BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

Toward Fast Determination of Protein Stability Maps: Experimental and Theoretical Analysis of Mutants of a *Nocardiopsis prasina* Serine Protease

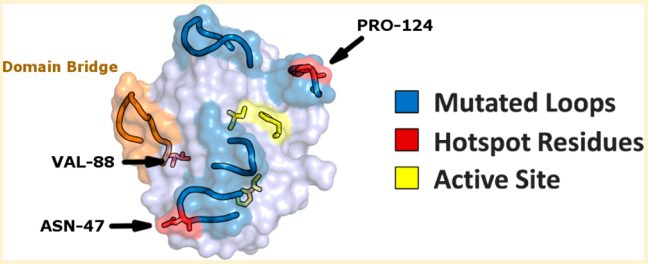
Damien Farrell,[†] Helen Webb,[†,‡,§] Michael A. Johnston,[†] Thomas A. Poulsen,[‡] Fergal O'Meara,[†] Lars L. H. Christensen,[‡] Lars Beier,[‡] Torben V. Borchert,[‡] and Jens Erik Nielsen*[,†]

[†]School of Biomolecular and Biomedical Science, Centre for Synthesis and Chemical Biology, UCD Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland

[‡]Protein Optimization, Novozymes A/S, Brudelysvej 26, 2880 Bagsværd, Denmark

Supporting Information

ABSTRACT: The stability of serine proteases is of major importance for their application in industrial processes. Here we study the determinants of the stability of a *Nocardiopsis prasina* serine protease using fast residual activity assays, a feature classification algorithm, and structure-based energy calculation algorithms for 121 micropurified mutant enzyme clones containing multiple point mutations. Using a multivariate regression analysis, we deconvolute the data for the mutant clones and find that mutations of residues Asn47 and Pro124 are deleterious to the stability of the enzyme. Both of these residues are situated in loops that are known to be important for the stability of the highly homologous α-lytic protease. Structure-based energy calculations with PEATSA give a good general agreement with the trend of experimentally measured values but also identify a number of clones that the algorithm fails to predict correctly. We discuss the significance of the results in relation to the structure and function of closely related proteases, comment on the optimal experimental design when performing high-throughput experiments for characterizing the determinants of protein stability, and discuss the performance of structure-based energy calculations with complex data sets such as the one presented here.



The engineering of protein stability continues to be of major importance for the use of enzymes in a wide range of industrial applications,[1] including starch liquefaction,[2] biomass conversion,[3] bread baking,[4] and laundering.[5] While efficient strategies for engineering protein stability have been established over the past 30 years, it is nevertheless still challenging to increase the stability of a very stable protein. In particular, it is often difficult to increase the stability of an enzyme while maintaining or increasing the catalytic activity of the enzyme,[6] although notable successes have been produced in this area using rational protein engineering techniques.[7,8] In the past two decades, enzyme engineering and optimization studies have increasingly been performed using directed evolution strategies.[9,10] With such strategies, stabilized mutant enzymes are identified using screening or selection-based assays, and enzymes with impressive increases in both stability and activity have been developed in this way.[9]

In contrast to the well-behaved model proteins frequently studied in academic laboratories, industrially relevant enzymes are often much larger, cofactor-dependent, highly stable proteins that unfold irreversibly and are kinetically stable,[11] with the production of aggregates or partially degraded species in the case of proteases. Irreversibly unfolding proteins typically unfold via a three-step [native (N) $\leftrightarrows$ unfolded (U) $\rightarrow$ final (F)] process, where the rate-limiting step can be either the initial (often partial) unfolding (N $\leftrightarrows$ U) or the final irreversible step (N $\rightarrow$ F). Indeed, for a highly stable thermolysin-like protease, it has been found that the partial unfolding of a single loop constitutes the main determinant of stability, and that stabilization of other regions in the protein has a negligible effect.[7] This presents a problem when engineering the stability of kinetically stable proteins because one has to specifically stabilize the region that is rate-determining for the deactivation of the enzyme. In this paper, we investigate the use of fast stability measurements for determining protein stability hot spots for an industrially relevant, kinetically stable, irreversibly unfolding, 188-residue serine protease from *Nocardiopsis prasina* [*N. prasina* serine protease SP1 (NCPP)].[12]

The NCPP is a chymotrypsin-like (trypsin-like type S1) serine protease, whose structure adopts the typical two-domain β-barrel fold.[13] The active site is situated in a shallow cleft on the surface of the enzyme with the catalytic triad being formed by residues Asp 61, His 35, and Ser 143. The melting

Received: December 27, 2011
Revised: June 2, 2012
Published: June 6, 2012

temperature of the wild-type enzyme is 73 °C at pH 7, as measured by circular dichroism (CD)−temperature denaturation scans (see Materials and Methods), and NCPP is thus a highly thermostable enzyme.

We use this protease as a model system because an extensive data set was available to us from a previous protein optimization study, and because there is significant interest in understanding the determinants of protease stability in general.

NCPP and its close homologues, such as *Nocardiopsis alba* protease (NAPase) and α-lytic protease (αLP), have evolved to be active under harsh, proteolytic conditions. The proposed stability model for these proteases[13,14] states that the unfolding process is kinetically controlled. Even when the native state is thermodynamically favored, irreversible processes can lead the protein into a terminal unfolded state. This scenario is described by the Lumry−Eyring model.[15] To prevent irreversible unfolding, the enzyme employs a large kinetic barrier separating its unfolded (U) and final (F) states or native (N) and unfolded (U) states. In some cases, the native state is not energetically favored and folding requires a pro-region (a distinct polypeptide chain synthesized along with the protein precursor). When the pro-region is cleaved off the polypeptide, the protein is trapped in the native state because of the large energy barrier for the unfolding reaction. Thus, the folding−unfolding reactions of the kinetically stable enzyme are no longer in equilibrium, and classic thermodynamic relations for protein stability do not apply.[11,16]

The acid stability of NAPase[17] has been discussed in detail[13] as being partly due to the intradomain-only location of salt bridges providing acid kinetic stability. The high structural similarity with NAPase means that the arguments put forth for kinetic stability are likely to apply to NCPP, and that the catalytically active state of NCPP therefore is a kinetically stabilized state that is not in equilibrium with its unfolded state.

We employ fast residual activity assays on 121 micropurified clones at pH 7 for measuring the effect of mutations on the residual activity of NCPP. We analyze the resulting ΔT_{50} values using a simple feature selection method to identify residues in the data set that are of particular importance for the stability of the enzyme. We continue to compare the measured residual activity ΔT_{50} values of 72 clones to the predictions produced by a typical structure-based protein stability prediction method. Finally, we discuss the implications of our results for the construction of algorithms for predicting protein stability and suggest experimental and theoretical procedures that should be used for determining stability hot spot maps of industrial enzymes.

MATERIALS AND METHODS

Directed Evolution Strategy. Mutant clones were constructed using error prone polymerase chain reaction and a library-based combination of mutations designed using visual structural analysis with randomly chosen mutations in loops 2 and 3. The libraries were constructed using standard molecular biology techniques or purchased from Morphosys AG as a Slonomics library and transformed into *Bacillus subtilis* using standard procedures. Following plating and colony picking, all 5500 picked clones were sequenced to uniquely identify the mutations present in each clone.

Fermentation. Clones containing mutant enzymes were stored in 96-well plates containing 200 μL of TB glycerol and 100 μM CAM. Each variant was stored in triplicate in the same well position on three different 96-well plates to ensure three biological replicates. Each plate contained wild-type (WT) samples and controls for a comparison of the assays. Plates were covered with porous covers and incubated at 30 °C for 96 h with shaking at 250 rpm. Following incubation, plates were stored at −80 °C.

Selection of the 121 Variants. We measured the catalytic activity of each clone in triplicate using the casein-based EnzChek protease assay kits at pH 8.0 and 5.6. We used a final concentration of the casein substrate in the activity assay of 4.9 μg/mL and measured the increase in fluorescence at 590 nm. We selected 121 variants from a pool of 5500 variants using two selection strategies as outlined below.

Selection Method 1: pH−Activity Ratio. Two-thirds of the 121 variants were selected on the basis of an $(act_{pH\ 5.6})/(act_{pH\ 8.0})$ ratio (r) calculated from the two EnzChek activity measurements described above; we chose the 40 variants with the highest value of r and the 40 variants with the lowest value of r. In all cases, we required the relative activity of a clone to be at least 20% of that of the WT for it to be selected.

Selection Method 2: Calculated ΔpK_a Values. We chose a further 41 variants on the basis of ΔpK_a calculations using PEATSA.[18,19] We performed the calculations using a protein dielectric constant of 8 and an ionic strength of 0.144 M and selected the 41 variants that displayed an absolute calculated ΔpK_a of His 35 of >0.4 unit. We preferentially selected clones containing single-point mutations that had not been selected using the first selection strategy and furthermore ensured that we sampled as many different mutations as possible.

The selection of variants using both methods was completely nonbiased in terms of stability.

Fermentation of Selected Clones. The 121 selected variants were refermented in 10 mL well plates containing 4 mL of TB glycerol and 100 μM CAM. Each variant was fermented in four separate wells to ensure a sufficient yield. Plates were covered with porous covers and incubated at 30 °C for 96 h while being shaken at 250 rpm. Plates were removed and the four wells pooled. Pooled fermentations were centrifuged at 4500 rpm for 15 min, and the supernatant was filtered using 0.2 μm filters and the cell pellet discarded.

Micropurification. MEP HyperCel chromatographic medium (100 μL) was suspended in 20% ethanol and added to each well of a Whatman Unifilter plate. The liquid was removed by vacuum. The MEP HyperCel was washed using a 100 mM Tris/10 mM $CaCl_2$/1 mM Triton mixture (pH 8.5). The protease supernatant (2 × 400 μL) was bound to the MEP HyperCel in two steps using 0.5 M Tris (pH 8.5). The plate was left to incubate at room temperature for 1 h while being vigorously shaken at 1200 rpm. Following incubation, the MEP HyperCel chromatographic medium was washed several times using Tris (pH 8.5), and the unbound material was removed by vacuum. The micropurified protease was eluted from the medium using 50 mM sodium acetate (pH 4.3).

Residual Activity. Variants were incubated at pH 7.0 for 15 min at five temperatures (room temperature and 60, 65, 70, and 75 °C). Following incubation, suc-AAPF-pNA was diluted to 0.48 mM in a 100 mM Tris/10 mM $CaCl_2$/1 mM Triton mixture (pH 8); 180 μL of the substrate dilution was added to 20 μL of the incubated variants, and the residual activity was immediately read every 10 s for 3 min at 405 nm. The final concentration of suc-AAPF-pNA in the activity assay was 0.43 mM.

CD Spectroscopy. WT NCPP was dissolved in 2.5 mM potassium phosphate (pH 7.0) to a final concentration of 0.1

mg/mL. The conditions were as follows: temperature range of 25−90 °C, wavelength range of 180−280 nm, scan rate of 0.25 s/nm, and wait time of 0.3 s/°C. The protease sample was then cooled to 25 °C for a final wavelength scan.

Differential Scanning Calorimetry. WT NCPP was dissolved in a 100 mM Tris/10 mM $CaCl_2$/1 mM Triton mixture (pH 7.0) to a final concentration of 1 mg/mL. The sample was degassed for approximately 40 min. The sample temperature was increased from 20 to 90 °C at a rate of 1 °C/min. The protease sample was then cooled to 20 °C, and the temperature scan was repeated.

■ ANALYSIS

Initial Analysis of Thermal Inactivation Data. The loss of enzyme activity normalized to the room-temperature (25 °C) value provides the residual activity as a function of temperature. These data were fit to the following general sigmoid:

$$\text{activity}_{\text{obs}} = \frac{1}{1 + \left(\frac{T}{T_{50}}\right)^{m}} \quad (1)$$

where m is the slope, T is the temperature, and T_{50} is the midpoint of thermal inactivation, the temperature at which the activity is 50% of the activity at room temperature. Note that this equation merely provides a reasonable empirical fit to the normalized data for the small number of data points and has no physical basis. Changes in stability are estimated by subtracting the clone T_{50} from the wild-type T_{50}. All curves, in duplicate, at all pH values were automatically fit using a specially written software pipeline in PEATDB.[20] Filtering of the clones according to fit quality was then performed to extract the most reliable T_{50} values. Figure S5 of the Supporting Information shows a selection of curve fits for some of the variants.

Obtaining ΔE_a Values from Residual Activity Data. The deactivation of NCPP is irreversible as shown in Figure S1 of the Supporting Information, and the deactivation process can therefore be considered a first-order reaction described by

$$[\mathrm{N}] = [\mathrm{N}]_0 e^{-kt} \quad (2)$$

where $[\mathrm{N}]$ is the concentration of the native state at time t, $[\mathrm{N}]_0$ is the concentration of the native state at time zero, and t is the incubation time for the assay. We now assume that k follows classic Arrhenius behavior and eq 2 can thus be rewritten as

$$[\mathrm{N}] = [\mathrm{N}]_0 e^{-Ae^{-E_a/RT}t} \quad (3)$$

where E_a is the activation energy, A is the Arrhenius pre-exponential factor, T is the temperature, and R is the gas constant. Residual activity curves can readily be fit to eq 3, with A and E_a being the only two unknown variables. We determined A and E_a from three separate WT residual activity curves and used the WT value of A in fits of all mutant residual activity curves. We thus assume that the chemical nature of the transition state for the process remains relatively unperturbed by the mutations and that only the height of the transition state barrier changes.

The error in the E_a values was determined by fitting to data for both replicates and finding the standard error of the mean value. The experimental error was deemed well below the error in the calculated stability values.

In the following, we assume that the ΔE_a values ($E_{a,\text{mutant}} - E_{a,\text{WT}}$) originate from energy changes in the ground state for the rate-limiting step and that the energy of the transition state thus does not change.

Structure-Based Stability Predictions. We compare the ΔE_a values determined by eq 3 to calculated stability values from PEATSA (Protein Engineering Analysis Tool-Structural Analysis).[18] PEATSA uses the mutant-induced changes in the protein structure to determine the differences between the reference and mutant for each energy term. These are summed to find a free energy difference for both folded states. The mutant structure is modeled without backbone relaxation using a rotamer-based optimization as detailed previously.[18]

It is not immediately obvious that we can hope to achieve a good correlation between $\Delta\Delta G_{\text{fold}}$ values calculated with PEATSA and the ΔE_a values derived from the analysis described above. However, if the terminal inactivation is limited by the rate of unfolding, then $\Delta\Delta G_{\text{fold}}$ and ΔE_a will be correlated as discussed in the following.

A mutation can change the activation energy for the unfolding reaction by perturbing the energy difference between the transition state (TS) and the native (N) state. If there is a low degree of structural similarity between the N state and the TS, a mutation is most likely to do so by destabilizing the native state without changing the energy of the transition state. If the N state and the TS are very similar, a mutation is likely to affect both states equally, and it is also possible to have mutations affect only the TS. Mutations that selectively destabilize the folded state will change ΔG_{fold} and E_a by the same amount, thus giving an overall 1:1 correlation between ΔE_a and $\Delta\Delta G_{\text{fold}}$.

In the following, we assume that mutations change only the energy of the folded state and not the energy of the transition state and the unfolded state. In such a scenario, the change in the activation barrier for unfolding will be equivalent to the change in ΔG_{fold}. Only if mutations target regions that play a role in transition state stabilization or in stabilization of the unfolded state will this correlation break down.

■ RESULTS

In the following, we analyze the residual activity T_{50} data from the suc-AAPF-pNA experiment for NCPP to identify residues that are important for the stability of the enzyme. Our secondary object is to evaluate how well a classic structure-based stability prediction algorithm (PEATSA) performs on a data set from a kinetically stable enzyme such as NCPP. We begin our analysis by examining the data acquired for the unfolding of wt NCPP to establish the characteristics of the inactivation process. We then describe the effect of the NCPP mutations and use a feature classification algorithm to identify residues that are particularly important for the stability of NCPP. We conclude our analysis by examining the ability of PEATSA to predict the change in T_{50} for the mutant NCPP clones examined.

Stability of wt NCPP. DSC measurements and CD thermal denaturation scans demonstrated that NCPP unfolds irreversibly at pH 7.0 (see Figures S1 and S2 of the Supporting Information). The wild-type transition midpoint temperature, T_{50}, was found to be 69 °C when measured using the thermal inactivation assay, 80 °C using DSC, and 73 °C using CD. The difference in these numbers presumably reflects the effect that the experimental conditions (heating rate and buffer composition) have on the unfolding process of NCPP. We found an irreversible model to be the best fit to the DSC

dx.doi.org/10.1021/bi201926f | Biochemistry 2012, 51, 5339−5347

melting scans.[21] This particular model incorporates terms for both calorimetric and van't Hoff heat changes and yields a ΔH_{cal} of 953 kJ/mol.

NCPP and NAPase are highly similar enzymes (87% identical sequences, C_α root-mean-square deviation (rmsd) from X-ray structure of 0.31 Å), and the unfolding mechanism of NCPP is thus likely to be identical to the kinetically controlled unfolding of NAPase.

Stability Changes of Mutant NCPP Clones. We selected 121 mutant clones of NCPP from an initial pool of 5500 mutant clones produced as part of a directed evolution study (see Materials and Methods). After regrowing and micropurification, 72 clones yielded experimental stability data of sufficient quality for the data analysis. Denaturation curves for 65 of the clones were of sufficient quality to obtain accurate ΔT_{50} values (estimated standard error of 1.5 °C), whereas the seven remaining curves yielded only a lower limit for T_{50}. The mutations are all neutral or destabilizing with ΔT_{50} values of less than −12 °C (below which it was determined that an accurate value cannot be measured, as illustrated in Figure S6 of the Supporting Information). The most destabilized clones are listed in Table 1.

Table 1. Ten Most Destabilizing Variants in the Data Set[a]

mutations	$\Delta T_{50(\text{mutant}-\text{wild type})}$ (°C)
A12GH/A38RA/A47NL/A165RH	−19.0
A12GH/A22TV/A38RT/A47NA/A54QM/A165RH	−16.5
A22TV/A23ND/A38RQ/A47NR/A54QT/A165RH	−13.8
A17VA/A42QR	−13.3
A88VA/A122SR/A125ES	−13.2
A38RH/A47NI/A165RH	−12.8
A123YR/A124PR/A125ET	−12.5
A122SN/A123YT/A125EH	−11.1
A38RH/A47NQ/A54QE/A165RH	−10.1
A47NH/A48GD/A165RH	−10.0

[a]Note that the reliable detection limit of ΔT_{50} is approximately −12 °C. Therefore, a lower limit of −12 °C was applied to the first seven mutants in the calculation of weights to avoid biasing the results. See Figure S6 of the Supporting Information and section IV of the Supporting Information for further discussion.

The average number of mutations per clone is 2.8 with 67 of 72 clones (93%) containing multiple mutations. The data set contains a total of 100 unique single mutations amounting to a coverage of 2.7% compared to a full single-mutant saturation mutagenesis experiment. In total, 48 of 188 residue positions (26%) have been mutated, with the distribution of mutations skewed toward mutating loops 1, 3, and 4 (Figures 1 and 2).

Pinpointing Stability Hot Spots Using Mutational Data. The NCPP data set consists primarily of clones with multiple point mutations, and it is therefore not immediately possible to pinpoint mutations that cause large changes in T_{50}. To identify the effect of mutating specific residues, we apply a multivariate fitting analysis. All 48 residues mutated in the data set are treated as dichotomous (having two possible values, present or not) variables. For each sample (representing a mutation and corresponding stability), the variable is set to 1 or 0 depending on whether the residue is mutated or not mutated in the clone. Thus, each sample consists of a set of features (residues) associated with a response variable (stability). When

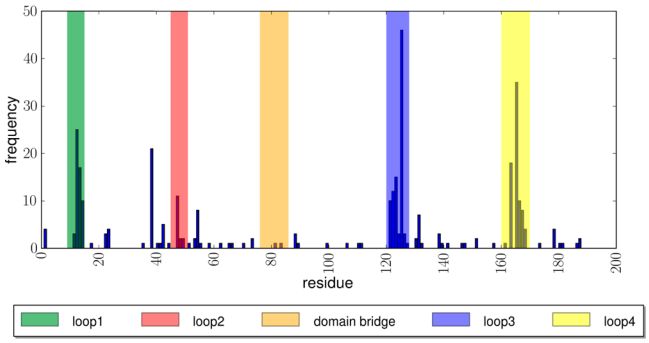


Figure 1. Frequency of each residue (the number of times it is present in all mutants) for the 121 clones analyzed in this paper. Mutations in loops 1, 3, and 4 are heavily over-represented in the data set. In total, 48 of 188 (26%) residues in the protein are mutated, with the average number of mutations per clone being 2.8.

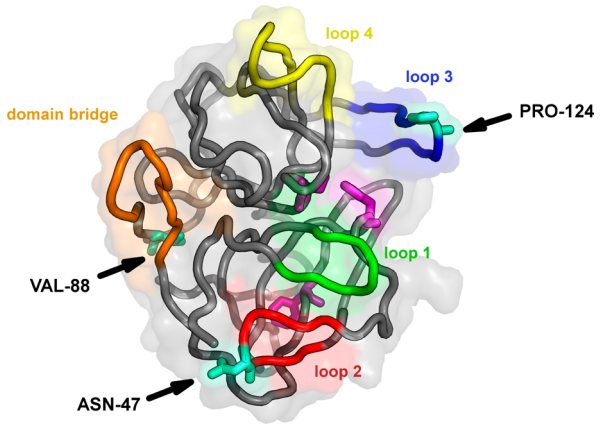


Figure 2. Location of loop regions and the "domain bridge" mapped on the structure, color-coded according to Figure 1. Active site residues are colored purple.

the data are cast in this form, they can be fit using a linear model with multiple variables:

$$\Delta T_{50} = \alpha_0 + \alpha_1 r_1 + \alpha_2 r_2 + \ldots + \alpha_n r_n \quad (4)$$

where r_n is 1 or 0 based on its presence in the mutant and α_n is the coefficient or weight for this feature. The fitted weights correspond to the importance of each residue for stability. Obtaining accurate weights is essentially a feature selection problem,[22] which is often encountered in supervised machine learning algorithms. There are various approaches to feature selection, and in this case, we applied a linear regression algorithm to fit the data and to find the weights for each residue (i.e., each feature). We used the least angle regression (LARS)[23] algorithm as implemented using the pyMVPA toolkit[24] because this algorithm is optimized to give robust variable weights for a low ratio of the number observations to the number of variables. See section III of the Supporting Information for a more detailed explanation of the method.

In applying eq 4, we make the assumption that any mutation of a specific residue has the same effect on the stability of NCPP. This is not the case as demonstrated by many studies of protein stability, but this assumption provides a useful first approximation that allows us to identify positions that strongly influence the stability of NCPP.

To estimate the general reliability of our algorithm in connection with eq 4 for selecting important features, we tested

dx.doi.org/10.1021/bi201926f | Biochemistry 2012, 51, 5339−5347

the procedure on a simulated data set. The test results show that the fitting can reproduce a set of weighted features, and we can furthermore use the feature weights to reconstruct stabilities for unknown mutants. Conclusions from this simulated testing are given in Discussion and detailed in section III of the Supporting Information.

The algorithm identified 30 residues with non-zero weights; three residues stood out as having very high negative values and hence were found to have a large influence on the stability of NCPP.

Table 2 lists the weights for the 10 most negatively weighted residues, and Figure 3 shows that all ΔT_{50} values in the data set

Table 2. Residue Weights Derived from the Feature Classification Algorithm for the Three Most Negative Weights[a]

residue	weight (pH 7)	error[b]	coverage[c]	total residue frequency	conservation score	loop
Pro 124	−8.1	1.3	1	3	1.5	3
Val 88	−5.2	0.6	1	2	−0.68	db
Asn 47	−7.4	0.37	7	9	−0.38	2

[a]Of a total of 30 features, 20 of which occur more than once in the mutations. The frequency column shows how many times the residue appears in all mutations. A mean weight, including other available pH data, is also given. Also given are the conservation scores found using the ConSurf server.[34] [b]Error derived from the cross validation test. [c]Coverage refers to the number of possible substitutions sampled for that residue.

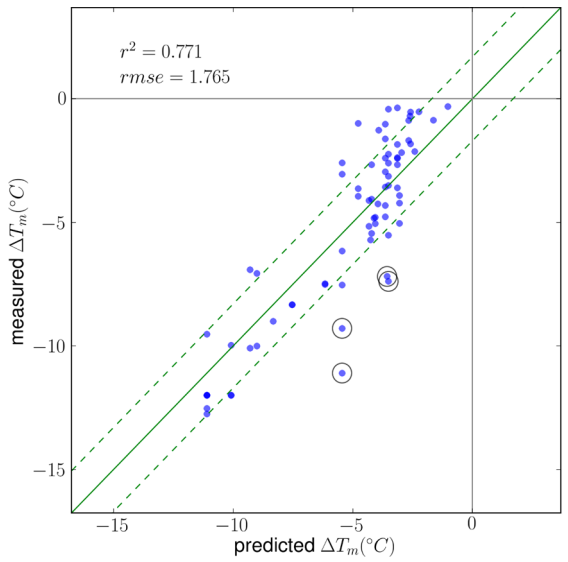


Figure 3. Stabilities reconstructed using the classifier algorithm plotted vs the experimental values. In cases of reasonable agreement, mutations either are having an additive effect or are dominated by one mutation (with the largest weight). Note that if the weights exactly reflected each (linear) contribution to stability, the correlation would be 1:1, within the baseline error. Where there is poor agreement, this may indicate that the mutations have a cooperative effect on the stability that is not considered by the linear model. The dashed line shows the rmsd of 1.7. Outliers in circles are those for which the error > 2 × rmsd.

can be reconstructed using eq 4 with an rmsd of 1.7 °C (correlation coefficient of 0.87) using only the 30 weights determined above. Figure 3 contains four mutant enzymes that have a residual (observed T_{50} − reconstructed T_{50}) that is >2 times greater than the rmsd. These four outliers do not represent a consistent pattern of mutations.

Note that 22 of the original 48 mutated positions appear only once in the data set and contribute much less than the remaining residues. Removal of the 22 residues followed by linear regression to optimize the weights of the 26 remaining residues yields an rmsd of 2.3 °C and a correlation coefficient of 0.76.

The feature classification method is thus able to pinpoint specific residues that are important for the stability of NCPP in the current data set. However, the particular variables selected and the weights to which they are assigned are dependent on the data set used. To test the reliability of the identified weights, we performed a cross validation by random subsampling. Specifically, we repeated the fitting process 50 times with a random 20% of the original mutants removed per iteration. This analysis yields a measure of the variability in the residue position weights for our particular data set. Features that show high variability in detection and weight are less reliable, and features that show little variability are robust descriptions of the characteristics of the data set. The errors for the features in Table 2 are calculated from the standard error of the mean weight from the 50 runs and show that the descriptors for positions 47, 88, and 124 are robust and that these residues therefore can be trusted to be of major importance for the stability of NCPP.

Hot Spot Residues. The residues identified from the analysis described above are Asn 47, Val 88, and Pro 124, which all are located in the surface loops shown in Figure 2.

Asn 47 is situated at the bend of the hairpin of loop 2 and makes hydrogen bonds to the side chain of Asn 70 and to the N-terminus of the protein (Ala 1). Loop 2 is involved in the enhanced stability of αLP and *Thermobifida fusca* protease A (TFPA), an extreme thermophile.[17] The loop consists of three residues in TFPA and contains a very tight turn facilitated by a *cis*-proline, which limits the conformational flexibility of the loop and is thought to be a major determinant of the increased thermostability of TFPA compared to that of αLP. In NCPP and αLP, the loop consists of three additional residues (Gly 46-Asn 47-Gly 48), and the loop residues display significantly above-average *B* factors in the X-ray structure of αLP (Protein Data Bank entry 1SSX[25]) and slightly higher *B* factors in the X-ray structure of NCPP (K. Wilson, unpublished data).

Val 88 is located close to the domain bridge loop,[17] which very likely plays a role in kinetic stability. However, the residue is only present in two mutants and always in combination with two other residues (Ser 122 and Glu 125). We cannot therefore rule out the possibility that the weight attributed to Val 88 is due to mutations of Ser 122 and Glu 125, further strengthening the notion that loop 3 plays an important role in determining the stability of NCPP. This illustrates a limitation with the use of automated methods for analyzing mutational data, the inability of the algorithms to consider structural arguments when identifying hot spots.

The last residue, Pro 124, is in loop 3, a β-hairpin loop in the C-terminal domain that is known in closely related proteases to be important in pro-region binding.[26] However, there is only one sample of this residue present in several of the most destabilized mutants (see Table 1). In contrast, other residues in this loop (Ser 122 and Glu 125) are sampled better but do not appear with significant weights in our analysis. Pro 124 is

dx.doi.org/10.1021/bi201926f | *Biochemistry* 2012, 51, 5339−5347

therefore likely to stabilize NCPP by reducing the entropy loss as loop 3 adopts its native conformation.

Despite extensive sampling, we do not find destabilizing residues in loops 1 and 4 and therefore conclude that these loops are not important for the stability of NCPP.

Deducing the Effect of Individual Positions and Mutations Using a Reductive Algorithm. The regression algorithm yields fitted weights for each position in the protein, but these weights simply represent the best fit to the experimental data and can incorrectly identify residues as shown above. To further validate the weights found by the LARS fit, we analyzed the NCPP mutant clones using a simple reductive algorithm. The NCPP data set contains seven single-point mutations that yield information about the ΔT_{50} values of six positions. (The single-site mutations are Pro 58 → Asp, Glu 125 → Arg, Gln 173 → His, His 35 → Asp, Gly 12 → Glu, Gly 167 → Asp, and Glu 125 → Met.) Using these six ΔT_{50} values, we can now deduce the ΔT_{50} for a further 26 positions by using the argument that if mutation of position A gives a ΔT_{50} of −5 °C and mutation of positions A and B gives a ΔT_{50} of −7 °C, then mutation of position B must give a ΔT_{50} of −12 °C. This argument breaks down if the mutated positions interact in the wild-type or mutant proteins and depends on the accuracy of the ΔT_{50} measurements, but the procedure is analogous to the classic deductive reasoning used when analyzing sets of point mutations in proteins. If we plot the ΔT_{50} values from this analysis against the weights from our regression algorithm, we obtain a very good correlation [correlation coefficient of 0.78 (Figure S13 of the Supporting Information)] with the most destabilizing position identified as Asn 47. Position 124 cannot be isolated using the deductive algorithm as it occurs in the 15 mutant proteins that cannot be successfully decomposed using the deductive analysis. Overall, this analysis lends strong support to the use of our regression technique in the analysis of protein stability data sets.

Predicting ΔE_a Values Using a Structure-Based Thermodynamic Energy Function. PEATSA is a structure-based energy function optimized to predict mutant-induced $\Delta\Delta G_{fold}$ values for two-state folding proteins. PEATSA yields a standard error of 4−6 kJ/mol (correlation coefficient of 0.52−0.76) for currently available protein stability data sets.[18] While the kinetically determined ΔE_a values for NCPP are not strictly comparable to the $\Delta\Delta G_{fold}$ values calculated by PEATSA, it is nevertheless interesting to examine the ability of PEATSA to reproduce such values. Figure 4 shows the correlation for predicted and fitted experimental ΔE_a values and illustrates that although PEATSA is able to provide accurate predictions (within error) for a number of clones, the ΔE_a values are generally underestimated in comparison to the predicted value.

When Is PEATSA Wrong? The prediction error is defined as predicted $\Delta\Delta G$ − experimentally fit ΔE_a; there are two kinds of outlier clones in Figure 4 (those marked with empty circles). (1) Those predicted to be destabilizing but are actually not will have a large positive error, and (2) those that experimentally are destabilized but not predicted to be, will have a large negative error. In either case, this means that the mutation either cannot be described as a pure thermodynamic effect or reflects the limitations of the mutational model. The majority of the outliers contain mutations in loops 1 (residues 11−13) and 2 (residues 47−49). Both loops contain multiple glycines: Thr-Met-Gly-Gly-Arg and Ile-Gly-Asn-Gly-Arg-Gly. Mutations of glycine residues are known to be difficult to model[27] because the backbone geometry is likely to change significantly upon the insertion of a new residue. G13K/A138G, predicted to be very destabilizing but with no measured change from that of the wild type, is a good example.

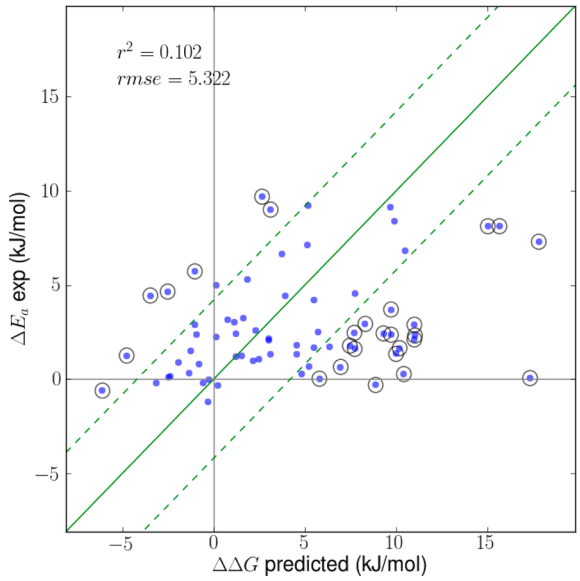


Figure 4. Correlation between predicted ΔE_a and $\Delta\Delta G$ values as predicted by PEATSA. The rmsd is 5.3 kJ/mol, and the correlation coefficient is 0.32. Outlier points marked with empty circles are those for which the residual is greater than the rmsd.

To provide an indication of which residues, if any, contribute consistently to high positive or negative errors, the feature classifier was applied to the error values. However, no residue had a significant weight in this analysis.

Two clones, N47H/G48D/A165H and G12H/N47H/G48D, were omitted from the correlation plot, as there are significant clashes in the mutated side chains as modeled by PEATSA. As discussed above, any changes to the glycines in this loop (loop 2) are likely to be disruptive, though surprisingly both of these clones are active. There are likely to be structural rearrangements in the backbone taking place that cannot be modeled because of the current limitations with PEATSA.

■ DISCUSSION

The engineering of protein stability continues to be of major importance for the application of enzymes in industrial processes, and it is therefore of interest to develop novel fast methods for obtaining a better understanding of the determinants of protein stability. In this work, we investigate the possibility of analyzing complex mutational data sets to obtain a detailed understanding of stability "hot spot" residues in a given protein. For a kinetically stabilized protein with a high T_{50}, the majority of point mutations identified will be destabilizing, and one may ask how the identification of such "negative" protein stability hot spots can aid in engineering a more stable protein. The answer to this question is connected with the deactivation processes governing the kinetically stabilized proteins, namely, that these proteins are deactivated irreversibly via a set of specific reaction pathways. For a kinetically stable protein, we will observe an effect for a mutation if it affects one of these deactivation pathways or if it opens a new deactivation pathway with a flow higher than or comparable to that of the other pathways.[28] It is reasonable to

dx.doi.org/10.1021/bi201926f | Biochemistry 2012, 51, 5339−5347

assume that the latter possibility is significantly less likely than the former, and consequently, stability hot spot residues identified by the approach put forth in this paper will therefore aid in identifying the structural elements that are involved deactivation steps. If we stabilize these structural elements by protein engineering, we can block or reduce the flow through one or more deactivation pathways and thus stabilize the protein further. Our analysis shows that loops 2 and 3 are major determinants of the stability of NCPP, and although our data set heavily oversamples loop 3 (see Figure 1) and we therefore can expect some bias toward residues in this loop, the method is nevertheless sensitive enough to identify Asn 47 in loop 2 that, on the basis of data from the αLP and TFPA systems, is very likely to be a major determinant of NCPP stability.

Sampling Error. Because our regression-based analysis uses data from multiple-mutation variants to determine a set of residue-specific weights, the robustness of the analysis will depend upon the characteristics of the particular variants that are analyzed. To ascertain that our method is able to identify meaningful weights, we studied its performance using a simulated experimental data set. In this data set, we know the residue-specific weights and can therefore generate data sets of any size to study the size dependence of the performance of the regression-based analysis. The results show that a given number of artificial weights (we used five significantly high weights of 40) can be reconstructed consistently by applying the fitting algorithm over multiple runs, each run representing a different set of random mutations. The standard error in weight estimates is typically ∼3.5% for a simulated data set with 120 samples and 40 features, though this will vary depending on (1) the relative significance of the true weight, (2) the amount of noise in the data, and (3) the number of samples and/or features (see Figure S9 of the Supporting Information). See section III of Supporting Information for further details. The results from true experimental data will be more complex, with nonlinear contributions for certain combinations of residues and/or amino acid substitutions. Nevertheless, we expect the method to be robust enough to successfully identify structural elements that are important for protein stability.

Constructing Structure-Based Protein Stability Hot Spot Maps. In this study, we have analyzed a data set that predominantly samples residues in surface loops of NCPP. However, in the general case, one can choose any subset of residues for hot spot analysis. Typically, one would choose only residues that are located in surface loops or areas that generally are thought to be potentially destabilizing. Indeed, selecting all the residues in a protein may mask the effect of a residue of interest by having too many highly destabilizing core mutations in the data set. In addition, choosing too many residues creates the problem of having to create a sufficient number of mutants to cover the whole protein. We therefore suggest the following rule of thumb for choosing a selection of variants when generating protein stability hot spot maps as outlined in this work.

Given that m is some algorithm to randomly generate a set of variants for a particular protein, it is a function of the following parameters:

$$m = f(N, M_{max}, S_{min}, R)$$

$$N \geq R \times 2;\ M_{max} \geq 3$$

where N is the number of variants (i.e., mutants produced), R is the subset of residues sampled (should be less than N for useful feature selection), M_{max} is the number of mutations per variant (should be allowed to vary among single, double, and triple mutations randomly up to a maximum of M_{max}; we recommended that this value should be at least 3 but lower than 8), and S_{min} is the minimum number of amino acid substitutions sampled per residue. This value will be limited for realistic experiments but should at least be 3−4 to avoid bias introduced by sampling only a specific substitution.

Additionally, the positions mutated in a single clone should be far from each other in the structure to minimize the chance of mutations affecting each other. Once the data set has been produced and characterized, the fitting and feature selection can be applied as described in the text. It may be useful to introduce a penalty to account for the effect of the particular set of amino acid substitutions sampled for each residue.

Note that this technique is not designed as a predictor of the effect of unknown mutations, but rather as a measure of the general effect of individual residues. We believe that the method can be a fast alternative to saturation mutagenesis experiments[29] or alanine scans, because the use of multiple mutations results in a significant decrease in the number of clones one has to characterize. An alanine scan of 100 residues requires characterization of as many clones, whereas a linear regression analysis using six mutant clones with an average of three samples per residue will require the characterization of only 50 clones [(100/6) × 3]. For saturation mutagenesis experiments, the possibility of the reduction being much larger depends on the exact nature of mutations chosen for each position. It should be mentioned that the method proposed here has some parallels to the ProSAR[30] methodology, and other regression-based QSAR methods for studying protein function.[31]

Using SPAs for Kinetically Stable Proteins. All current stability prediction algorithms (SPAs) predict mutational stability effects by calculating free energy changes between the folded states of wild-type and mutant proteins. The unfolded state is assumed to be unchanged, and thermodynamic stability is assumed. For kinetically stable proteins, this approach is clearly not valid, and we need to use a description of the protein that explicitly models different partial unfolding processes. In this respect, the COREX model of Hilser and coworkers[32] or a variant thereof can likely be used to identify likely unfolding intermediates, and the effect of mutations on the local unfolding equilibrium can then be assessed with current SPAs. In some cases, such an approach might, however, not be sufficient because some mutations inevitably would affect the activation barrier, and in these cases, it would therefore be necessary to model the transition state[33] of the inactivation process. There are thus significant obstacles to overcome in modeling the stability effects of mutations in kinetically stabilized proteins

■ CONCLUSION

We analyzed 121 mutants from thermal inactivation studies of the NCP protease and identify two surface loops that play a significant role in determining the stability of this enzyme. The identification of these two loops is corroborated by comparative studies of α-lytic protease and TFPA,[17] and they remain highly likely candidates for early unfolding events after visual inspection of the X-ray structure of NCP. This study shows that the analysis of 72 clones with an average of 2.8 mutations spread over a specific region of a 188-residue protein can yield a protein stability map from which likely early unfolding events

can be identified. Such maps may be generated for areas of proteins using a relatively small number of mutants in comparison to a full saturation mutagenesis experiment. Even given the simplistic assumptions of linearly additive stabilities, our method can give a convenient indication of the general stability of specific protein regions. We also demonstrate the performance of a structure prediction algorithm, PEATSA, on the same data set and find that although the correlation is significantly worse than for well-behaved proteins, the algorithm is still able to deliver predictions that indicate relative stabilities of many mutant clones.

■ ASSOCIATED CONTENT

Supporting Information

Figures detailing our experimental results from DSC, CD, and residual activity assays and details of our regression approach and how it was tested using simulated data. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Jens.Nielsen@gmail.com. Telephone: +45 30 77 16 92.

Present Address

§Section for Biomolecular Sciences, Department of Biology, Copenhagen Biocenter, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark.

Author Contributions

D.F. and H.W. contributed equally to this work.

Funding

This work was supported by the following grants to J.E.N.: Science Foundation Ireland PIYRA grant (04/YI1/M537), a project grant from the Irish Health Research Board (RP/2004/140), a Science Foundation Ireland Research Frontiers Program Award (08/RFP/BIC1140), and a Science Foundation Ireland Industry Supplement Award (04/YI.1/M537s1).

Notes

The authors declare the following competing financial interest(s): T.A.P., L.L.H.C., L.B., T.V.B., and J.E.N. are employees of Novozymes A/S.

■ REFERENCES

(1) Rubingh, D. N. (1997) Protein engineering from a bioindustrial point of view. *Curr. Opin. Biotechnol. 8*, 417−422.

(2) Richardson, T. H., Tan, X., Frey, G., Callen, W., Cabell, M., Lam, D., Macomber, J., Short, J. M., Robertson, D. E., and Miller, C. (2002) A novel, high performance enzyme for starch liquefaction. Discovery and optimization of a low pH, thermostable α-amylase. *J. Biol. Chem. 277*, 26501−26507.

(3) Clarke, N. D. (2010) Protein engineering for bioenergy and biomass-based chemicals. *Curr. Opin. Struct. Biol. 20*, 527−532.

(4) Henrick, K., Feng, Z., Bluhm, W. F., Dimitropoulos, D., Doreleijers, J. F., Dutta, S., Flippen-Anderson, J. L., Ionides, J., Kamada, C., Krissinel, E., Lawson, C. L., Markley, J. L., Nakamura, H., Newman, R., Shimizu, Y., Swaminathan, J., Velankar, S., Ory, J., Ulrich, E. L., Vranken, W., Westbrook, J., Yamashita, R., Yang, H., Young, J., Yousufuddin, M., and Berman, H. M. (2008) Remediation of the protein data bank archive. *Nucleic Acids Res. 36*, D426−D433.

(5) Maurer, K. H. (2004) Detergent proteases. *Curr. Opin. Biotechnol. 15*, 330−334.

(6) Bloom, J. D., Wilke, C. O., Arnold, F. H., and Adami, C. (2004) Stability and the evolvability of function in a model protein. *Biophys. J. 86*, 2758−2764.

(7) Van den Burg, B., Vriend, G., Veltman, O. R., Venema, G., and Eijsink, V. G. (1998) Engineering an enzyme to resist boiling. *Proc. Natl. Acad. Sci. U.S.A. 95*, 2056−2060.

(8) Gribenko, A. V., Patel, M. M., Liu, J., McCallum, S. A., Wang, C., and Makhatadze, G. I. (2009) Rational stabilization of enzymes by computational redesign of surface charge-charge interactions. *Proc. Natl. Acad. Sci. U.S.A. 106*, 2601−2606.

(9) Eijsink, V. G., Gaseidnes, S., Borchert, T. V., and van den Burg, B. (2005) Directed evolution of enzyme stability. *Biomol. Eng. 22*, 21−30.

(10) Farinas, E. T., Bulter, T., and Arnold, F. H. (2001) Directed enzyme evolution. *Curr. Opin. Biotechnol. 12*, 545−551.

(11) Sanchez-Ruiz, J. M. (2010) Protein kinetic stability. *Biophys. Chem. 148*, 1−15.

(12) Novozymes (2005) Process for production of low temperature active alkaline protease froma deep-sea fungus, Raghukumar, Chandralata, United States.

(13) Kelch, B. A., Eagen, K. P., Erciyas, F. P., Humphris, E. L., Thomason, A. R., Mitsuiki, S., and Agard, D. A. (2007) Structural and mechanistic exploration of acid resistance: Kinetic stability facilitates evolution of extremophilic behavior. *J. Mol. Biol. 368*, 870−883.

(14) Jaswal, S. S., Sohl, J. L., Davis, J. H., and Agard, D. A. (2002) Energetic landscape of α-lytic protease optimizes longevity through kinetic stability. *Nature 415*, 343−346.

(15) Lumry, R., and Eyring, H. (1954) Conformational changes of proteins. *J. Phys. Chem. 58*, 110−120.

(16) Baker, D., Sohl, J. L., and Agard, D. A. (1992) A protein-folding reaction under kinetic control. *Nature 356*, 263−265.

(17) Kelch, B. A., and Agard, D. A. (2007) Mesophile versus thermophile: Insights into the structural mechanisms of kinetic stability. *J. Mol. Biol. 370*, 784−795.

(18) Johnston, M. A., Søndergaard, C. R., and Nielsen, J. E. (2010) Integrated prediction of the effect of mutations on multiple protein characteristics. *Proteins: Struct., Funct., Bioinf. 79*, 165−178.

(19) Johnston, M. A., Farrell, D., and Nielsen, J. E. (2012) A Collaborative Environment for Developing and Validating Predictive Tools for Protein Biophysical Characteristics. *J. Comput.-Aided Mol. Des. 26*, 387−396.

(20) Farrell, D., Miranda, E., Webb, H., Georgi, N., Crowley, P., McIntosh, L., and Nielsen, J. E. (2010) Titration_DB: Storage and analysis of NMR-monitored protein pH titration curves. *Proteins 4*, 843−857.

(21) Microcal (1998) *DSC Data Analysis in Origin-Tutorial Guide.*

(22) Guyon, I., and Elisseeff, A. (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research 3*, 1157−1182.

(23) Efron, B. (2004) Least angle regression. *Annals of Statistics 32*, 407−499.

(24) Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., and Pollmann, S. (2009) PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics 7*, 37−53.

(25) Fuhrmann, C. N., Kelch, B. A., Ota, N., and Agard, D. A. (2004) The 0.83 Å resolution crystal structure of α-lytic protease reveals the detailed structure of the active site and identifies a source of conformational strain. *J. Mol. Biol. 338*, 999−1013.

(26) Peters, R. J., Shiau, A. K., Sohl, J. L., Anderson, D. E., Tang, G., Silen, J. L., and Agard, D. A. (1998) Pro region C-terminus:protease active site interactions are critical in catalyzing the folding of α-lytic protease. *Biochemistry 37*, 12058−12067.

(27) Feyfant, E., Sali, A., and Fiser, A. (2007) Modeling mutations in protein structures. *Protein Sci. 16*, 2030−2041.

(28) Vriend, G., Berendsen, H. J., van den Burg, B., Venema, G., and Eijsink, V. G. (1998) Early steps in the unfolding of thermolysin-like proteases. *J. Biol. Chem. 273*, 35074−35077.

(29) Reetz, M. T., and Carballeira, J. D. (2007) Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat. Protoc. 2*, 891−903.

(30) Fox, R. J., Davis, S. C., Mundorff, E. C., Newman, L. M., Gavrilovic, V., Ma, S. K., Chung, L. M., Ching, C., Tam, S., Muley, S., Grate, J., Gruber, J., Whitman, J. C., Sheldon, R. A., and Huisman, G.

W. (2007) Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol. 25*, 338−344.
(31) Sachan, N., Kadam, S. S., and Kulkarni, V. M. (2007) Human protein tyrosine phosphatase 1B inhibitors: QSAR by genetic function approximation. *J. Enzyme Inhib. Med. Chem. 22*, 267−276 , 371−373.
(32) Vertrees, J., Barritt, P., Whitten, S., and Hilser, V. J. (2005) COREX/BEST server: A web browser-based program that calculates regional stability variations within protein structures. *Bioinformatics 21*, 3318−3319.
(33) Anderson, D. E., Peters, R. J., Wilk, B., and Agard, D. A. (1999) α-Lytic protease precursor: Characterization of a structured folding intermediate. *Biochemistry 38*, 4728−4735.
(34) Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., and Ben-Tal, N. (2005) ConSurf 2005: The projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res. 33*, W299−W302.