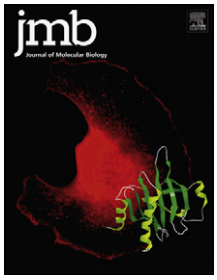

Predicting the Effect of Amino Acid Single-Point Mutations on Protein Stability—Large-Scale Validation of MD-Based Relative Free Energy Calculations

Thomas Steinbrecher[1], Chongkai Zhu[2], Lingle Wang[2], Robert Abel[2], Christopher Negron[2], David Pearlman[2], Eric Feyfant[2], Jianxin Duan[1] and Woody Sherman[2]

*1 - **Schrödinger GmbH,** Dynamostr. 13, 68165 Mannheim, Germany*
*2 - **Schrödinger Inc.,** 120 West 45th Street, 17th Floor, New York, NY 10036, USA*

***Correspondence to Thomas Steinbrecher:** thomas.steinbrecher@schrodinger.com*
http://dx.doi.org/10.1016/j.jmb.2016.12.007
Edited by Anna Panchenko

Abstract

The stability of folded proteins is critical to their biological function and for the efficacy of protein therapeutics. Predicting the energetic effects of protein mutations can improve our fundamental understanding of structural biology, the molecular basis of diseases, and possible routes to addressing those diseases with biological drugs. Identifying the effect of single amino acid point mutations on the thermodynamic equilibrium between the folded and unfolded states of a protein can pinpoint residues of critical importance that should be avoided in the process of improving other properties (affinity, solubility, viscosity, etc.) and suggest changes at other positions for increasing stability in protein engineering. Multiple computational tools have been developed for *in silico* predictions of protein stability in recent years, ranging from sequence-based empirical approaches to rigorous physics-based free energy methods. In this work, we show that FEP+, which is a free energy perturbation method based on all-atom molecular dynamics simulations, can provide accurate thermal stability predictions for a wide range of biologically relevant systems. Significantly, the FEP+ approach, while originally developed for relative binding free energies of small molecules to proteins and not specifically fitted for protein stability calculations, performs well compared to other methods that were fitted specifically to predict protein stability. Here, we present the broadest validation of a rigorous free energy-based approach applied to protein stability reported to date: 700+ single-point mutations spanning 10 different protein targets. Across the entire data set, we correctly classify the mutations as stabilizing or destabilizing in 84% of the cases, and obtain statistically significant predictions as compared with experiment [average error of ~1.6 kcal/mol and coefficient of determination (R^2) of 0.40]. This study demonstrates, for the first time in a large-scale validation, that rigorous free energy calculations can be used to predict changes in protein stability from point mutations without parameterization or system-specific customization, although further improvements should be possible with additional sampling and a better representation of the unfolded state of the protein. Here, we describe the FEP+ method as applied to protein stability calculations, summarize the large-scale retrospective validation results, and discuss limitations of the method, along with future directions for further improvements.

Introduction

Protein structure and energetics are at the heart of structural biology, dictating a vast array of biochemical functions, from enzymatic reaction rates to cellular signalling. A vast amount of data has been assembled on the relation between sequence mutations to human diseases [1–14], effectively connecting macroscopic phenotypic changes to details of atomic structure (see, e.g., the *Journal of Molecular Biology* special issue on "Understanding molecular effects of naturally occurring genetic differences" [15]). Protein structure determination, primarily by X-ray crystallography and NMR, sometimes supplemented by homology modeling, has been successful in providing atomic resolution 3D

structures, with tens of thousands of experimental molecular structures publicly available. However, our understanding of protein energetics—such as the thermodynamic properties of processes ranging from protein binding, folding/unfolding, or aggregation—is much more limited.

In this work, we focus on the energetics of protein stability, specifically on protein stability changes caused by single amino acid point mutations (SPMs). All results presented in the following refer to relative free energies of stability (i.e., we compute the effect of individual amino acid residue mutations on the stability of a protein); no absolute stability predictions are made. It is well established that protein sequences evolve to conserve stable functional folds [16] for a majority of proteins. While cases of functional disordered protein states are well known [17–20], a specific folded 3D structure is one of the defining characteristics distinguishing biologically functional proteins from most polymers. Despite this importance of folded structures, it has been estimated that the folded state of most proteins is only marginally more stable than the unfolded state [21–23]. The majority of polypeptide sequences do not correspond to stable protein folds [24,25] and individual amino acid point mutations can have large destabilizing effects [26,27]. The enormous size of the available protein sequence space affords nearly limitless mutation options, of which any particular set of mutations could stabilize or destabilize a given protein fold. In addition, it is well established that for practical purposes of protein functionality, kinetic stability is of great importance [28], but the majority of previous studies have emphasized thermodynamic stability, which will also be the focus of this work.

The ability to reliably predict protein stability changes would be extraordinarily valuable in understanding basic protein structure and function and in connecting evolutionary constraints in protein sequences to structural properties. As changes in protein stability are often connected to disease-causing mutations [29–33], such predictions can provide a direct link between molecular structure and human disease. Protein stability predictions would also provide value in the field of protein design and engineering [34,35]. In the latter case, the goal may be either directly to increase the thermostability of a protein to facilitate its production and handling, or may be concerned with maintaining protein stability while mutations are introduced to optimize other properties such as affinity, solubility, aggregation, or viscosity.

Unsurprisingly, a large number of tools to predict protein stability effects have been developed [35–55]. The majority of these methods use empirically fitted parameters to optimize predictions for a training set of known protein mutations. However, even with substantial parameterization, typically only moderate success is achieved in prospectively predicting protein stabilities due to the great complexities associated with the protein sequence–structure–energy relationship. Many factors, such as the detailed microenvironment of the mutations, hydration structure, unfolded state, and protein flexibility, are not accounted for in sequence-based models. Additionally, an overreliance on parameters can make predictions less transferable to systems beyond the training set and the parameterized models are often hard to interpret.

Ideally, protein stability predictions should aim to accurately represent the underlying protein folding energetics by incorporating physics-based methods [56] that explicitly consider the important factors affecting protein stability (e.g., protein flexibility, the unfolded state, solvation, etc.). If the molecular conformational space and potential energy function are modeled accurately enough, free energy changes can be directly calculated by principles from statistical thermodynamics. Such an approach, based on molecular dynamics-based free energy perturbation (MD-FEP) calculations has recently been successfully employed in the calculation of relative protein–ligand binding free energies [57–61]. In principle, the concepts of FEP calculations can be employed to compute a variety of free energy properties where the two relevant endpoints can be modeled (e.g., bound *versus* unbound for binding affinity, folded *versus* unfolded for protein stability, or solid *versus* aqueous for solubility); here we report the large-scale application of FEP+ to computing protein stability free energies.

Results

Data set

Multiple data sets to test protein stability predictions are available in the literature [35–37,62] and several databases specifically collect amino acid point mutation data [63–67]. To facilitate a comparison of our method with established tools, we have taken data from a previously published validation data set originally compiled to test the stability prediction tool FOLD-X [36]. The 964 mutations in the entire set available in the Supporting Information of Ref. [36] span 37 different proteins. Due to the large amount of manual effort required to carefully prepare each system for MD simulations, and because predictions for systems with relatively few total data points (e.g., <10) are often inconclusive even in good cases, we have selected the 10 protein systems with the largest numbers of point mutations in their respective sets, comprising 741 total data points (see Table 1), as our data set to test FEP+ protein stability predictions. Due to implementation limitations at the time this work was conducted,