2005).

4.5.2 Initial classifications and query domains

From CATH v4.0.0, we collected 23,682 representative single domains with mutual sequence identities $\leq 60\%$ and lengths ≥ 40 residues using CD-Hit (Fu et al. 2012). Among them, 21,309 representative domains existed in an older version of CATH (v3.5.0). These 21,309 domains are from 1,158 folds in CATH v3.5.0. A total of 141 of these folds each have at least 25 domains. From each of these 141 large folds, we randomly sampled 10% of domains; these 1623 queries sampled were subjected to classification by other methods. The remaining 15,420 domains in the 141 folds constitute the initial fold classification. This procedure was repeated 30 times, generating 30 random sets of queries. To examine the CATH classifications of newly solved domain structures, representative domains of the 136 large folds in CATH v3.5.0 were used as the initial classification, whereas the 8280 domains newly added to the 136 folds in CATH v4.0.0 were queries. With the same criterion, 6,476 representative domains in 89 large folds were collected from SCOP v1.73. Using only the large folds, 30 sets of query domains were randomly picked. Each of the sets contained 606 queries, and the other 5,870 domains in large folds were treated as the initial classifications. A total of 801 domains newly added to the 86 folds in SCOP v1.74 since v.1.73 were identified as newly solved queries.

4.5.3 Gaussian mixture model and posterior predictive distribution

The observed within-fold TM$_{\max}$-scores for a fold are denoted as $\boldsymbol{x} = (x_1, \dots, x_N)$, which are assumed to have been independently drawn from a mixture of k Gaussian components. *N* is the number of representative domains in the fold. The probability of observing an *x* is

$$p(x|k, \pi, \mu, \sigma^2) = \sum_{i=1}^{k} \pi_i \, f(x; \mu_i, \sigma_i^2), \tag{1}$$

where μ_i, σ_i^2, and π_i are the mean, variance, and mixture proportion of component i. Latent allocation data are referred to as $z = (z_1, \dots, z_N)$, in which z_i specifies the mixture component to the observation x_i. Here, z_i's are independently and identically distributed samples from the following probability mass function (PMF).

$$p(z_i = j|\pi, \mu, \sigma^2) = \pi_j. \tag{2}$$

Conditional on the allocation value z_i, the observed x_i is a random number from the following Gaussian probability density function.

$$p(x_i|z_i = j, \pi, \mu, \sigma^2) = p\left(x_i; \mu_j, \sigma_j^2\right) = f\left(x_i; \mu_j, \sigma_j^2\right) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{\left(x_i - \mu_j\right)^2}{2\sigma_j^2}} . \tag{3}$$

We assume the following priors of the parameters:

$$\pi|k \sim Dirichlet(\gamma_1, \dots, \gamma_k); \tag{4}$$

$$\mu_j \sim Normal(\xi, \kappa^{-1}); \tag{5}$$

$$\sigma_j^{-2} \sim Gamma(\alpha, \beta); \tag{6}$$

$$k \sim Poisson(\lambda) . \tag{7}$$

We set $\gamma_i = 1$, ξ = median of observed TM$_{\max}$-scores of a fold, and κ^{-1} = R^2,where R is the difference between the maximum and minimum of $\boldsymbol{x}$. These parameters make the prior distributions of π and μ_j rather flat. We set $\alpha = 2$ and $\beta \sim \Gamma(g, h)$, which is a gamma distribution with the shape parameter $g = 0.2$ and rate parameter h equal to $10/R^2$, to express the belief that the σ_j^{-2}s are similar. At last, k follows a Poisson distribution with parameter $\lambda = 1$. All the settings together render the priors weakly informative and thus allow observed TM$_{\text{max}}$-scores to dominate the parameter inference. The joint prior probability can be written as

$$P(k, \pi, \mu, \sigma^{-2} | \alpha, \beta, \gamma, \xi, \kappa^{-1}, \lambda) = P(k|\lambda)P(\pi|k, \gamma)P(\sigma^{-2}|k, \alpha, \beta)P(\mu|k, \xi, \kappa^{-1}), \quad (8)$$

and therefore the joint posterior probability is

$$P(k, \pi, \mu, \sigma^{-2}|\boldsymbol{x}) \propto P(\boldsymbol{x}|k, \pi, \mu, \sigma^{-2})P(k, \pi, \mu, \sigma^{-2}|\alpha, \beta, \gamma, \xi, \kappa^{-1}, \lambda). \quad (9)$$

Let $\theta = (k, \pi, \mu, \sigma^{-2})$ and $\tilde{x}$ denote unobserved TM$_{\text{max}}$-scores of the fold. The posterior predictive distribution is

$$f(\tilde{x}|\boldsymbol{x}) = \int P(\tilde{x}, \theta|\boldsymbol{x})\, d\theta = \int P(\tilde{x}|\theta)\, P(\theta|\boldsymbol{x}) d\theta, \quad (10)$$

which is the probability density function (PDF) of potential query-fold TM$_{\text{max}}$-scores ($\tilde{x}$) given the observed within-fold TM$_{\text{max}}$-scores ($\boldsymbol{x}$).

Due to the lack of a closed form, θ's are sampled from the posterior distribution $P(\theta|\boldsymbol{x})$ using reversible-jump Markov chain Monte Carlo implemented in the R package of miscF (Feng 2013). The simulation had 30,000 iterations, 5000 burn-in steps, and a thinning parameter of 5. Initial values unspecified previously are assigned automatically by the miscF package. Conditional on each θ, $\tilde{x}$ is sampled from the Gaussian mixture $P(\tilde{x}|\theta)$. This model was used to develop the C3P method for fold classifications. The C3P package including scripts to run CEP can be obtained at http://www.umich.edu/~zhanglab/download.htm.

4.6 ACKNOWLEDGEMENTS

We thank Bryan Moyers, Jian-Rong Yang, and Zhengting Zou for comments. This work was supported by the U.S. National Institutes of Health grant R01GM103232 to J.Z.

Figure 4.1 Comparing classifications by various TM-scores, CEP, C3P and CATH. Fractions of fold-level domain structure classifications by various TM-scores, CEP, and C3P that are inconsistent with the CATH classification. (**A**) 30 sets of 10% randomly chosen domains from CATH v3.5.0 and (**B**) 8280 newly added domains in CATH v4.0.0 since v3.5.0. Error bars show one standard deviation.

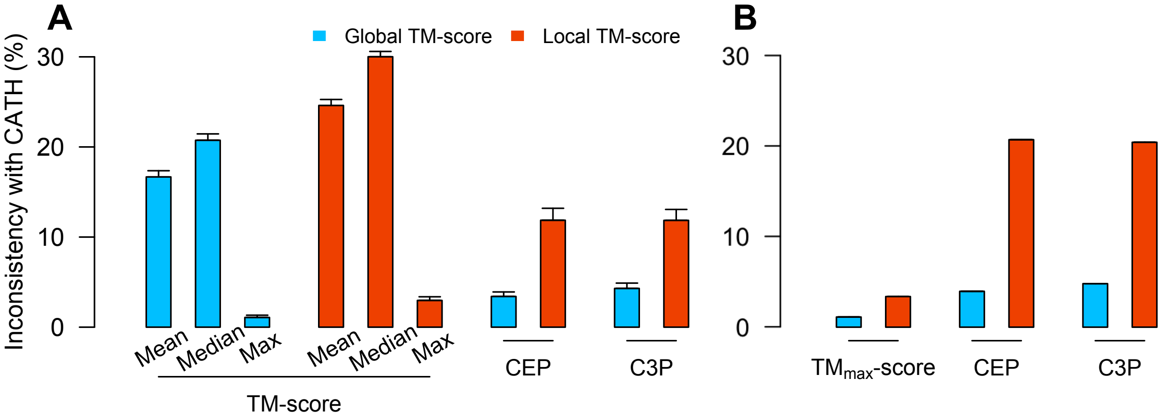


Figure 4.2 Within-fold structural heterogeneities of the 141 large folds in CATH version 3.5.0. (**A**) Mean within-fold TM$_{max}$-score of each fold. A whisker indicates the standard deviation (SD) of the within-fold TM$_{max}$-scores. (**B**) Correlation between the mean and SD of within-fold TM$_{max}$-scores across folds.

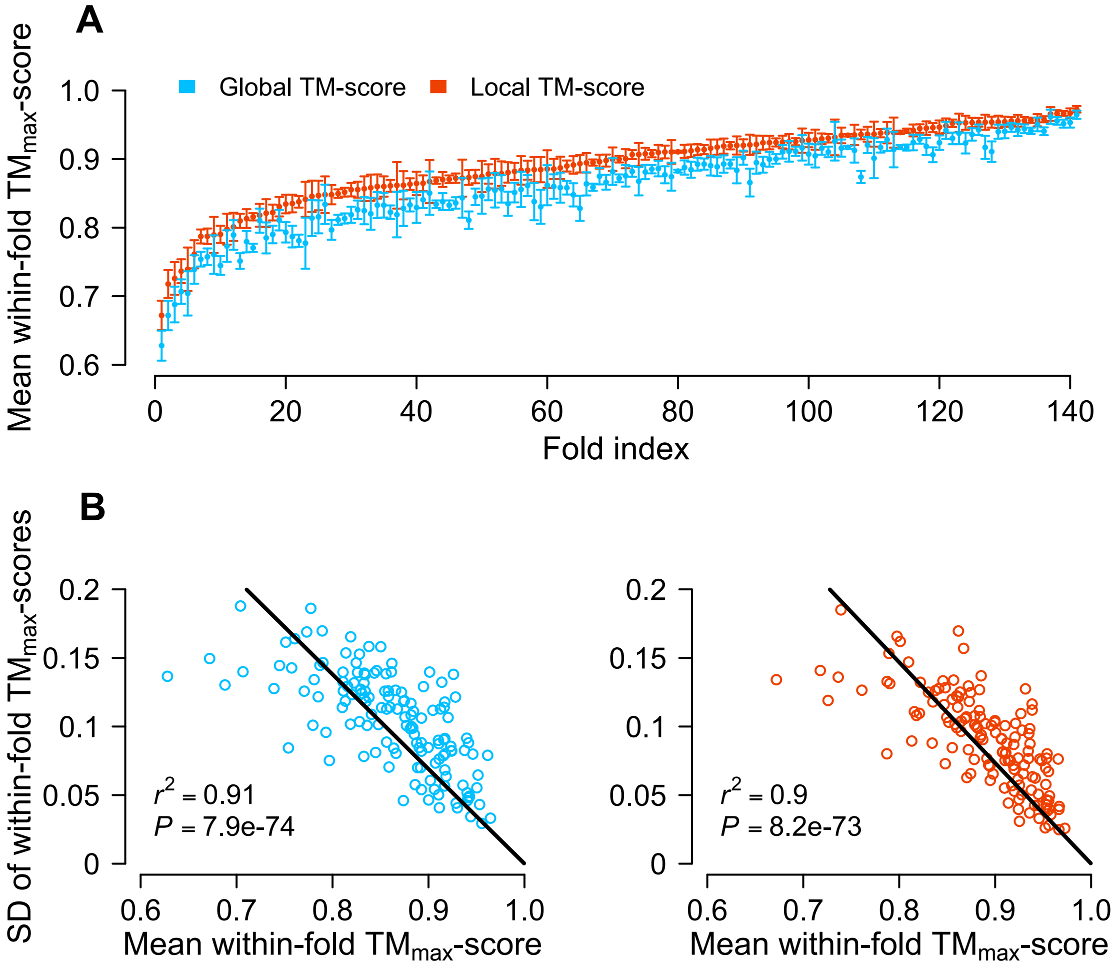


Figure 4.3 A hypothetical example contrasting fold classifications by TM$_{max}$-score and CEP. (**A**) query-fold TM$_{max}$-scores of a query to two folds. (**B**) Frequency distribution of within-fold1 TM$_{max}$-scores. (**C**) Frequency distribution of within-fold2 TM$_{max}$-scores. In (B) and (C), CEP is the area left to the vertical line under the distribution.

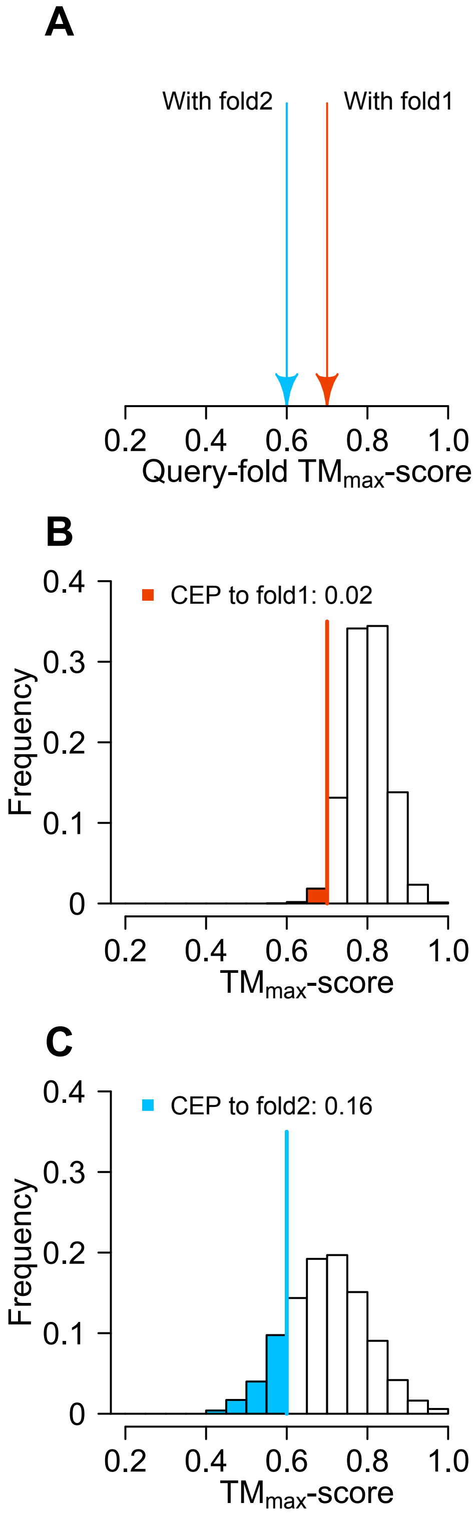


Figure 4.4 CEP and C3P classifications of randomly picked domains from large folds in CATH. (**A**) Reclassifications by global TM$_{max}$-score-based CEP and C3P. A vertical bar corresponds to a fold, and its width is proportional to the number of domains in the fold. Domains within a fold are sorted by length ascendingly. The red, green, and blue colors represent folds from α, β and $\alpha\beta$ classes in CATH, respectively. A line linking vertical bars of two horizontal bars connects the same domain that is classified into different folds by the two different methods, although a vertical bar with the same width of a line denotes 20 domains. This discrepancy is introduced for clearer visualization. (**B**) Venn diagram of classifications by CATH and global TM$_{max}$-score-based CEP and C3P. (**C**) Reclassifications by local TM$_{max}$-score-based CEP and C3P. (**D**) Venn diagram of classifications by CATH and local TM$_{max}$-score-based CEP and C3P. In (A) and (B), results from the first of the 30 sets of queries are presented. In (C) and (D), average results from the 30 sets of queries are presented.

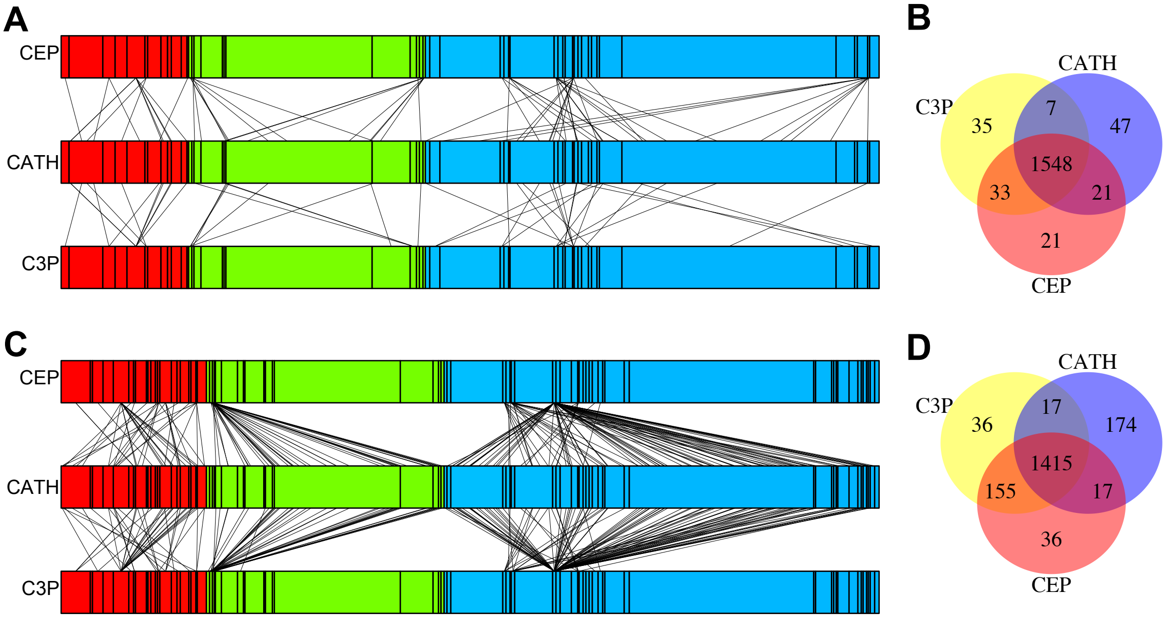


Figure 4.5 Rank correlations between various properties of a fold and the number of domains reclassified into the fold by CEP. The domains are reclassified by (**A-C**) global TM_{max}-score-based CEP and (**D-F**) local TM_{max}-score-based CEP. The lines show linear regressions. ρ, Spearman's rank correlation coefficient.

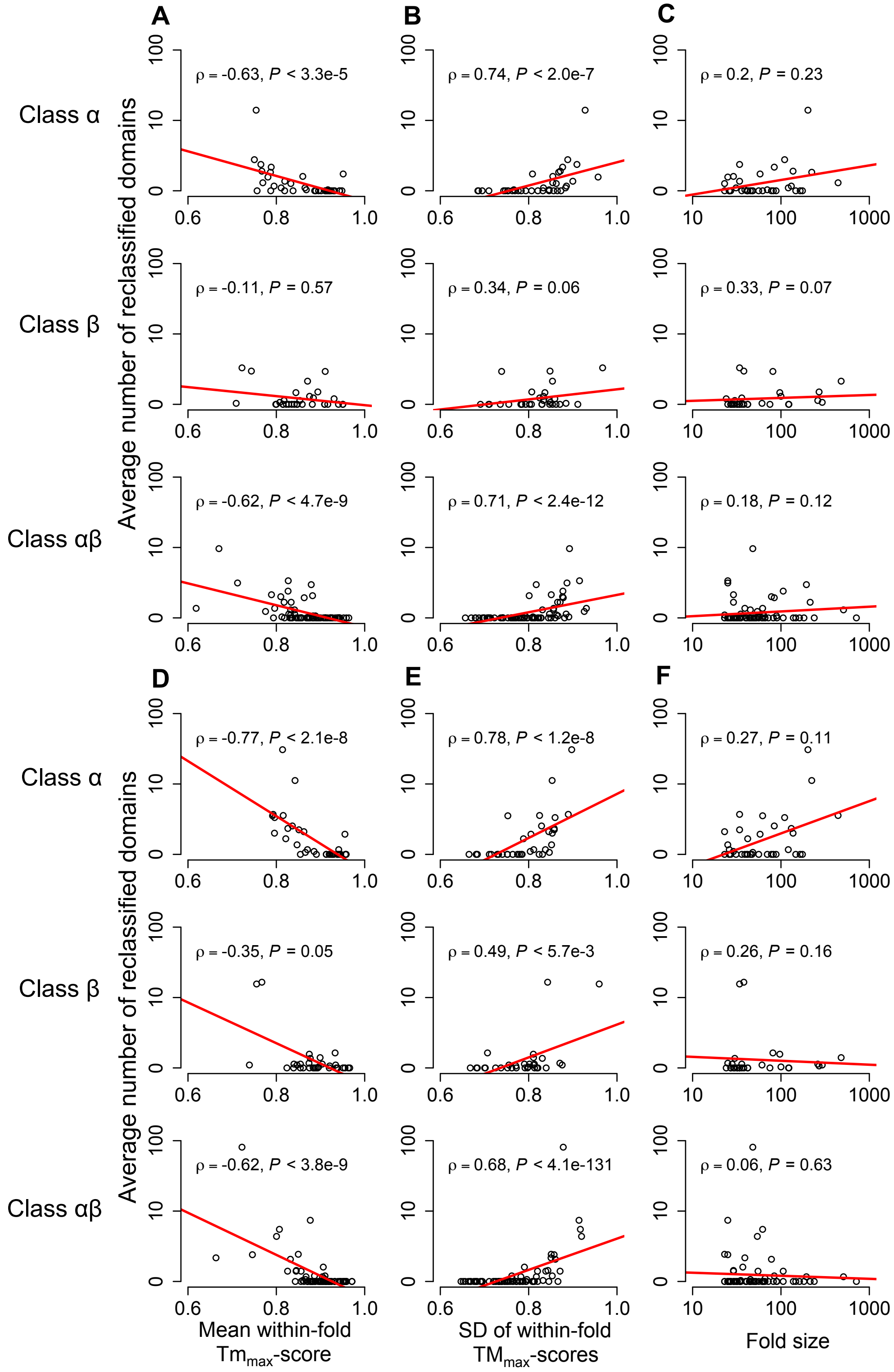


Figure 4.6 Reclassification of a CATH domain by global TM$_{max}$-score-based C3P. The green structure shows the query domain (CATH id = 1ny8A00). The query is reclassified into fold 3.30.460 from fold 3.30.300. (**A**) Structure alignment with 1egaA02 (red) of fold 3.30.300 (query-fold TM$_{max}$-score = 0.55). (**B**) Structure alignment with 1ylqA00 (red) of fold 3.30.460 (query-fold TM$_{max}$-score = 0.54). (**C**) Structure alignment of the query with 10 randomly picked domain members in fold 3.30.300 (C3P = 0.06, domain members in red). (**D**) Structure alignment of the query with 10 randomly picked domain members in fold 3.30.460 (C3P = 0.15, domain members in red). In (C)-(D), the red lines are 10 randomly selected domains from the fold. All structures are displayed by PyMOL

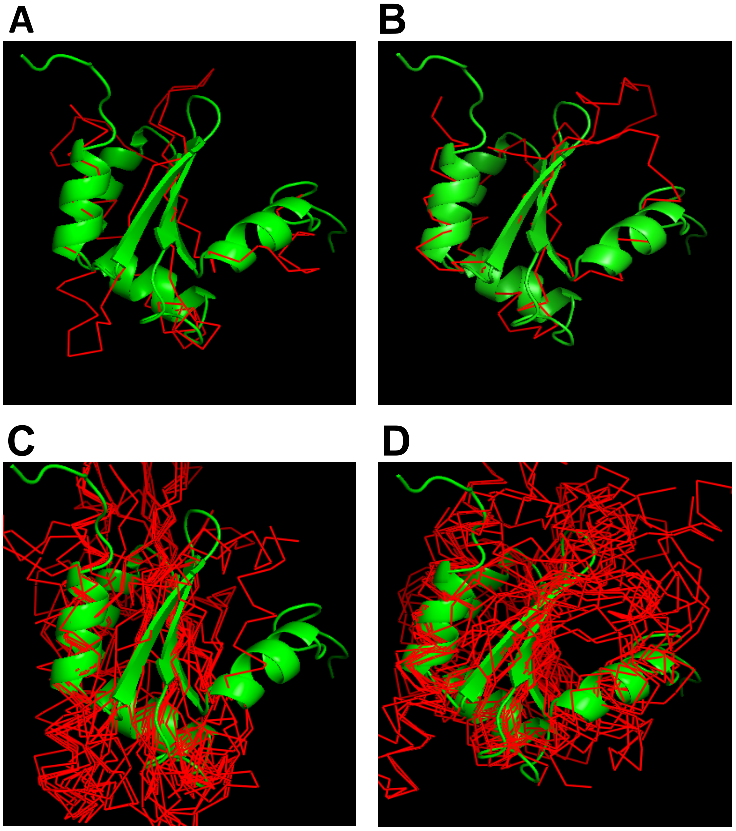


Figure 4.7 Comparing classifications by various TM-scores, CEP, C3P and SCOP. Fractions of fold-level domain structure classifications by various TM-scores, CEP, and C3P that are inconsistent with the SCOP classification for (**A**) 30 sets of 10% randomly chosen domains from CATH v3.5.0 and (**B**) 523 newly added domains in SCOP v1.75 since v1.73. Error bars show one standard deviation.

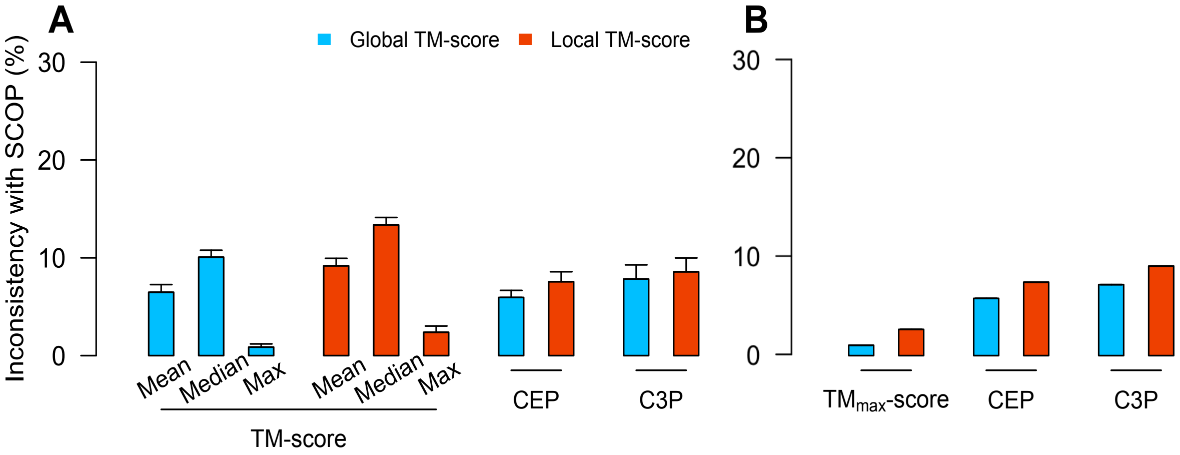
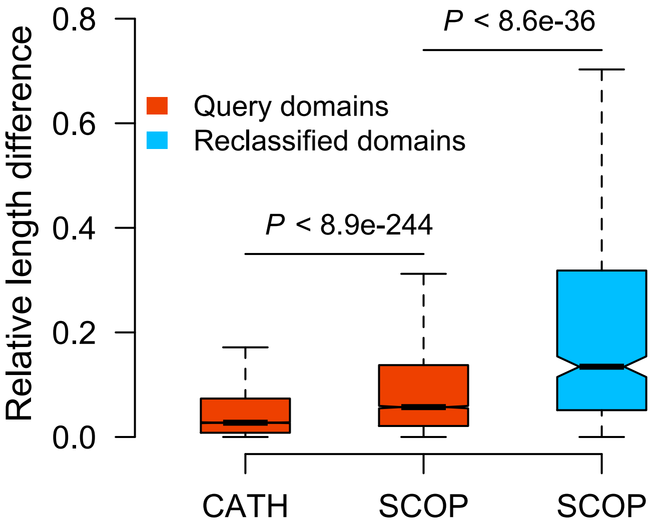


Figure 4.8 Relative length difference between domains within folds. The relative length difference is defined as the absolute length difference between a query from a fold and its best-matched domain (i.e., with the highest TM-score) in the fold, divided by the length of the shorter of the two. Domains included in the CATH bar are 16,173 nonredundant domains from the 30 sets of random queries in CATH v3.5.0. Domains included in the red SCOP bar are 6,134 nonredundant domains from the 30 sets of random queries in SCOP v1.73. Domains included in the blue SCOP bar are 456 queries reclassified by global $\mathrm{TM_{max}}$-score-based CEP. In this bar plot, the notch indicates the median and the bar corresponds to the interquartile range (IQR), covering from the first quartile to the third quartile of the sample. The two whiskers of the bar show the minimum value not smaller than the 1st quartile minus 1.5 times IQR and the maximum value not greater than the 3rd quartile plus 1.5 times IQR, respectively. These values are default in the boxplot package of *R*. *P* values are from Mann-Whitney *U* tests.



4.7 REFERENCES

Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. 2008. Data growth and its impact on the SCOP database: new developments. *Nucleic acids research* **36**(Database issue): D419-425.

Caetano-Anolles G, Caetano-Anolles D. 2003. An evolutionarily structured universe of protein architecture. *Genome research* **13**(7): 1563-1571.

Camoglu O, Can T, Singh AK, Wang YF. 2005. Decision tree based information integration for automated protein classification. *Journal of bioinformatics and computational biology* **3**(3): 717-742.

Cheek S, Qi Y, Krishna SS, Kinch LN, Grishin NV. 2004. 4SCOPmap: automated assignment of protein structures to evolutionary superfamilies. *BMC bioinformatics* **5**: 197.

Cheng J, Baldi P. 2006. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* **22**(12): 1456-1463.

Choi IG, Kim SH. 2006. Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences of the United States of America* **103**(38): 14056-14061.

Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, Pellegrini-Calace M, Jones D, Thornton J, Orengo CA. 2011. Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic acids research* **39**(Database issue): D420-426.

Feng D. 2013. miscF: Miscellaneous Functions.

Fox NK, Brenner SE, Chandonia JM. 2014. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic acids research* **42**(Database issue): D304-309.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**(23): 3150-3152.

Getz G, Vendruscolo M, Sachs D, Domany E. 2002. Automated assignment of SCOP and CATH protein structure classifications from FSSP scores. *Proteins* **46**(4): 405-415.

Grishin NV. 2001. Fold change in evolution of protein structures. *Journal of structural biology* **134**(2-3): 167-185.

Harrison A, Pearl F, Mott R, Thornton J, Orengo C. 2002. Quantifying the similarities within fold space. *Journal of molecular biology* **323**(5): 909-926.

Harrison A, Pearl F, Sillitoe I, Slidel T, Mott R, Thornton J, Orengo C. 2003. Recognizing the fold of a protein structure. *Bioinformatics* **19**(14): 1748-1759.

Holm L, Sander C. 1995. Dali: a network tool for protein structure comparison. *Trends in biochemical sciences* **20**(11): 478-480.

Jo T, Cheng J. 2014. Improving protein fold recognition by random forest. *BMC bioinformatics* **15 Suppl 11**: S14.

Kabsch W. 1978. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **34**(5): 827-828.

Kim YJ, Patel JM. 2006. A framework for protein structure classification and identification of novel protein structures. *BMC bioinformatics* **7**: 456.

Kolodny R, Petrey D, Honig B. 2006. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Current opinion in structural biology* **16**(3): 393-398.

Krishna SS, Grishin NV. 2005. Structural drift: a possible path to protein fold change. *Bioinformatics* **21**(8): 1308-1310.

Nepomnyachiy S, Ben-Tal N, Kolodny R. 2014. Global view of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America* **111**(32): 11691-11696.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH--a hierarchic classification of protein domain structures. *Structure* **5**(8): 1093-1108.

Orengo CA, Pearl FM, Bray JE, Todd AE, Martin AC, Lo Conte L, Thornton JM. 1999. The CATH Database provides insights into protein structure/function relationships. *Nucleic acids research* **27**(1): 275-279.

Osadchy M, Kolodny R. 2011. Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America* **108**(30): 12301-12306.

Pascual-Garcia A, Abia D, Ortiz AR, Bastolla U. 2009. Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLoS computational biology* **5**(3): e1000331.

Pearl FM, Martin N, Bray JE, Buchan DW, Harrison AP, Lee D, Reeves GA, Shepherd AJ, Sillitoe I, Todd AE et al. 2001. A rapid classification protocol for the CATH Domain Database to support structural genomics. *Nucleic acids research* **29**(1): 223-227.

Richardson S, Green PJ. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)* **59**(4): 731-792.

Rogen P, Fain B. 2003. Automatic classification of protein structure by using Gauss integrals. *Proceedings of the National Academy of Sciences of the United States of America* **100**(1): 119-124.

Shindyalov IN, Bourne PE. 2000. An alternative view of protein fold space. *Proteins* **38**(3): 247-260.

Siew N, Elofsson A, Rychlewski L, Fischer D. 2000. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* **16**(9): 776-785.

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 research* **41**(Database issue): D490-498.

Swindells MB, Orengo CA, Jones DT, Hutchinson EG, Thornton JM. 1998. Contemporary approaches to protein structure classification. *BioEssays : news and reviews in molecular, cellular and developmental biology* **20**(11): 884-891.

Taylor WR, Orengo CA. 1989. Protein structure alignment. *Journal of molecular biology* **208**(1): 1-22.

Wang M, Caetano-Anolles G. 2009. The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure* **17**(1): 66-78.

Xu J, Zhang Y. 2010. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**(7): 889-895.

Yan RX, Si JN, Wang C, Zhang Z. 2009. DescFold: a web server for protein fold recognition. *BMC bioinformatics* **10**: 416.

Zemla A. 2003. LGA: A method for finding 3D similarities in protein structures. *Nucleic acids research* **31**(13): 3370-3374.

Zhang Y, Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**(4): 702-710.

-. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**(7): 2302-2309.

CONCLUSION

In this thesis, I addressed three questions of molecular evolution: (1) testing the compensatory hypothesis that explains why some human disease-associated residues (DARs) appear as wild-types in other species; (2) testing the functionality of translated pseudogenes and (3) classifying protein structure domains into folds in a continuous space continuity.

In Chapter 2, I identified potential compensatory residues for human DARs using structure information, and then demonstrated that they alleviate destabilizing effects of the DARs. Because reducing protein stability is a primary cause of disease by DARs, this observed alleviation in destabilizing effects indicates that the potential compensatory residues mitigate the harm of DARs. I also demonstrated that on average, the observed compensatory effects might be sufficient to restore normal phenotypes. Moreover, because compensatory residues are not necessarily close to the DARs, using only local compensatory residues here underestimates the compensatory effects. Taken together, my results strongly suggest that compensatory residues alleviate the disease-causing effects of DARs, supporting the compensatory hypothesis.

It would be interesting to know how a DAR and their compensatory residues get fixed in a population. In general, there are two scenarios. First, the compensatory residues are neutral or beneficial, and thus reach fixation in the population. Their fixations create a genetic background that alleviates deleterious effects of the DAR, allowing the spread of the DAR to the entire population. Second, the DAR exists with a low allele frequency in the population due to its mild fitness effect. During this time, compensatory residues appear and make the originally deleterious DAR beneficial or neutral. Therefore, the DAR together with the compensatory residues reach fixation. These two scenarios may be distinguished by examining the order with which the compensatory residues and the DAR appear in a phylogeny.

In Chapter 3, due to advances of mass spectrometry, low concentration proteins can be identified. The high-resolution human proteomes include 322 translated pseudogenes. Seventy-eight of them are qualified for evolutionary studies. The median nonsynonymous and synonymous rate ratio (ω) of these translated pseudogenes is significantly lower than that of transcribed pseudogenes and that of non-transcribed pseudogenes, respectively. Fifteen percent of the translated pseudogenes have ω significantly lower than 1. These results indicate purifying selection acting on this subset of translated pseudogenes, suggesting their functionality. I found that this subset of translated pseudogenes either have uninterrupted coding ORFs as their parental genes or use alternative start codons to express relatively complete protein domains, which are often independent functional units. The remaining 85% of translated pseudogenes have their median ω similar to non-translated pseudogenes. This indicates that their translations are either non-functional or falsely discovered by mass spectrometry.

The finding that 15% of translated pseudogenes are potentially functional is contradictory to their annotation as pseudogenes. Moreover, 8% of transcribed pseudogenes have ω significantly lower than 1. These selected pseudogenes might be missed by the mass spectrometry or expressed in tissues that were not screened. The selected translated and transcribed pseudogenes have at most 11% and 27% false discovery rates on average when multiple tests are corrected. Therefore, the truly selected pseudogenes still account for 11% and 6% of translated and transcribed pseudogenes, respectively. These results indicate that a sizeable fraction of annotated pseudogenes are potentially functional. Consequently, the current standard for pseudogene annotation needs to be modified. I suggest that ω should be calculated for pseudogene candidates. If the ω is significantly lower than 1, the pseudogene candidate is probably still functional even with signs of pseudogenization, and thus requires further inspection before being classified as a pseudogene.

In Chapter 4, I developed CEP/C3P to classify a domain structure into existing folds by considering the continuity of fold space. Depending on local/global similarity scores and fold schemes, CEP/C3P classifications differ from the current classifications for 4-12% of all domains. The differences confirm the continuous nature of the fold space and demonstrate impacts of the continuity on current fold classifications. Although CEP/C3P classification is in principle more reasonable than current fold classifications, its advantage is difficult to demonstrate in practice due to the lack of gold standards. A compromised benchmark is the classification based on sequence homology. However, there are two caveats for this validation. One is that many domains have no detectable homology to one another, and thus cannot be classified using sequence homology. The second is that homologous proteins do not necessarily have similar structures. These two facts impair the use of homology relationships for validating CEP/C3P classifications.

CEP/C3P takes a current classification as the initial fold classification, which was generated without considering fold space continuity. Thus, CEP/C3P is not able to remove the ignorance of the continuity completely from its classification. To fix this problem, a *de novo* classification is needed. For example, a Gaussian mixture distribution can be used to model the whole structure space. The number of Gaussian components in the mixture distribution corresponds to the number of folds. For instance, a Gaussian mixture distribution with K components partitions all the domains into K folds. Within each fold, a domain is assigned as the center (C) of the fold. The similarities between the central domain and the other domains in the fold are modeled by a Gaussian distribution. The key is to find the values of parameters *K*, *C*s and memberships of other domains to best fit the mixture distribution. This analytically impossible task can be achieved using Markov Chain Monte Carlo (MCMC) technique under Bayesian frameworks. Similarly to CEP and C3P, this *de novo* method classifies domains into folds by considering distributions of within-fold similarities, and thus takes fold space continuity into account for classifications.

In sum, Chapter 2 and 3 used structural genomic and proteomic data respectively to address important questions in molecular evolution, which are otherwise unapproachable. They demonstrate the usefulness of structural genomic and proteomic data in molecular evolution. In Chapter 4, I demonstrated the importance of considering fold space continuity in fold classifications, and developed CEP and C3P to assist current fold classification by considering this important factor. The more reasonable fold classifications with my method are expected to play important roles in the study of protein structures, functions, and evolution.

APPENDIX 1

HOW SIGNIFICANT IS A PROTEIN STRUCTURE SIMILARITY WITH TM-SCORE=0.5?

ABSTRACT

Motivation: Protein structure similarity is often measured by RMSD, GDT-score, and TM-score. However, the scores themselves cannot provide information on how significance the structural similarity is. Also, it lacks a quantitative relation between the scores and conventional fold classifications. This paper aims to answer two questions: (1) what is the statistical significance of TM-score? (2) What is the probability of two proteins having the same fold given a specific TM-score? We first made an all-to-all gapless structural match on 6,684 non-homologous single-domain proteins in the PDB and found that the TM-scores follow an extreme value distribution. The data allow us to assign each TM-score a P-value that measures the chance of two randomly selected proteins obtaining an equal or higher TM-score. With a TM-score at 0.5, for instance, its P-value is 5.5×10^{-7}, which means we need to consider at least 1.8 millions of random protein pairs to acquire a TM-score no less than 0.5. Second, we examine the posterior probability of the same fold proteins from three datasets SCOP, CATH and the consensus of SCOP and CATH. It is found that the posterior probability from different datasets has a similar rapid phase transition around TM-score=0.5. This finding indicates that TM-score can be used as an approximate but quantitative criterion for protein topology classification, i.e. protein pairs with TM-score>0.5 are mostly in the same fold while those with TM-score<0.5 are mainly not in the same fold

INTRODUCTION

Protein structure comparison is essential in almost every aspect of modern structural biology, ranging from experimental protein structure determination to computer-based protein folding and structure prediction, from protein topology classification to structure-based protein function annotation, and from protein-ligand docking to new compound screening and drug design (Kuntz 1992; Murzin et al. 1995; Orengo et al. 1997; Zhang 2009). The most commonly used means to compare protein structures is to calculate the root mean squared deviation, RMSD, of all the equivalent atom pairs after the optimal superposition of the two structures (Kabsch 1978). However, because all atoms in the structures are equally weighted in the calculation, one of the major drawbacks of RMSD is that it becomes more sensitive to the local structure deviation than to the global topology when the RMSD value is big. For example, the RMSD of two protein structures can be high if the tails or some loops have a different orientation even though the global topology of the core part is the same; this cannot be distinguishable, based on the RMSD value alone, from the case where two structures have completely different topologies.

Aiming at developing protein topology-sensitive measures, Zemla et al proposed a global distance test score, GDT-score (Zemla et al. 1999; Zemla 2003), which count the number of Cα pairs which have a distance below 1 Å, 2 Å, 4 Å, and 8 Å after the optimal superposition. This measurement has been used as one of the major criteria in the community-wide CASP experiments for assessing the modeling accuracy of structure predictions (Zemla et al. 1999; Moult et al. 2007). However, the distance cutoffs in the GDT score are subjective and may need to be manually tuned for different categories of modeling targets (Kopp et al. 2007). Moreover, similar as RMSD, the magnitude of GDT scores for random structure pairs has a power-law dependence with the protein length (Zhang and Skolnick 2004), which renders the absolute value of GDT-score less meaningful.

To address these issues, Zhang and Skolnick recently developed a template modeling score, TM-score (Zhang and Skolnick 2004), which counts all residue pairs using the Levitt-Gerstein weight (Levitt and Gerstein 1998) and therefore does not need discrete distance cutoffs. Since the short distance in the Levitt-Gerstein matrix is weighted stronger than the long distance, TM-score is more sensitive to the global topology. Moreover, because it adopts a protein size-dependence scale to normalize the residue distances, the magnitude of TM-score of random structure pairs is protein size independent.

Despite the advantage and usefulness of RMSD, GDT- and TM-scores in quantitatively measure of protein structure similarities, the scores themselves does not tell the statistical significance of each score value, which is essential in many of the statistical studies of protein structure comparisons and alignment analysis (Levitt and Gerstein 1998; Sadreyev et al. 2009). As another highly-related issue, proteins have been categorized into various structural families based on the structure and/or evolutionary similarities, using either human visual intuition (Murzin et al. 1995) or semi- or fully-automated structural comparisons (Holm et al. 1992; Orengo et al. 1997). These hierarchical databases provide important facilities to our understanding of protein structure and function, and gauge the structural comparison and categorizations. However, it generally lacks a quantitative correspondence between the structural similarity scores and the various levels of protein structure categorizations. For example, a simple but often-asked question in protein structure prediction and assessment is: does the predicted model have the correct fold (compared to the native structure) given all the RMSD, GDT-score, and TM-score?

In this work, we try to address these issues by answering two questions: (1) what is the statistical significance of each TM-score value; and (2) what is the probability of two proteins to have the same fold given the TM-score. Here, the reason for us to choose TM-score is that the magnitude of TM-score is protein size independent, which facilitates the attainment of length-independent analytical results of the calculations. Although our focus in the second question is on the fold level, the results can be easily extended to other level of structural similarities.

MATERIALS AND METHODS

Definition of TM-score

TM-score is defined to assess the topological similarity of two protein structures (Zhang and Skolnick 2004):

$$\text{TM-score} = \frac{1}{L}\left[\sum_{i=1}^{L_{ali}} \frac{1}{1+d_i^2/d_0^2}\right]_{\max} \quad (1)$$

where L is the length of the target protein, L_{ali} is the number of the equivalent residues in two proteins. d_i is the distance of the *i*th pair of the equivalent residues between the two structures, which depends on the superposition matrix; the 'max' means the procedure to identify the optimal superposition matrix to maximize the sum in Eq. (1). The scale $d_0 = 1.24\sqrt[3]{L-15} - 1.8$ is defined to normalize the TM-score in a way that the average magnitude of the TM-score for random protein pairs is independent on the size of the proteins. TM-score stays in (0, 1] with a higher value indicating a stronger similarity.

Dataset of random protein structure pairs

6,684 single domain structures were culled from the PDB database (Berman et al. 2002). These proteins share a low pair-wise sequence similarities (with sequence identity < 25%), as filtered by PISCES (Wang and Dunbrack 2003), with the protein length between 80-200 amino acids.

Each protein structure in this dataset is used as the target protein to compare with all the other proteins in the dataset with the same or longer length. For each protein pair, the target protein is first superposed by the TM-score program on the N-terminal fragment of the bigger protein structure with the TM-score normalized by the target protein. The target sequence then slides gaplessly along the sequence of the bigger protein with a window size of 20 residues until less than 20 amino acids remain on the larger protein; a TM-score is obtained with each of the gapless alignments. This procedure on the dataset ends up with a total of 71,583,085 random and protein-like structure pairs.

It should be mentioned that the TM-score superimpositions are obtained from a set of gapless sliding alignments rather than from the optimal structural alignments of the two proteins. The purpose of the gapless alignment is for generating random structure background, because a structural alignment, produced by such as Dali (Holm and Sander 1995) and TM-align (Zhang and Skolnick 2005), for instance, usually represents an optimal match of a given pair of protein structures that is selected from a huge number of possible combinations of corresponding residues assignments. Thus, a structural alignment does not constitute random structural comparisons even though the non-homologous protein pairs are randomly selected.

Dataset of proteins with same/different folds

To estimate the posterior probability for structure pairs at given TM-scores to share the same topology, a collection of protein pairs in both the same and the different folds is necessary. For this purpose, we borrow the Fold and Topology definition from the standard protein structure classification databases: SCOP (Murzin et al. 1995) and CATH (Orengo et al. 1997), to generate the same and different fold datasets.

Three sets of same fold structure pairs. The first set of protein domains (Set-I) are collected from the SCOP 1.73 database (Murzin et al. 1995). After filtering out the redundant proteins with a sequence identity > 95% and the small proteins with length below 80 amino acids, 11,239 protein domains remain, which cover 551 main Fold families in SCOP. An all-to-all pairing is then carried out for the proteins within the same Fold family and ends up with a total of 746,420 protein pairs which are considered as sharing the same fold in SCOP. Many of the same fold domains, however, have only similar core regions but include some long-tails and outlier super-secondary structures that have different orientation and structures. To rule out the possible contamination of the outliers on the structure scores, we further remove those domain pairs that have a radius gyration difference larger than 10%. Thus, 449,281 valid structure pairs are finally obtained from the SCOP library at the "Fold" level.

The second set of protein domains (Set-II) are from CATH 3.2.0 (Orengo et al. 1997). The structure pairs are generated from the proteins in the same "Topology", a structural level equivalent to the "Fold" in SCOP (Hadley and Jones 1999). After the same redundancy and length filtering, 12,248 domains covering 598 main Topologies in CATH are obtained. An all-to-all pairing among proteins of the same Topology families result in 2,769,868 domain pairs. By applying the radius gyration cut-off, 1,360,782 pairs were left for the domains at the CATH "Topology" level.

The third protein pair set (Set-III) is a consensus of the SCOP and CATH databases. Due to the different domain splitting system, SCOP and CATH may have protein domains with the same ID (the same PDB names and chains) but having different sequence segments. To ensure that SCOP and CATH deal with the same structures, we filter out those inconsistent domains and collect only the structures which have the same IDs in the SCOP and CATH and meanwhile have a sequence identity >90% between the SCOP and CATH domains. By these criteria, 5,105 domain structures are culled from SCOP with counterparts in CATH, which cover 328 different fold families. An all-to-all pairing is carried out among the proteins which are consistently defined by SCOP and CATH as the same fold, resulting in 186,359 protein pairs. After the radius gyration filtering, 117,446 pairs are finally collected and used.

Three sets of differnt fold structure pairs. There are three sets of different (or non-same) fold protein pairs corresponding to the same fold pairs in Set I, II and III. Due to the big size of the protein sets, we found that the TM-score distributions for non-same fold proteins are very similar for different protein sets. Therefore, we generate all the non-same fold protein pairs from the well-defined and consensus set of the 5,105 protein domains.

The first non-same fold protein set is named Set-I'. It contains an all-to-all pairings of the 5,105 protein domains but excluding all the pairs that are in the same SCOP Fold family, which results in 12,815,737 protein pairs. The Set-II' is similar as Set-I' but excluding the domain pairs that are in the same CATH Topology family, which results in 12,507,855 protein pairs. To generate Set-III', any pairs which are either in the same SCOP Fold family or in the same CATH Topology family are excluded. This results in 12,497,203 protein pairs.

RESULTS

Statistical significance of TM-score

Extreme Value Distribution (EVD) is often used to model the smallest or largest value among a large set of independent, identically distributed random values (Embrechts et al. 1997).

It has been shown that both sequence and structure comparison scores of proteins follow the EVD (Levitt and Gerstein 1998). The general function of EVD is described as

$$y = f(x|\mu,\sigma) = \sigma^{-1}\exp\left(\frac{\mu-x}{\sigma}\right)\exp\left(-\exp\left(\frac{\mu-x}{\sigma}\right)\right) \quad (2)$$

where μ is the so-called location parameter and σ is the scale parameter.

In Figure A.1.1, we show the distribution of TM-score values calculated from 71,583,085 random protein pairs which are collected from 6,684 non-homologous proteins in the PDB library by gapless threading. The distribution matches well to the Eq. (2) with the best fitting parameter μ=0.1512 and σ=0.0242 estimated by the Maximum Likelihood method. We also split the protein samples into 4 groups according to the protein size, i.e. [80, 100], [101, 120], [121, 160], [161, 200]; all of them follow well the same EVD. This data on one hand demonstrates the robustness of the extreme value distribution for the TM-score distribution of random protein pairs. On the other hand, it confirms the previous conclusion that the TM-score magnitude and distribution of random proteins are independent of protein size (Zhang and Skolnick 2004).

We are interested in the probability of having a TM-score equal to or greater than a certain value (*x*) among random protein pairs, i.e. P-value of a TM-score. The P-value can be obtained by integrating Eq. (2) from *x* to 1, i.e.

$$P\text{-value}(x) = \int_x^1 f(x|\mu,\sigma) = 1 - \exp\left(-\exp\left(\frac{\mu-x}{\sigma}\right)\right) \quad (3)$$

Figure A.1.2 shows the overall shape of the P-value versus TM-score with μ and σ taken from the data in Figure A.1.1. In general, the probability to find a TM-score ≤0.17 from random structural pairs is close to 1. The P-value then decreases rapidly with TM-score >0.17; it becomes significantly below 1 when TM-score >0.3. In the Inset of Figure A.1.2, we plot the P-value for the TM-score range in [0.3, 1], which follows approximately an exponential regression. When TM-score=0.5, it corresponds to a P-value of 5.5×10^{-7}.

Many authors have demonstrated that the magnitude of RMSD, GDT-score, and several other matrices are all protein length-dependent (Levitt and Gerstein 1998; Betancourt and Skolnick 2001; Ortiz et al. 2002; Zhang and Skolnick 2004). A basic assumption of this work is that the magnitude of TM-score is protein length-independent, which gives the opportunity to express the P-value as a sole function of TM-score. Figure A.1.3 shows explicitly the average TM-score value and the deviations with different protein size, where a bin-width of protein size=10 residues is taken. The data again confirm the size independence of the TM-score values in random protein pairs.

As an intuitive explanation of the P-value, we also present in Figure A.1.3 the number of random protein pairs which are needed to achieve or surpass certain TM-score values; this is converted from the P-value data shown in Figure A.1.2. For TM-score=0.5, for example, it needs at least 1.8 millions of random structural matches so that one structure match can hit a TM-score equal to or higher than 0.5. When TM-score=0.72, this number increases to 10 billion.

TM-score of proteins with the same fold

Although the P-value can give a quantitative measure of the statistical significance of each TM-score value, researcher often wants to know what TM-score approximately corresponds to the protein pairs sharing the same fold. For example, an often-asked question in *ab initio* and template-based protein structure prediction is how to judge whether a predicted model has the same fold or topology as the experimental structure (Jauch et al. 2007; Kopp et al. 2007; Zhang 2009). Here, we address this issue by calculating the Posterior Probability for proteins at certain

TM-score to have the same or different folds. We will examine the results of the posterior probabilities using three different fold/topology standards.

TM-score corresponding to the SCOP Fold level. According to the Bayesian rules, for a given TM-score, the posterior probability of proteins having the same or different Fold can be expressed as:

$$\begin{cases} P(F|TM)=\dfrac{P(TM|F)P(F)}{P(TM|F)P(F)+P(TM|\overline{F})P(\overline{F})} \\ P(\overline{F}|TM)=\dfrac{P(TM|\overline{F})P(\overline{F})}{P(TM|F)P(F)+P(TM|\overline{F})P(\overline{F})} \end{cases} \tag{4}$$

Here, *TM* stands for the TM-score of the compared proteins as calculated by the structural alignment program TM-align (Zhang and Skolnick 2005); *F* and $\overline{F}$ represent the events that the proteins belong to the same and different Fold in SCOP, respectively; $P(F)$and $P(\overline{F})$ are the prior probabilities of proteins in same and different folds; $P(TM\mid F)$ and $P(TM\mid\overline{F})$ are the conditional probabilities of TM-score when the two proteins are in the same or different Fold families, respectively.

In the Set I and Set-I’, 449,281 pairs of proteins are considered by SCOP1.73 as the same Fold and 12,815,737 are as not in the same Fold. Thus, the conditional probabilities can be calculated by

$$\begin{cases} P(TM\mid F)=\dfrac{N(TM)}{\sum\limits_{TM=0}^{1}N(TM)} \\ P(TM\mid\overline{\mathrm{F}})=\dfrac{\overline{N}(TM)}{\sum\limits_{TM=0}^{1}\overline{N}(TM)} \end{cases} \tag{5}$$

where $N(TM)$ is the number of protein pairs with a certain TM-score (*TM*) in the Set-I, and $\overline{N}(TM)$ is the number of protein pairs with the TM-score in the non-same fold protein Set-I’. The denominators are the summation of the same and non-same fold protein pairs for all TM-scores in (0, 1], which equals to the total number of protein pairs in Set I and Set-I' respectively.

In Figure A.1.4 ('squares'), we divide the TM-score space into 20 bins and present the conditional probability for both the same and non-same fold proteins. As expected, the same Fold and the non-same Fold proteins are well grouped in different TM-score ranges. However, since TM-score and SCOP fold is not a one-to-one correspondence, there is a small overlap of TM-score between the two protein data sets.

To calculate the prior probabilities of $P(F)$ and $P(\overline{\mathrm{F}})$, for the purpose of minimizing statistical bias, we collect all 85,685 protein domains in the SCOP database. An all-to-all pairing is then carried out on these proteins. The prior probability can be calculated by

$$\begin{cases} P(F) = \dfrac{N(\mathrm{F})}{N(F) + N(\overline{F})} \\ P(\overline{F}) = 1 - P(F) \end{cases} \tag{6}$$

where $N(\mathrm{F})$ and $N(\overline{F})$ are the number of the same Fold and the non-same Fold pairs according to the SCOP definition. Overall, $P(F)$=0.0142 and $P(\overline{\mathrm{F}})$=0.958 in our counting.

Figure A.1.5 ('squares') is the posterior probability for proteins of certain TM-score to be in the same SCOP Fold, when we integrate the data of Eqs. (5) and (6) into Eq. (4). When TM-score <0.4, there almost never have a pair of proteins which are in the same SCOP Fold family. On the other hand, when TM-score >0.6, the probability of the two proteins in the same SCOP Fold rapidly increases to >65%. There is a clear phase transition occurring around the half score of TM-score.

TM-score corresponding to the CATH Topology level. Since the fold definition can be dependent on subjective choices, to examine the robustness of the TM-score distribution, we calculate the posterior probability using another widely-used database, CATH (Orengo et al.

1997). 1,360,782 protein pairs are considered by CATH as of the same Topology in Set-II and 12,507,855 pairs as of different Topology in Set-II'.

In Figure A.1.4 (triangles), we show the conditional probabilities of the same and non-same fold protein pairs in the CATH database. Compared with the SCOP data, there is a clear shift of the distribution towards smaller TM-score, which indicates that the fold definition in CATH Topology is on average broader than that in SCOP Fold, although the non-fold protein distribution of CATH is similar as that of SCOP. Correspondingly, the prior probability $P(F)$ of CATH Topology calculated from all the 114,125 CATH domains based on Eq. (6) is 0.0299, which is higher than that of SCOP (0.0142), because more protein pairs are categorized into the same Topology families due to the broader structural cut-off in CATH. The prior probability of the non-same Fold proteins $P(\bar{F})$=0.97.

Figure A.1.5 (triangles) shows the posterior probability of protein pairs to be in the same CATH Topology families with given TM-scores. There is a slight shift of CATH compared with SCOP towards smaller TM-score as well; but a similar rapid phase transition is observed in TM-score between 0.4 and 0.6.

TM-score corresponding to the consensus SCOP&CATH fold families. Due to the slight inconsistence between SCOP and CATH database, we here consider a consensus set of protein pairs, Set-III, where the proteins are considered as the same Fold by both SCOP and CATH, which covers 328 consensus structural families. The non-same fold protein pairs (Set-III') are those where both SCOP and CATH categorize them into different structural families.
As shown in Figure A.1.4 (stars), the conditional probabilities of TM-score for proteins in the same families in the consensus set are slightly shifted towards larger TM-score related to SCOP,

because of the even tighter definition of the fold family. Similarly, the prior probability for the same fold $P(F)$= 0.0149 while $P(\bar{F})$=0.985 for the non-same fold proteins.

In Figure A.1.5 (stars), we present a posterior probability of proteins at the same fold and non-same Fold families based on the stricter and consensus definition from both SCOP and CATH. There is again a rapid phase transition around TM-score=0.5. Compared with SCOP and CATH, however, this transition is more rapid.

Combining the results of the three different datasets, it seem quite safe to assign TM-score=0.5 as a rough but quantitative cutoff for protein Fold/Topology definition, i.e. most of proteins with TM-score > 0.5 can be considered as of the same topology whereas most proteins with a TM-score < 0.5 should be of different topology. When the TM-score is further away from the cutoff value, the conclusion becomes gradually safer. When TM-score=0.4, for example, > 99.9% of proteins are not in the same fold according to the consensus definition of SCOP and CATH; when TM-score = 0.6, > 90% of proteins are in the same fold based on the consensus criterion.

DISCUSSION AND CONCLUSION

We first examined the TM-score distribution of randomly-selected non-homologous protein pairs using gapless threading and found that it follows a simple extreme value distribution. This allows us to calculate P-value to estimate the statistical significance of each TM-score value. When TM-score <0.17, the P-value is close to 1, which means that any protein structures or computer models at this level of similarity is indistinguishable from random structure pairs. The P-value decreases rapidly below 0.001 when TM-score>0.3, a region of structural similarity which is significantly different from random structures. When TM-score=0.5, the P-value is reduced to 5.5×10^{-7}, meaning that at least 1.8 millions of random protein pairs are needed to achieve on average one of this similarity.

It should be noted that this data does not contradict with a previous study (Zhang et al. 2006) where the average TM-score of the structural alignment by TM-align on random structure pairs is around 0.3, because the structural alignment in TM-align (Zhang and Skolnick 2005) has been optimally selected from a huge number of possible structural matches, which is therefore far from random structural matches although the protein structure pairs are randomly selected. Interestingly, in the recent CASP7 and CASP8 blind protein structural predictions, the average TM-score of the worst three models for each target are 0.161±0.041 and 0.168±0.042, respectively (data taken from http://zhang.bioinformatics.ku.edu/casp7 and http://zhang.bioinformatics.ku.edu/casp8); both are below and near 0.17. This means that the predicted models from these bottom groups are not more than a random pickup of structures from the PDB library.

Second, we developed an approach for estimating the posterior probability of proteins with given TM-scores to be in the same or different fold family. Using three different datasets which has Fold/Topology defined from the standard SCOP and CATH databases, we observed a similar rapid phase transition of the probability around TM-score=0.5. This indicates that TM-score may be used as a quantitative criterion for assessing whether protein structures or model predictions are of the same topology, i.e. for TM-score <0.5, proteins are mostly not in the same fold while for TM-score >0.5, proteins are generally in the same fold. This criterion becomes gradually safer when the actual TM-score reaches a value further away from the cut-off.

It should be mentioned that this TM-score cutoff may not be directly applicable for comparing protein structures in the CATH and SCOP databases, i.e. there are actually considerable proteins in the same fold family in SCOP and CATH which may have a TM-score below 0.5; this is mainly due to the outlier of protein structures, such as long tails. In our culled dataset and calculations, we only focused on the topology of the core regions and had the abnormal protein domains filtered out. Therefore, when applying TM-score in real protein comparison, one should either cut the structural outliers or normalize the TM-score in the correct target length that only corresponds to the important core structures.

The second part of the studies in this paper has been focused on the fold level of protein structures, which is mainly because this concept of topology has been most generally used in protein folding and protein structure prediction. Also, this category of structure similarity is clearly defined and has equivalency in both SCOP and CATH databases (Hadley and Jones 1999). Nevertheless, the extension of our approach to other levels of homologous family and super-family should be straightforward.

ACKNOWLEDGEMENTS

The project is supported in part by the Alfred P. Sloan Foundation, NSF Career Award (DBI 0746198), and the National Institute of General Medical Sciences (GM083107, GM084222).

Figure A.1.1. TM-score distribution of gapless comparisons among non-homologous protein structures. The continuous curve represents an extreme value distribution with the location parameter and the scale parameter being 0.1512 and 0.0242 respectively; the Chi-square error of fitting is 0.001. The TM-score distributions of 4 subdivisions are from proteins with length in [80, 100], [101, 120], [121, 160], and [161, 200], respectively.

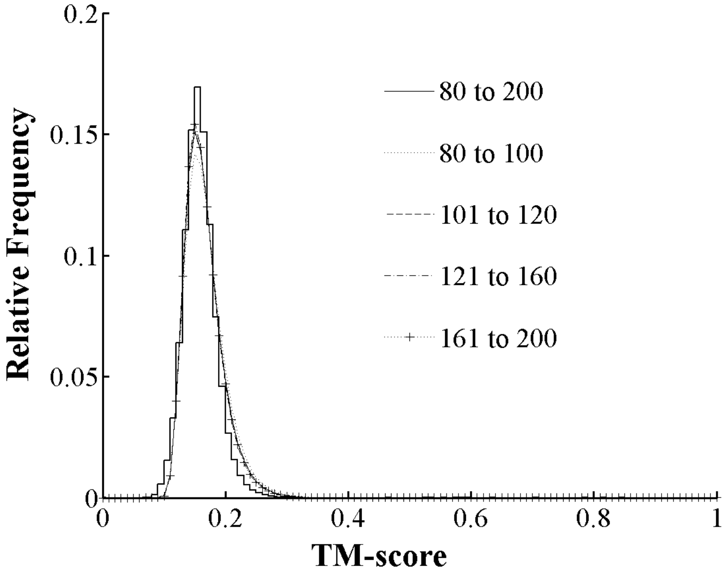


Figure A.1.2. The p-value versus TM-score. The curve is a sigmoid like Boltzmann function with chi square equal to 0.0001. Inset: P-value (in logarithm scale) vs. TM-score in [0.3, 1].

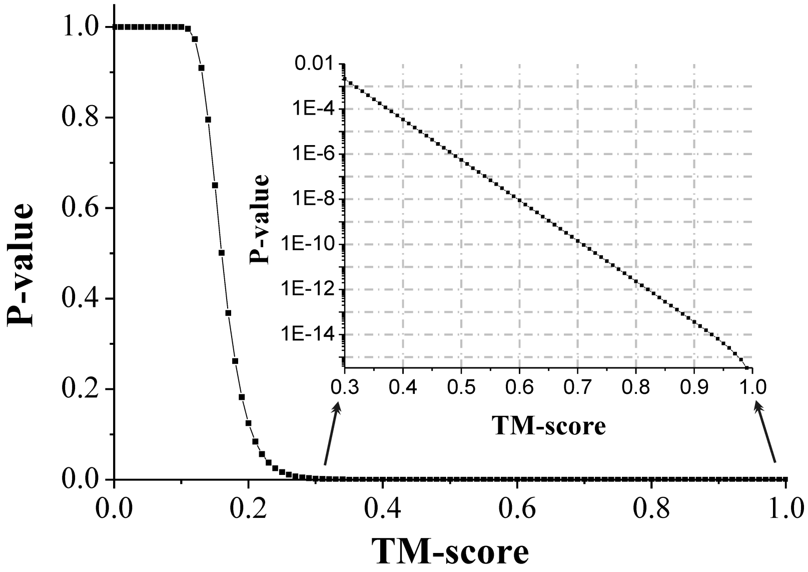


Figure A.1.3. The average TM-scores (with error bars) of random structure matches with protein length from 80 to 200 amino acids. The straight and dash lines above TM-scores=0.2 show the number of random protein pairs (values on the right-hand side) needed to achieve or surpass a certain TM-score level. By doing random structure comparisons in 10^2, 10^4, 10^{10}, and 10^{16} times, one can hit a match with a TM-score equal to or above 0.263, 0.374, 0.709, and 0.977, respectively. 1.8×10^6 random matches are needed to achieve a TM-score ≥0.5.

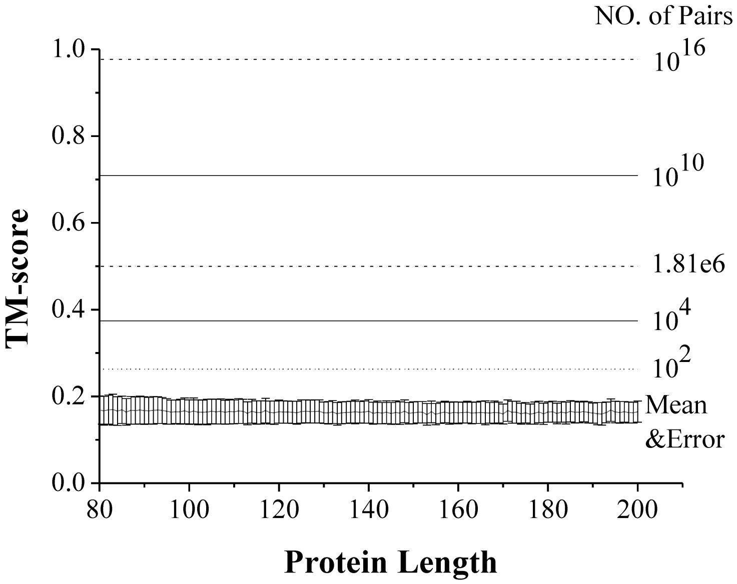


Figure A.1.4. Conditional probability versus TM-score. The conditional probabilities of TM-score for proteins in the same fold and different fold families as defined by SCOP (Set I, Set I'), CATH (Set II, Set II') and SCOP&CATH (Set III, Set III').

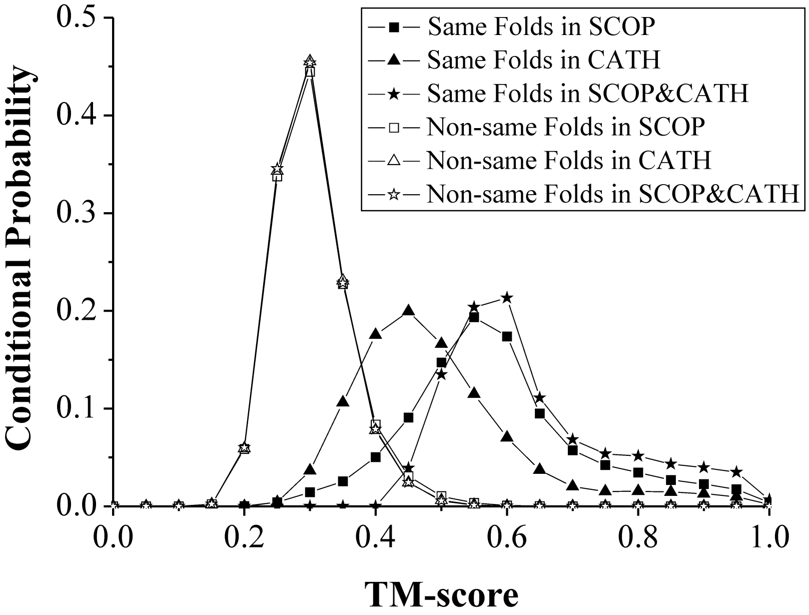
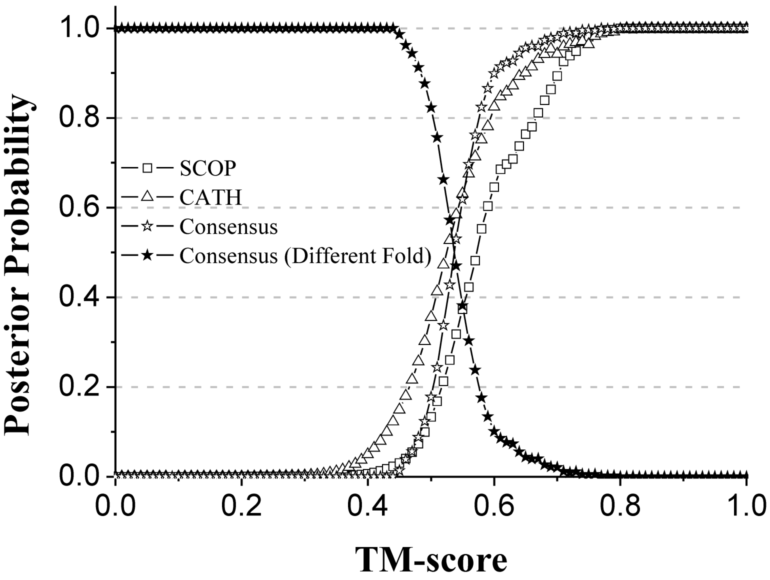


Figure A.1.5. **The posterior probability of proteins with a given TM-score to be in the same Fold (open points) or non-same Fold family (solid points).** The Fold family is defined based on either the SCOP Fold level, or the CATH Topology level, or a consensus of SCOP Fold and CATH Topology families.



REFERENCES

Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S. 2002. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography* **58**(6): 899-907.

Betancourt MR, Skolnick J. 2001. Universal similarity measure for comparing protein structures. *Biopolymers* **59**(5): 305-309.

Embrechts P, Kluppelberg C, Mikosch T. 1997. *Modelling extremal events for insurance and finance*. Berlin: Spring Verlag

Hadley C, Jones DT. 1999. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure* **7**(9): 1099-1112.

Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. 1992. A database of protein structure families with common folding motifs. *Protein Sci* **1**(12): 1691-1698.

Holm L, Sander C. 1995. Dali: a network tool for protein structure comparison. *Trends Biochem Sci* **20**(11): 478-480.

Jauch R, Yeo HC, Kolatkar PR, Clarke ND. 2007. Assessment of CASP7 structure predictions for template free targets. *Proteins* **69**(S8): 57-67.

Kabsch W. 1978. A discussion of the solution for the best rotation to relate two sets of vecotrs. *Acta Cryst* **A 34**: 827-828.

Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. 2007. Assessment of CASP7 predictions for template-based modeling targets. *Proteins* **69**(S8): 38-56.

Kuntz ID. 1992. Structure-based strategies for drug design and discovery. *Science* **257**(5073): 1078-1082.

Levitt M, Gerstein M. 1998. A unified statistical framework for sequence comparison and structure comparison. *Proc Natl Acad Sci U S A* **95**(11): 5913-5920.

Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. 2007. Critical assessment of methods of protein structure prediction-Round VII. *Proteins* **69 Suppl 8**: 3-9.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**(4): 536-540.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH--a hierarchic classification of protein domain structures. *Structure* **5**(8): 1093-1108.

Ortiz AR, Strauss CE, Olmea O. 2002. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci* **11**(11): 2606-2621.

Sadreyev RI, Kim BH, Grishin NV. 2009. Discrete-continuous duality of protein structure space. *Curr Opin Struct Biol* **19**(3): 321-328.

Wang G, Dunbrack RL. 2003. PISCES: a protein sequence culling server. Vol 19, pp. 1589-1591. Oxford Univ Press.

Zemla A. 2003. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* **31**(13): 3370-3374.

Zemla A, Venclovas C, Moult J, Fidelis K. 1999. Processing and analysis of CASP3 protein structure predictions. *Proteins* **Suppl 3**: 22-29.

Zhang Y. 2009. Protein structure prediction: when is it useful? *Current opinion in structural biology* **19**(2): 145-155.

Zhang Y, Hubner I, Arakaki A, Shakhnovich E, Skolnick J. 2006. On the origin and completeness of highly likely single domain protein structures *Proc Natl Acad Sci USA* **103**: 2605-2610.

Zhang Y, Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**: 702-710.

-. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**(7): 2302-2309.

APPENDIX 2

SUPPLIMENTARY FIGURES AND TABLES

Figure A.2.1.1 Linear regression between predicted and experimentally measured stability reductions caused by point mutations. Experimentally determined stability reductions caused by 1201 mutations are provided by the authors of Rosetta (Kellogg et al. 2011). Predicted stability reductions are calculated by Rosetta using the parameters described in Materials and Methods. The linear regression forced to go through the origin is obtained using the robust package in R (http://cran.r-project.org/web/packages/robust/index.html). REU, Rosetta Energy Unit.

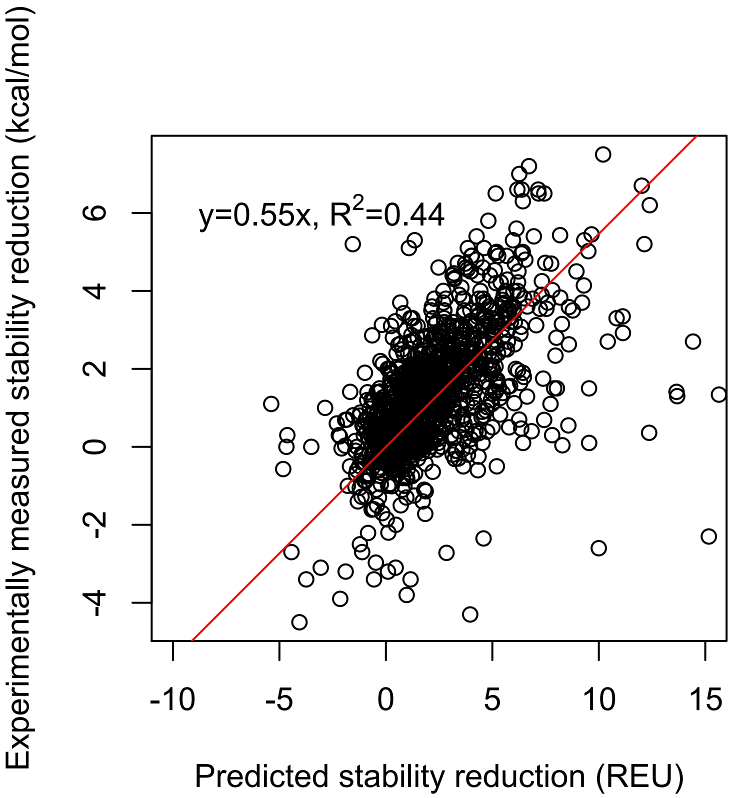


Figure A.2.1.2 Correlation between log(*p*-value) and log(sample size) in the comparison between mutations to wt-DARs and rg-DARs. Each circle represents a mutation type listed in Table A.2.1.1. For each mutation type, protein stability reductions upon mutations to wt-DARs and rg-DARs are compared by a one-tail Mann-Whitney U test. Mutation types with *p*-values < 0.05 are shown as red circles.

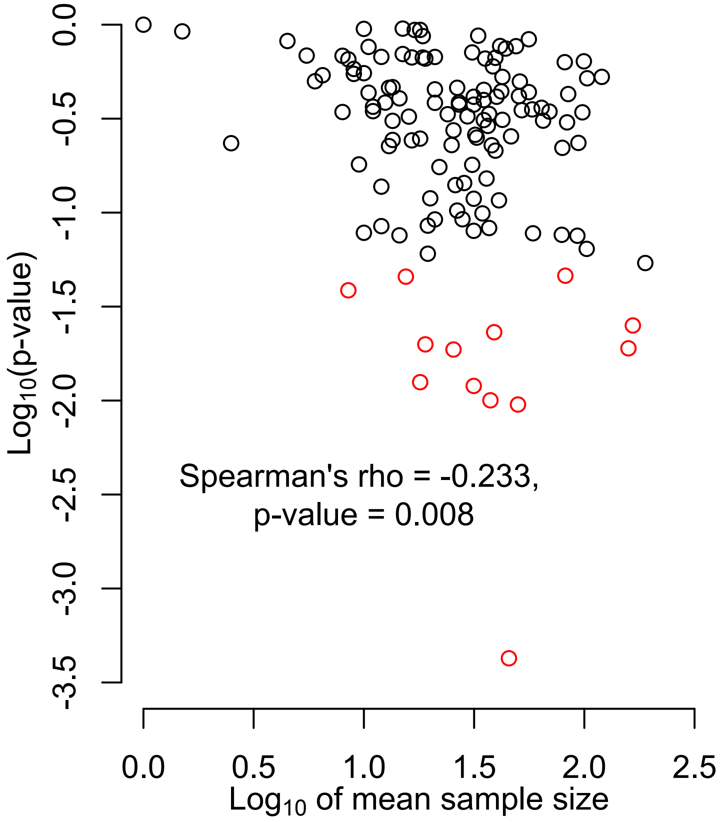


Figure A.2.1.3 Frequency distribution of the mean difference between $\Delta\Delta G_1$ and $\Delta\Delta G_2$ averaged over protein. The larger the difference, the greater the compensation effect. Destabilizing wt-DARs have $\Delta\Delta G_1 > 1$ kcal/mol. Arrows indicate median values of the corresponding 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. This figure is the same as Fig. 2.3 except that, when a protein harbors multiple wt-DARs, we respectively averaged $\Delta\Delta G_1$ and $\Delta\Delta G_2$ values from different wt-DARs in the same protein before comparison.

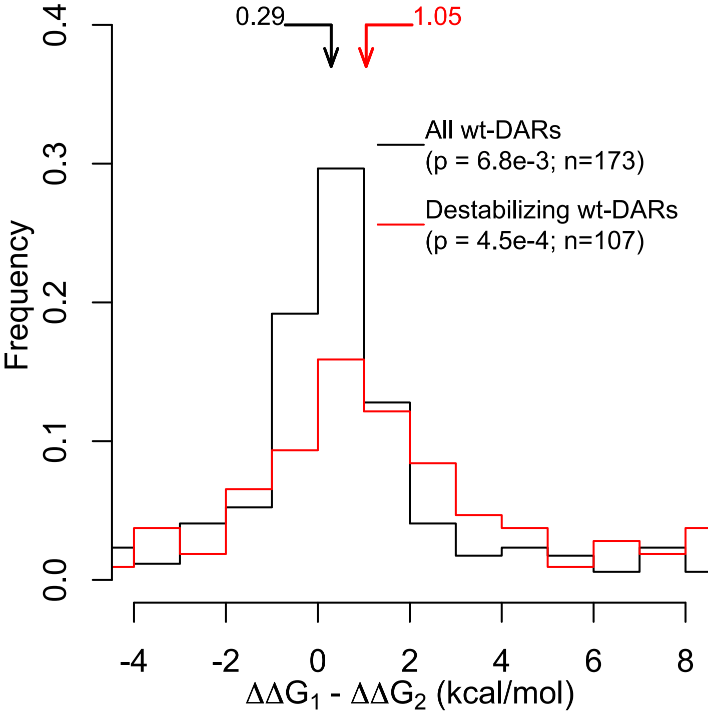


Figure A.2.1.4 Frequency distribution of the mean difference between $\Delta\Delta G_1$ and $\Delta\Delta G_2$ for non-redundant protein set. The larger the difference, the greater the compensation effect. Destabilizing wt-DARs have $\Delta\Delta G_1 > 1$kcal/mol. Arrows indicate median values of the 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. This figure is the same as Fig. 2.3 except that only a subset of protein structures with pairwise sequence identity <60% are used.

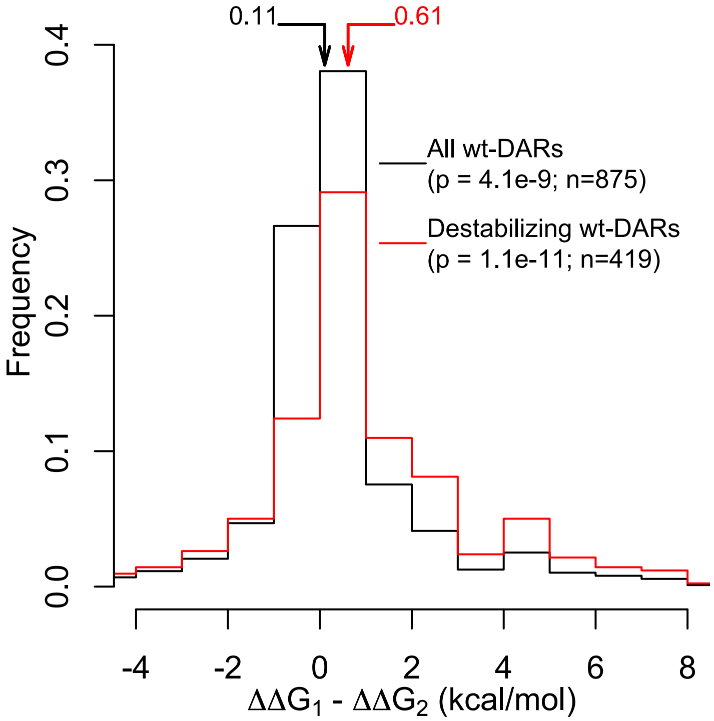


Figure A.2.1.5 Radial distribution of residue densities in protein structures. The residue density in each bin (bin size is 0.1 Å) for a focal residue is the number of residues that fall into the bin divided by the volume corresponding to the bin. This density is then averaged over all residues of all proteins.

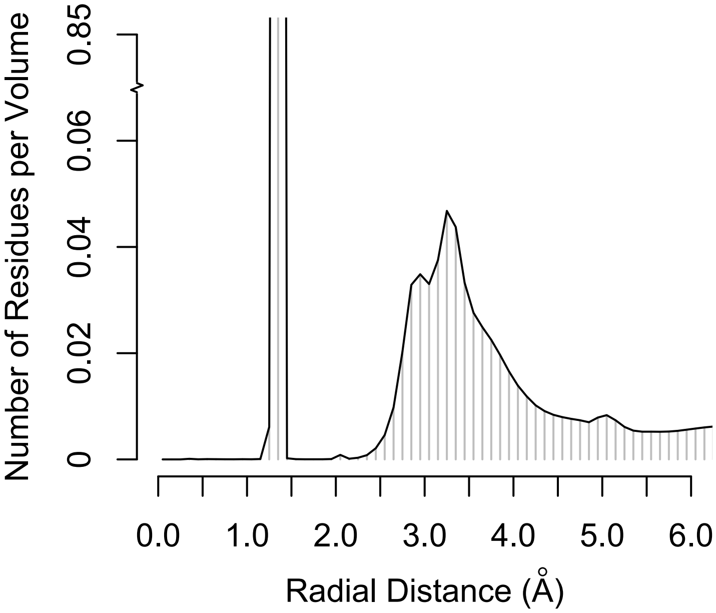


Figure A.2.1.6 Alignment of plasminogen sequences. The disease-associated residue (DAR) is at the underlined position, where the wild-type residue in human is R and the DAR is H. The residues highlighted in green are the potential compensatory neighboring residues in panda. The residues in red are the non-compensatory neighboring residues in horse. All protein sequences are from UniProt. Structure 2KNFA in PDB is used to model the structure of human plasminogen. The residue positions are assigned according to P00747 in UniProt.

```
                                           4
                                           9
                              78901234567890123456789012345678901234567890123456
Structure (Human)             SNADCMFGNGKGYRGKRVTTVTGTPCQDWAAQEPHRHSIFTPETNPRAGL
P00747 (Human)                SEEDCMFGNGKGYRGKRATTVTGTPCQDWAAQEPHRHSIFTPETNPRAGL
G1MBX3 (Panda)                SETDCMFGNGKGYRGKKATTVLGIPCQEWTAQEPHKHSIFTPETNPRAGL
F6USP9 (Horse)                SEPDCMLGIGKGYQGKKATTVTGTRCQAWAAQEPHRHSIFTPEANPWANL

                                 5
                                 3
                              7890123456789012345678901234567890----123
Structure (Human)             EKNYCRNPDGDVGGPWCYTTNPRKLYDYCDVPQC----AAP
P00747 (Human)                EKNYCRNPDGDVGGPWCYTTNPRKLYDYCDVPQC----AAP
G1MBX3 (Panda)                EKNVSHFDFLDVNGPWCYTTNPRKLFDYCDIPQCVCATASG
F6USP9 (Horse)                EKNYCRNPDGDVNGPWCYTMNPQKLFDYCDVPQC----ESS
```

Figure A.2.2.1 Rank correlations between various properties of a fold and the number of domains reclassified into the fold by C3P. The domains are reclassified (**A-C**) global TM_{max}-score-based C3P and (**D-F**) local TM_{max}-score-based C3P. The lines show linear regressions. ρ, Spearman's rank correlation coefficient.

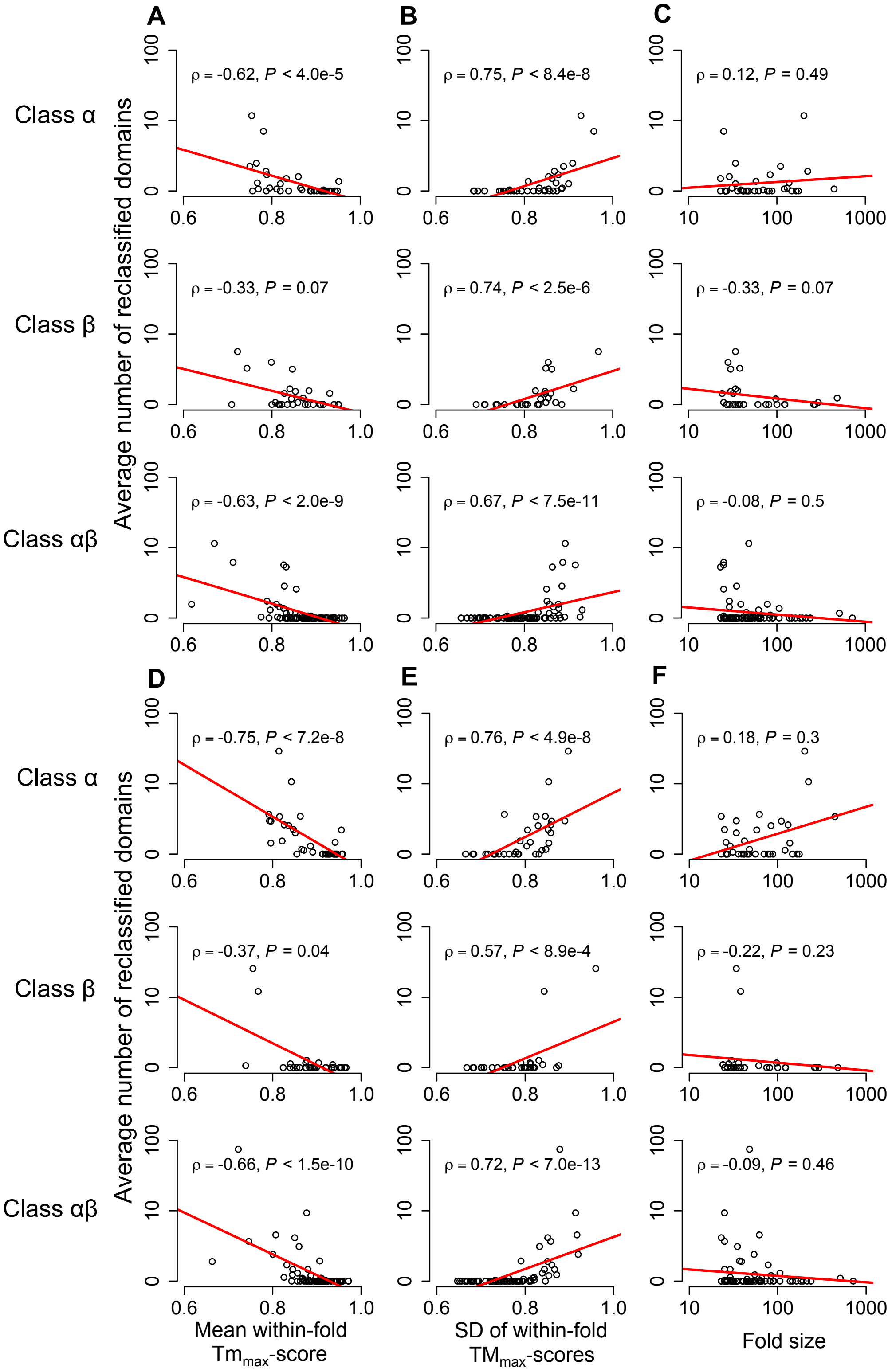


Table A.2.1.1. List of the 128 mutation types shared by mutations to wt-DAR and rg-DAR. *p*-values are from Mann-Whitney U test (one-tail).

Wild type to Mutant	Average Sample Size	P-value
R to P	45.5	4.00E-04
V to A	50	0.0095
S to F	37.5	0.01
S to N	31.5	0.012
A to E	18	0.0125
V to F	25.5	0.0187
G to R	158.5	0.019
R to K	19	0.0199
N to K	39	0.0231
E to K	166	0.0251
K to Q	8.5	0.0386
L to Q	15.5	0.0456
G to D	82	0.0462
L to P	189	0.0539
H to D	19.5	0.0604
A to T	102.5	0.0641
P to L	93	0.0753
T to N	14.5	0.0757
I to T	79	0.0762
G to V	58.5	0.0776
T to S	10	0.0781
P to R	31.5	0.08
C to R	37	0.0828
M to L	12	0.0847
F to C	19.5	0.0853
Q to H	21	0.092
R to S	28	0.0922
K to N	34.5	0.0991
H to Q	26.5	0.1025
M to T	41	0.1163
W to C	31.5	0.1185
V to E	20	0.1192
R to T	12	0.1375
E to D	26	0.1401
I to N	28.5	0.1437
A to D	36	0.1518
K to R	22	0.1748
Q to R	31	0.1796

W to S	9.5	0.1805
W to R	39.5	0.2137
V to M	79.5	0.2212
S to T	13	0.2259
D to V	38	0.2283
M to R	25	0.2289
L to I	2.5	0.2341
D to N	94	0.235
I to S	16.5	0.2424
Y to D	13.5	0.2437
C to F	18	0.2476
M to V	32.5	0.2507
R to G	46.5	0.2545
V to G	32	0.2595
L to S	25.5	0.2741
D to E	36.5	0.2894
R to W	83.5	0.3022
N to T	13.5	0.3079
F to L	65	0.3086
S to R	35	0.3108
D to G	42.5	0.3117
E to A	16	0.3242
A to S	29.5	0.3252
P to T	24	0.334
T to P	37	0.3349
R to H	98	0.3417
F to Y	8	0.343
T to I	69.5	0.3448
T to K	11	0.3467
P to S	52	0.3507
S to P	58	0.3543
G to E	64	0.3612
K to T	11	0.3659
G to A	31.5	0.3756
T to A	27	0.3758
N to D	21	0.3839
N to H	12.5	0.3848
M to I	27	0.3859
V to L	35	0.3973
F to I	14.5	0.4055
R to L	31.5	0.4134
Y to H	40	0.4136
L to R	50.5	0.4168

Y to C	84.5	0.4275
S to I	10.5	0.4344
A to P	56	0.4377
F to S	42	0.4416
Q to P	35	0.4501
F to V	21	0.453
Y to S	13	0.4617
G to C	26.5	0.4617
D to A	13.5	0.4651
N to S	51	0.4983
L to W	6	0.5
R to Q	103	0.5187
C to Y	42.5	0.5274
R to C	120	0.5275
W to L	6.5	0.5385
H to L	9	0.5472
C to S	10	0.5521
E to V	9	0.5817
T to M	38.5	0.6001
G to S	81.5	0.6317
A to V	99.5	0.6384
Q to L	8.5	0.6544
E to Q	19	0.6598
E to G	35.5	0.6604
L to V	39.5	0.6664
V to D	18.5	0.6688
M to K	16.5	0.6694
I to M	21	0.6734
P to A	12	0.6739
L to M	8	0.6832
I to L	5.5	0.6848
A to G	15	0.698
H to Y	31	0.7119
H to R	44	0.7464
W to G	10.5	0.7619
K to E	49	0.7684
S to L	41.5	0.7699
Y to F	4.5	0.8189
L to F	56	0.8367
H to P	18.5	0.8695
V to I	33	0.8738
I to R	1.5	0.9214
I to F	17	0.9378

I to V	18	0.9381
S to G	10	0.9498
Q to E	15	0.9531
V to Q	1	1

A.2 REFERENCES

Kellogg EH, Leaver-Fay A, Baker D. 2011. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**(3): 830-838.