HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Development of computational approaches for knowledge-driven protein engineering aimed at improving thermostability

Inaugural dissertation

for the attainment of the title of doctor
in the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

presented by

Prakash Chandra Rathi
from Pokaran, India

Düsseldorf, October 2014

from the Institute for Pharmaceutical and Medicinal Chemistry
at the Heinrich Heine University Düsseldorf

Published by permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich Heine University Düsseldorf

Supervisor: Prof. Dr. Holger Gohlke
Co-supervisor: Prof. Dr. Karl-Erich Jaeger

Date of the oral examination: 31.10.2014

I declare under oath that I have compiled my dissertation independently and without any undue assistance by third parties under consideration of the 'Principles for the Safeguarding of Good Scientific Practice at Heinrich Heine University Düsseldorf'.

This dissertation has not been submitted in its present or a similar form in any other institution. I have not made any successful or unsuccessful attempt to obtain a doctorate before.

Düsseldorf, 31.10.2014

(Prakash Chandra Rathi)

TABLE OF CONTENTS

LIST OF PUBLICATIONS INCLUDED IN THIS THESIS

(Contribution in parentheses)

Peer-reviewed publications

I. **Prakash Chandra Rathi (70%)**, Hans Wolfgang Höffken, and Holger Gohlke
Quality matters: Extension of clusters of residues with good hydrophobic contacts stabilize (hyper)thermophilic proteins. *Journal of Chemical Information and Modeling*, **2014**, 54: 355–361.
Impact factor reported for 2012: 4.30

II. **Prakash Chandra Rathi (65%)**, Sebastian Radestock, and Holger Gohlke.
Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *Journal of Biotechnology,* **2012**, 159:135-144.
Impact factor reported for 2012: 3.18

III. Christopher Pfleger[§], **Prakash Chandra Rathi (35%)**[§], Doris L. Klein, Sebastian Radestock, and Holger Gohlke
Constraint Network Analysis (CNA): A Python software package for efficiently linking biomacromolecular structure, flexibility, (thermo)stability, and function. *Journal of Chemical Information and Modeling*, **2013**, 53:1007-1015.
Impact factor reported for 2012: 4.30

IV. Dennis M. Krüger[§], **Prakash Chandra Rathi (30%)**[§], Christopher Pfleger, and Holger Gohlke
CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo)stability, and function. *Nucleic Acids Research*, **2013**, 41:W340-348.
Impact factor reported for 2012: 8.27

Other publications

V. **Prakash Chandra Rathi (40%)**[§], Daniel Mulnaes[§], and Holger Gohlke
VisualCNA: A GUI for interactive Constraint Network Analysis and protein engineering for improving thermostability, **2014** (submitted manuscript).

VI. **Prakash Chandra Rathi (75%),** Karl-Erich Jaeger, and Holger Gohlke
Structural rigidity and protein thermostability, **2014** (submitted manuscript).

VII. **Prakash Chandra Rathi (40%)**, Alexander Fulton, Karl-Erich Jaeger, and Holger Gohlke
Application of rigidity theory to the thermostabilization of proteins, **2014** (submitted manuscript).

VIII. **Prakash Chandra Rathi** (**30%**), Christopher Pfleger, Simone Fulle, Doris L. Klein, and Holger Gohlke
Statics of biomacromolecules. in: "Modeling of Molecular Properties" , P. Comba (ed.), S. 281-299, Wiley-VCH, Weinheim, 2011.

Publications I-IV and VIII have been reprinted with kind permission from their respective publishers.

[§] Both authors share first authorship of the respective publication.

ABBREVIATIONS

AK	Adenylate Kinase
*Bs*LipA	*Bacillus subtilis* lipase A
CD	Circular Dichroism
CNA	Constraint Network Analysis
CoSM	Computational Saturation Mutagenesis
C_p	Heat capacity at constant pressure
CS	Citrate Synthase
DCM	Distance Constraint Model
DHFR	DiHydroFolate Reductase
DSC	Differential Scanning Calorimetry
DSF	Differential Scanning Fluorimetry
e.g.	*exempli gratia* (Latin for "for example")
ENT^{FNC}	Ensemble of Network Topologies derived from Fuzzy Non-covalent Constraint definitions
et al.	*et alii* (Latin for "and others")
FIRST	Floppy Inclusion and Rigid Substructure Topography
F_{mxw}	Number of floppy modes (using Maxwell's formula)
GUI	Graphical User Interface
H	Cluster configuration entropy
HDX	Hydrogen/deuterium exchange
HEWL	Hen Egg White Lysozyme
i.e.	*id est* (Latin for "that is")
ICE	Improved Configurational Entropy
KINARI	KINematic And RIgidity analysis of proteins
MD	Molecular Dynamics
mDCM	minimal Distance Constraint Model
MSA	Multiple Sequence Alignment
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PDF	Probability Density Function
$\widetilde{rc}_{ij,neighbor}$	Median stability of rigid contacts between residue-neighbors
SWIG	Simplified Wrapper and Interface Generator
TKSA-GA	Tanford–Kirkwood Surface Accessibility Genetic Algorithm
TLP	Thermolysin-Like Protease
T_m	Melting temperature
T_{og}	Optimal growth temperature
T_{opt}	Temperature of maximum activity
T_p	Phase transition temperature calculated using CNA
T_{50}^t	Temperature required for half inactivation of a protein in time t
vdW	van der Waals
VPG	Virtual Pebble Game
vs.	versus
WT	Wild-type
Å	Ångström
ºC	Degree Celsius

ΔG_{FU}	Difference between the Gibbs free energy of folded and unfolded state of a protein

Amino acid	**Three letter code**	**Single letter code**
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

ZUSAMMENFASSUNG

Hohe Thermostabilität ist eine erwünschte Eigenschaft von Proteinen, insbesondere für deren Einsatz in der industriellen Biokatalyse. Ein Biokatalysator mit erhöhter Thermostabilität ermöglicht die Durchführung einer Katalyse bei höheren Temperaturen und dadurch die Erhöhung der Reaktionsgeschwindigkeit. Allerdings sind die meisten Proteine in der Natur nicht dafür optimiert, harten industriellen Prozessbedingungen, einschließlich hoher Temperaturen, zu widerstehen. Daher wendet man häufig Protein-Engineering an, um thermostabile Varianten existierender Proteine zu erzeugen. Jedoch kann nur ein kleiner Teil der vielen theoretisch möglichen Varianten eines Proteins experimentell hinsichtlich Thermostabilität getestet werden. Daher sind computergestützte Ansätze zur Vorhersage von *Weak Spots*, Aminosäureresten deren Mutation die Thermostabilität eines Proteins voraussichtlich erhöht, vielversprechend für Protein-Engineering-Projekte zur Verbesserung der Thermostabilität. In dieser Arbeit entwickelte ich solche computergestützten Ansätze, eine Aufgabe die drei Teile umfasste: I) Die Identifikation der wichtigsten nicht-kovalenten Wechselwirkungen, welche die Thermostabilität eines Proteins bestimmen. II) Die Entwicklung und Verbesserung von Ansätzen zur Vorhersage thermostabilisierender Protein-Mutationen auf Grundlage der Erkenntnisse aus Teilaufgabe I. III) Die Validierung dieser Ansätze durch ihre retrospektive und prospektive Anwendung auf Testsysteme.

In dieser kumulativen Arbeit zeigte ich, dass die hydrophobe Wechselwirkungsenergie der wichtigste Faktor bei der Unterscheidung von mesophilen und (hyper)thermophilen Protein-Homologen ist. Auf Basis dieser Information entwickelte ich einen Ansatz zur Vorhersage von *Weak Spots* auf Grundlage der hydrophoben Wechselwirkungsenergien einzelner Aminosäurereste (**Publikation I**). Erstmals zeigte ich, dass die Größe von Aminosäurerest-Clustern, die anhand hydrophober Wechselwirkungsenergien einzelner Aminosäurereste identifiziert wurden, ein wesentlich besseres Unterscheidungsmerkmal für mesophile und (hyper)thermophile Proteine ist als allein das Vorhandensein oder die Größe hydrophober Aminosäurerest-Cluster. Auf Grundlage dieser Erkenntnis verbesserte ich den auf der Rigiditäts-Theorie basierenden *Constraint Network Analysis* (CNA) Ansatz, um durch die temperaturabhängige Modellierung hydrophober Interaktionen und zusätzlich Ensemble-basierte CNA die Thermostabilität und Weak Spots von Proteinen vorherzusagen (**Publikation II**). Zur einfachen Vorbereitung und ausführlichen Analyse von CNA-Berechnungen, wurde ein Software-Paket, ein Web-Service als auch eine graphische Benutzeroberfläche entwickelt, welche das Protein-Engineering zur Verbesserung der

Thermostabilität unterstützen. (**Publikationen III–V**). Ich verwendete sodann CNA zur retrospektiven Untersuchung der Beziehung zwischen der strukturellen Rigidität von Proteinen und deren thermodynamischer Thermostabilität, wobei Lipase A von *Bacillus subtilis* (*Bs*LipA) als Testfall mit literaturbekannten thermodynamisch thermostabilisierten Varianten diente (**Publikation VI**). Meine vergleichenden Untersuchungen von *Bs*LipA Varianten zeigten, dass eine thermodynamische Thermostabilisierung eindeutig mit einer erhöhten strukturellen Rigidität einhergeht. Diese Erkenntnis wurde in einer prospektiven Studie ausgenutzt: um den CNA-basierten Ansatz zu validieren, entwickelte ich eine computergestützte Strategie zur Vorhersage thermostabilisierender Mutationen und wendete diese auf *Bs*LipA an. Experimentelle Tests bestätigten die vorhergesagte Thermostabilisierung für drei von zwölf mutierten *Bs*LipA Varianten (**Publikation VII**). Diese Studie zeigte erstmals, dass CNA prospektiv zur Anreicherung thermostabilisierender Protein-Mutationen verwendet werden kann.

Ich bin davon überzeugt, dass die Geschwindigkeit und Vorhersagegenauigkeit dieser neuen und verbesserten computergestützten Ansätze es ermöglichen werden, die Grundlagen der Thermostabilität von Proteinen zu erforschen, thermostabilisierende Mutationen vorherzusagen und dadurch die Wirksamkeit und Effizienz von Protein-Engineering-Projekten zu verbessern.

ABSTRACT

High thermostability is a desired property for proteins, in particular for their use in industrial bio-catalysis. A bio-catalyst with elevated thermostability allows carrying out catalysis at higher temperatures and, thus, to increase the rate of a reaction. However, most proteins in nature are not optimized to withstand harsh industrial process conditions including high temperatures. Therefore, engineering existing proteins by mutagenesis is often employed in order to produce thermostable variants. However, only a minor fraction of the large number of theoretically possible mutated variants of a protein can be tested experimentally for their thermostability. Computational approaches that predict *weak spots*, residues that are more likely to increase a protein's thermostability upon mutation, are hence promising for protein engineering projects aimed at improving thermostability. In this thesis, I developed such computational approaches, a task that was subdivided in three main parts: I) To identify the most significant non-covalent interactions that determine a protein's thermostability. II) To develop and improve approaches for predicting thermostabilizing mutations on a protein based on the outcome of part I. III) To validate these approaches by their retrospective and prospective application on test systems.

In this compilation thesis, I showed that hydrophobic interaction energy is the most discriminating factor between mesophilic and (hyper)thermophilic protein homologs. Using this information, I developed an approach for predicting weak spots based on residue-wise hydrophobic interaction energies (**Publication I**). For the first time, I showed that the size of residue clusters that are identified based on residue-wise hydrophobic interaction energies discriminates mesophilic and (hyper)thermophilic proteins much better than the existence or size of clusters of hydrophobic residues alone. Based on this finding, I improved the rigidity theory-based Constraint Network Analysis (CNA) approach for predicting protein thermostability and weak spots by modeling hydrophobic interactions in a temperature-dependent manner, in addition to performing an ensemble-based CNA (**Publication II**). For an easy setup and extensive analysis of CNA calculations, we developed a software package, a web service, as well as a graphical user interface that facilitate protein engineering for improving thermostability (**Publications III-V**). Next, I used CNA to study the relation of a protein's structural rigidity and its thermodynamic thermostability using *Bs*LipA as a test case for which thermodynamically thermostabilized variants are reported in the literature (**Publication VI**). For the first time, my systematic comparative study of *Bs*LipA variants revealed that thermodynamic thermostabilization is unequivocally accompanied by increased

structural rigidity, leading to a significant and good correlation between structural rigidity and thermodynamic thermostability of these variants. Finally, in order to validate the CNA approach, I developed a computational strategy for predicting thermostabilizing mutations and applied it to lipase A from *Bacillus subtilis* (*Bs*LipA) prospectively. Experimental testing confirmed the predicted thermostabilization for three out of twelve mutated *Bs*LipA variants (**Publication VII**). This study demonstrated, for the first time, that CNA can be applied prospectively for enriching thermostabilizing mutations on a protein.

I am confident that the speed and prediction accuracy of these new and improved computational approaches will allow to investigate the basis of protein thermostability, to predict mutations that increase thermostability, and thus to improve the efficacy and efficiency of protein engineering projects.

1 INTRODUCTION

Proteins are central in carrying out biological functions. In order to carry out their functions, proteins are required to fold in a complex three-dimensional structure in order to interact with their binding partners.[1] Several extrinsic factors including temperature,[2-8] pressure,[9-13] and solvent[14-20] affect a protein's stability by disrupting its folded structure.[21] Among these factors, effect of temperature has been extensively studied owing to its significance in industrial applications.[22-27] Thermostable enzymes are sought after in biotechnological industry because these allow carrying out biocatalysis at elevated temperatures.[28,29] Enzymes found in nature are not always optimized to withstand extreme industrial process conditions including high temperature.[30] Therefore, the identification and development of thermostable enzymes is an important aspect of research in biotechnology.[25]

In general, enzymes from (hyper)thermophiles*, *i.e.*, organisms that grow optimally at temperatures above 50°C (85°C), show a higher temperature tolerance than their orthologs from mesophiles, *i.e.*, organisms with an optimal growth temperature (T_{og}) of 25-50°C.[28,31] Consequently, identifying thermostable enzymes by screening metagenomes is an obvious approach. Screening large metagenomes is often highly cumbersome, and a thermostable variant of the desired protein is not necessarily available in nature, however.[32] This makes engineering of existing enzymes for improving their thermostability a valuable alternative.[33]

Protein engineering approaches for improving thermostability include random mutagenesis and recombination followed by screening for thermostable mutants,[34] rational design,[35] and data-driven approaches.[36] On the one hand, random mutagenesis has a limitation in that only a restricted sequence space can be experimentally tested owing to the large combination of mutations that are theoretically possible;[37] on the other hand, rational design of a protein requires a thorough knowledge of the mechanisms underlying thermostabilization of the protein.[35] As a compromise, data-driven approaches are being pursued for reducing the library size for mutagenesis based on suggestions of interesting residue positions that, when mutated, would lead to a more thermostable variant of the protein.[36] These data-driven approaches rely upon knowing the most important intrinsic factor(s) that improve thermostability of a protein. Comparisons of pairs of meso- and (hyper)thermophilic proteins have revealed several such factors,[38,39] including improved hydrogen bonding,[40] ion pair and

* Hereinafter, "mesophilic protein" is used as a synonym for a protein from a mesophilic organism and, similarly, "(hyper)thermophilic protein" is used as a synonym for a protein from a (hyper)thermophilic organism.

salt bridge networks,[39] better hydrophobic packing,[41] shortening of loops,[42] and higher secondary structure content.[43] Factors contributing to protein thermostability are reviewed by several authors.[4,44-47] The multitude of the factors that contribute to protein thermostability poses a pertinent question: which of these factor(s) is (are) most significant for protein thermostability? The answer to the question will enable us to improve data driven computational approaches by emphasizing the most significant interactions while modeling a protein.

Data-driven approaches that restrict the mutant library size for a protein using the available knowledge are frequently being used for improving protein thermostability.[36] These approaches exploit information derived from sequence and/or structure of proteins. In one of the sequence-based data-driven approaches, amino acids in a target mesophilic protein are substituted with the ones from a (hyper)thermophilic homologue using site-directed mutagenesis.[48,49] Other methods involve generating a consensus sequence of the target protein from a multiple sequence alignment (MSA) of several mesophilic sequences of the same protein by keeping consensus amino acids from the alignment at every sequence position.[50-52] Several structure-based data-driven approaches predict the change in ΔG_{FU} (difference between the Gibbs free energy of folded and unfolded state of a protein) upon mutation of a target protein by employing empirical or knowledge-based potentials[53-58] or machine learning.[59,60] These methods can be employed to identify a set of potential mutations that can be experimentally evaluated. Taking advantage of high performance computing, free energy calculations for mutations using thermodynamic integration is also pursued to improve protein thermostability.[61]

In another structure-based data-driven approach introduced by M. T. Reetz *et al.*,[62] saturation mutagenesis was performed in an iterative manner on those residues of *Bacillus subtilis* lipase A (*Bs*LipA) with the highest crystallographic B-factors leading to development of significantly more thermostable variants compared to wild type (WT) by screening less than 8000 colonies. This followed the guiding principle that thermostable proteins usually show a higher degree of structural rigidity than their counterparts from mesophilic organisms; hence, preferentially stabilizing the most mobile regions should increase thermostability.

Following the same guiding principle to exploit the link between rigidity and protein thermostability, S. Radestock and H. Gohlke developed a graph theory-based rigidity analysis approach termed Constraint Network Analysis (CNA) for protein thermostability prediction

and identification of weak spots.[63,64] In CNA, a protein is modeled as a constraint network where atoms (sites) are connected by covalent and noncovalent interactions (constraints).[65] Then rigidity analysis of the protein network is performed using the *pebble game* algorithm wherein the network is decomposed into rigid parts and flexible links between them.[66,67] By successively removing hydrogen bonds from the network in increasing order of strength, a thermal unfolding simulation is carried out.[64,68] So far, hydrophobic interactions have not been considered in a temperature-dependent manner and, accordingly, hydrophobic constraints were kept constant throughout the thermal unfolding simulation. Next, a phase transition point T_p at which a largely rigid network becomes largely flexible is identified that relates to the (thermodynamic) thermostability of the protein and can be compared to T_m values. In addition to this, weak spots are identified as the residues that become flexible during the phase transition.[45,46]

It has been found in several studies that thermophilic proteins have a more rigid fold than their mesophilic homologs.[63,64,69,70] As an opposing view, proteins from thermophilic organisms have been reported to be as flexible as or even more flexible than their homologs from mesophilic organisms.[71-74] These different views on the relation between protein thermostability and structural rigidity have been a matter of ongoing discussion,[69,75-80] and may be related to that, from a mechanistic point of view, the general term "protein thermostability" embraces at least two different meanings:[30,81] (1) *thermodynamic* thermostability describes the folded-unfolded equilibrium of a protein, and (2) *kinetic* thermostability refers to the length of time a protein remains active before undergoing irreversible denaturation at an elevated temperature. When the folded structure of a protein is energetically more stable than its unfolded structure, the protein is said to be *thermodynamically* thermostable. A *kinetically* thermostable protein is less prone to precipitation and aggregation and, hence, is not readily inactivated at high temperatures. We hypothesize that a lack of differentiation between thermodynamic and kinetic thermostabilization contributes in part, to the opposing views of "increased *vs.* decreased structural rigidity" and "protein thermostability.

In the present thesis, I identified the most significant determinants of protein thermostability and used this information to improve constraint modeling in the CNA approach. The CNA approach was also improved by devising a way to incorporate a structural ensemble as input, rather than a single protein structure as done before. Next, we developed software packages

for performing CNA calculations and analyzing results from CNA in an interactive manner. Furthermore, a web service was developed for allowing setting up and running CNA calculations from a web browser. Next, we studied the link between structural rigidity and protein thermostability. This was done using *Bs*LipA as test case because several kinetically and thermodynamically thermostabilized mutants of *Bs*LipA have been reported in the literature.[62,82-88] Finally, the CNA approach and software packages were validated by predicting thermostabilizing mutants of *Bs*LipA and subsequent experimental thermostability estimation.

2 BACKGROUND

2.1 Protein thermostability

Proteins are complex three-dimensional molecular machines that carry out biological functions. A folded form of a protein brings key amino acids together to catalyze an enzymatic reaction. However, several adverse conditions including high temperature and presence of denaturants, solvents, and extreme pH can transform a protein into a form that is not active. The term *thermostability* refers to the ability of a protein to resist adverse effects of high temperature by preserving its tertiary active structure. Unfolding of a protein in to an inactive, disordered polypeptide chain, which can be reversible, is termed *denaturation*.[81] Another term *inactivation* refers to an irreversible loss of activity of a protein over time due to physical, biological, or chemical factors.[81] These factors include precipitation and aggregation,[89-93] deleterious reactions on amino acid side-chains,[94] and proteolysis.[95,96]

Two concepts of thermostability arise out of these deleterious phenomena: *thermodynamic* and *kinetic* thermostability. Thermodynamic thermostability refers to the stability of the folded form of a protein in comparison to its unfolded form: Thermodynamically thermostable proteins have lower free energy of unfolding (ΔG_{FU}) than less thermostable proteins. In contrast, a kinetically thermostable protein resists the process of inactivation, and does not necessarily need to be thermodynamically stable.[30,81] However, a thermodynamically less stable protein is usually also kinetically less stable because an unfolded protein is more susceptible to factors responsible for kinetic instability than the folded form. These concepts of thermostability can be expressed in a simplistic scheme (eq. (1))[30,81]

$$\boldsymbol{N \overset{K}{\rightleftharpoons} U \overset{k}{\rightarrow} I} \tag{1}$$

where N, U, and I represent the native, unfolded and inactivated form of a protein, respectively. The equilibrium constant K belongs to the reversible folded-unfolded transition (related to thermodynamic thermostability), and the rate constant k governs the transformation of an unfolded form to an inactivated form (related to kinetic stability). However, multiple (n) relatively stable intermediates can occur on the path of the folded-unfolded transformation with distinct equilibrium constants K_1 to K_n, and each stable intermediate can convert into an inactivated form with a distinct rate constant k_1 to k_n.[30]

2.2 Protein thermostability and rigidity

Structural rigidity (or flexibility) is an important property of proteins, and has been associated with their thermostability,[69] in addition to molecular recognition as well as catalysis.[97,98] Rigidity of proteins has been frequently characterized experimentally using X-ray crystallography, cryo-electron microscopy, single molecule fluorescence, nuclear magnetic resonance (NMR) spectroscopy, amide hydrogen/deuterium exchange (HDX) etc.[99-104] It has been observed that thermophilic proteins have an increased structural rigidity than their mesophilic counterparts.[63,64,69,70] However, a delicate balance in rigidity and flexibility is required for an optimal function of a protein. On the one hand, an overall rigid fold provides resistance to the unfolding at high temperatures; on the other hand, a flexible active site is required for carrying out catalysis.[105-108] This phenomenon has a profound implication in the field of protein engineering aimed at improving thermostability in that a thermostable mutant that improves global rigidity of the fold should retain the active site flexibility to allow its function. As an opposing view, proteins from thermophilic organisms have been reported to be as flexible as or even more flexible than homologs from mesophilic organisms.[71-74] G. Hernández and D. M. LeMaster observed that the HDX rates are comparable for most parts of mesophilic and thermophilic rubridoxin at 23°C with latter showing even higher flexibility in multiple-turn region.[71] Similarly, J. Fitter and J. Heberle found comparable rates of slowly exchanging amide protons for mesophilic and thermophilic α-amylase and a higher flexibility for the latter with respect to motions on shorter time scales.[73] These different views on the relation between protein thermostability and structural rigidity have been a matter of ongoing discussion.[69,75-80]

The difference in the temporal resolution of the experimental technique or computational analysis used to detect protein flexibility[99,100,102-104,109,110] contribute, in part, to the opposing views of "increased *vs.* decreased structural rigidity" and "protein thermostability". In the present thesis, we address the question of the relation between protein thermostability and structural rigidity using the rigidity theory-based CNA approach, which characterizes protein rigidity and flexibility as static properties thereby in a time-independent manner. (**Publication VI**).

2.3 Factors contributing to protein thermostability

Proteins from (hyper)thermophilic organisms (T_{og} > 50ºC(85ºC)) often have a very similar fold to and an identical function as the mesophilic homologs (T_{og} < 50ºC).[28,31] However, they

retain their activity at higher temperatures where their counterparts from mesophilic organisms are usually denatured or inactivated. By comparing proteins from mesophilic organisms with their thermophilic homologs, several intrinsic factors have been identified that improve thermodynamic thermostability of proteins. These factors include improved hydrogen bonding,[40] ion pair and salt bridge networks,[39] better hydrophobic packing,[41] shortening of loops,[42] higher secondary structure content[43], and increased rigidity of a protein.[63,64,69,70,111] Overall, an optimized network of these interactions/determinants makes the fold of a thermodynamically thermostable protein energetically more favorable than its unfolded form.[69,112] These determinants can also improve kinetic thermostability of a protein by reducing the rate of unfolding, when incorporated at unfolding initiation sites of the protein.[113]

The multitude of the factors that contribute to protein thermostability poses a pertinent question: which of these factor(s) are most significant for protein thermostability. The answer to the question will enable us to more accurately tune data-driven approaches aimed at predicting *weak spots* on a protein *i.e.*, residues that are more likely to improve protein thermostability upon mutation. We hypothesize that the reason why different determinants of thermostability have been revealed in previous studies[39-43,63,64,69,70,111] lies in that the focus of these analyses has been on structural factors. At variance from these studies, in the present thesis, we identify the most significant factors responsible for improved thermostability of proteins using a large test set of 132 pairs of mesophilic/thermophilic and 149 pairs of mesophilic/hyperthermophilic homologous protomers by comparing the quality (energy) of different non-covalent interactions (**Publication I**).

2.4 Experimental measures and methods of thermostability characterization

Thermodynamic stability refers to the stability of the folded form of a protein in comparison to its unfolded form. Hence, thermodynamic stability can be measured by monitoring the reversible folded-unfolded transition of a protein when applying a gradient temperature ramp. The most commonly used technique for measuring thermodynamic stability is circular dichroism (CD) spectroscopy that estimates secondary structure content of a protein by measuring change in the ellipticity using a circularly polarized light.[114] On the ellipticity *vs.* temperature curve, the melting temperature (T_m) is then identified as the temperature that indicates when half of the protein is unfolded. The folded-unfolded transition can also be identified by measuring the intensity of fluorescence emitted by hydrophobic amino acids,

particularly Trp.[115-117] Another method is differential scanning fluorimetry (DSF), which measures the intensity of fluorescence emitted by a dye upon binding to hydrophobic residues that usually form the core of a globular protein.[118-121] Protein unfolding can be quantitatively estimated by measuring the intensity of the fluorescence: the more a protein unfolds the more of the dye binds to hydrophobic residues that are exposed and, hence, more intensity is recorded. Similar to CD spectroscopy, T_m values can be calculated on the intensity *vs.* temperature curves in the case of fluorescence and DSF experiments. The DSF can be carried out in a high throughput manner in any machine that provides a temperature ramp and accurate measurement of fluorescence intensity, e.g., it can be performed in well plates using a real time polymerase chain reaction machine.[119,122] Differential scanning calorimetry (DSC) is another technique used to characterize protein thermostability. In DSC, heat capacity at constant pressure (C_p) is measured as a function of temperature. The resulting C_p *vs.* temperature curve is used to calculate enthalpy of unfolding and T_m.[123-125] Free energy of unfolding (ΔG_{FU}) is a thermodynamic measure of a protein's stability. Using eq. (2), ΔG_{FU} can be calculated from a protein's denaturation curve (*e.g.*, fluorescence *vs.* denaturant) obtained using thermal or chemical (*e.g.*, urea/guanidinium HCl) denaturants.[126] Since ΔG_{FU} is a thermodynamic measure of protein stability, it is essential that the measurements are done when equilibrium is attained and the unfolding reaction is reversible.[126]

$$\Delta G_{FU} = -RT \ln K_{eq} = -RT \ln \left(\frac{f_D}{1 - f_D}\right) \quad (2)$$

Here R, T, $K_{eq,}$ and f_D are the gas constant, temperature, folded-unfolded equilibrium constant, and the fraction of protein that is unfolded, respectively.

When a protein denatures irreversibly due to events leading to a rapid deformation of the unfolded form, such as aggregation or proteolytic degradation, the kinetic stability becomes a more important indicator of its thermostability than the thermodynamic stability. In such cases, the free energy difference between the folded and the transition states on the folded-unfolded path becomes more important because once a protein becomes unfolded, it is exposed to an irreversible denaturation. Kinetic thermostability of a protein is measured by the rate at which the protein is inactivated. One of the measures of kinetic stability, the *half-life* is the time required to reduce activity of a protein to its half-maximal activity at a given temperature (or other condition).[127] Other measures for thermostability include the temperature required for half inactivation of a protein in time t (T_{50}^{t}),[62] residual activity at a

fixed time and temperature,[128,129] and temperature of maximum activity (T_{opt}).[2] A higher T_{50}^{t} indicates that the protein can resist (half) inactivation at a higher temperature when a larger fraction of the protein is in unfolded form than at a lower temperature. T_{opt} indicates kinetic stability indicating a trade-off between the gain in activity and the increased susceptibility for inactivation as temperature rises.

2.5 Modes of protein thermostabilization

A protein can be either intrinsically stabilized by incorporating thermostabilizing mutations or alternatively, by modulating extrinsic factors. Frequently, as an extrinsic mode of protein stabilization, engineering buffer conditions of a protein with respect to salt concentration, salt type, and pH is carried out leading to an increase in thermostability of a protein.[130] Intrinsic stabilization involves mutating residues of a protein with an aim to identify thermostable variant; this process is termed *protein engineering*. In nature, over the years, proteins have evolved by selecting advantageous mutations to suit environmental conditions of the host organisms. In an approach of protein engineering termed *directed evolution,* the natural evolution is mimicked in that the proteins are randomly mutated under a selection pressure (*e.g.*, ability to be active at high temperatures).[34] This process can be carried out iteratively generating variants that are more thermostable than the members of their parent generation in each cycle. Several successful attempts of directed evolution aimed at improving thermostability have been reported.[88,131-137] Directed evolution has a limitation, however, in that only a restricted sequence space can be tested for the desired activity.[37] Based on extensive (preferably structural) knowledge about the protein of interest and its mechanism of thermostability, *rational design* involves site-directed mutagenesis of residues that destabilize the protein.[138-141] Rational design requires an understanding of important factors of protein thermostability identified by studying structures and sequences of thermophilic proteins and comparing those against mesophilic proteins (see Chapter 4). However, such detailed knowledge is not always available for each protein.[35] Data-driven approaches are being pursued as a promising alternative to directed evolution and rational design. In data-driven approaches, the library size for mutagenesis is reduced by suggesting interesting residue positions that, when mutated, would lead to a more thermostable protein.[36] Several successful data-driven approaches for protein thermostability have been reported in literature.[62,142,143] Apart from protein engineering and medium engineering, immobilization of protein on a solid support is also pursued to impart thermostability to a protein. Immobilization is achieved either via covalent or ionic interactions with the solid support or via cross-linking or

entrapment.[130] Modes of protein thermostabilization are reviewed in detail by A. S. Bommarius and M. F. Paye.[130]

2.6 Computational approaches for protein thermostabilization

Experimental testing of mutants of a protein for a desired thermostability in directed evolution[34] is limited by the large number of theoretically possible mutations a protein can harbor: The number of possible single point site-saturation mutants is $n\times19$ for a protein sequence of length n. The number exponentially grows when one considers mutations at multiple sites together: For a double mutant, the number of possible mutants would be $n\times(n-1)\times19^2/2$ = 7,183,900 for a protein of 200 residues. This necessitates data-driven approaches that restrict the library size for experimental testing by employing all available structural and sequence information of the target protein and computational models.[36] Data-driven approaches differ from the rational protein design in that it does not require a deep understanding of the forces contributing to the free energy of unfolding and the mechanism of thermostabilization of the target protein.[35] Computational data-driven approaches that can predict the effect of a mutation on the stability of a protein or to suggest weak spots mutating which would more likely improve a protein's thermostability are frequently employed to assist protein engineering projects. Computational approaches aimed at improving thermostability can be broadly divided into structure-based and sequence-based approaches; selected approaches are briefly described here.

2.6.1 Structure-based computational approaches

Several structure-based computational approaches rely upon calculating the change in ΔG_{FU} upon mutations using energy functions. Six of such methods, CC/PBSA,[144] EGAD,[145] FoldX,[146] I-mutant2.0,[147] Rosseta,[148] and Hunter[149] were surveyed by V. Potapov *et al.*[150] These six methods use three different classes of energy functions: I) Physical-based potentials, which are based on the analyses of the forces between atoms (CC/PBSA and EGAD).[151] II) Knowledge-based potentials, which rely on statistical analysis of geometric properties extracted from a large set of protein structures (FoldX and Hunter).[152] III) Support vector machine-based regression, which is a supervised machine learning technique (I-Mutant2.0).[153] Rosetta uses a hybrid physical-based and knowledge-based potential.[148] EGAD energy function employs OPLS-AA force field[154] along with generalized Born continuum model for polar solvation energy,[155] and a solvent-accessible surface area-dependent term for hydrophobic effect. CC/PBSA energy function differs from that of EGAD

in that it uses GROMOS96 force field[156] and Poisson-Boltzmann equation for calculating polar solvation energy[157] apart from using an entropy term based on quasi-harmonic approximation.[158] V. Potapov *et al* used a large data set of 2156 mutations that were modeled by these six methods independently, although all methods employ a "fixed backbone" approach while modeling a mutation. A single structure of the mutated variant was used for calculating the change in ΔG_{FU} in all methods except for CC/PBSA in which structural ensemble of mutated variants was built using Concoord.[159] Finally, a different sub-set of mutations were used evaluating each method due to three reasons: I) For each method, the mutations that were originally used for training were discarded; II) Disallowed mutations e.g., mutations to or from Cys, Gly or Pro are not allowed in EGAD owing to the fixed backbone approach; III) Modeled structures with steric clashes were discarded in EGAD and Rosetta. The authors found that the correlation coefficients (r) between experimental and predicted changes in ΔG_{FU} for the six methods ranged between 0.26 (Rosetta) and 0.59 (EGAD). However, the methods were found to be *good on average*: the correlation coefficients increase up to 0.96 when the change in ΔG_{FU} is compared in bins of 1 kcal mol^{-1}. In summary, these methods could not correctly predict the stability changes in detail; however, they were able to predict a correct trend. A. Fischer *et al.* introduced a computational saturation mutagenesis protocol (CoSM) for predicting a stability change upon a single point mutation at a protein interface using an artificial $(\beta\alpha)_8$-barrel protein as a test case.[160] The unfolded structures of the WT and the mutants were approximated by virtually splitting the $(\beta\alpha)_8$-barrel into two $(\beta\alpha)_4$-half barrels, and energies of the folded and unfolded forms were calculated using the MAB force field.[161] For fourteen single point mutants at residue V234 covering all amino acid classes, the authors obtained a good correlation between changes in ΔG_{FU} and the calculated stability differences ($R^2 = 0.85$).[160] However, the method has a shortcoming when it comes to smaller proteins or proteins with a complicated topology that cannot be virtually cut into two stable parts. At a distinction with methods based on calculation of changes in stability upon mutation, A. V. Gribenko *et al.* focused on improving surface charge-interactions for improving protein thermostability.[162] In their approach, surface charge distribution was optimized using the Tanford–Kirkwood surface accessibility genetic algorithm (TKSA-GA).[163-165] The authors demonstrated that the sequences of human acylphosphatase and cell-division cycle factor 42 GTPase predicted to have the largest increase in the favorable energy of charge-charge interactions also showed increased thermostability without compromising activity.[162] Since their method only

considers charge-charge optimization of surface residues, mutations that thermostabilize a protein by other mechanisms cannot be identified.

2.6.2 Sequence-based computational approaches

The most common sequence-based approach for thermostabilization involves substituting amino acids in a mesophilic protein sequence with amino acids found at the same positions in a (hyper)thermophilic homolog.[48,49] This approach follows the hypothesis that (hyper)thermophilic proteins have evolved from mesophilic proteins.* Hence, mutations in the (hyper)thermophilic proteins with respect to the mesophilic proteins confer thermo-tolerance to the (hyper)thermophilic proteins. However, this approach can only be used when a sequence of a thermophilic homolog is available. This limitation is overcome by a consensus sequence approach that does not necessarily require sequence(s) of (hyper)thermophilic homolog(s) of the target protein. Here, using a MSA of several mesophilic sequences, a new consensus sequence of the target protein is generated keeping consensus amino acid at every sequence position. This approach follows the idea that every amino acid of a protein contributes to some extent to the protein's overall stability, and therefore, optimizing a considerable number of residues together would be advantageous. For optimizing the sequence for improved thermostability, residues that have proven their fitness in several homologous proteins, *i.e.*, consensus amino acids are considered for each sequence position. This is based on the hypothesis that consensus amino acids contribute more to the stability of a protein than the non-consensus amino acids. Employing the consensus sequence approach, highly thermostable phytase variants were developed by M. Lehmann *et al.*[50-52]

E. Bae *et al.* developed an approach termed "improved configurational entropy" (ICE) for designing thermostable sequences of a protein using a MSA; the fitness of the designed sequences is then assessed using structural information of a large dataset of proteins.[167] The approach requires two or more homologous sequences to the target protein as an input. In ICE, all possible sequences are created by varying the variable sequence positions with the substitutions observed at that position in an MSA. Finally, sequences with the lowest average local structural entropy[168] calculated for all tetramers (four consecutive amino acids in the sequence) based on their frequency of occurrence in different secondary structure elements are considered thermostable. Using ICE, two thermostable variants of adenylate kinase (AK)

* The hypothesis that (hyper)thermophilic proteins have evolved from mesophilic proteins does not hold for all proteins; there are examples of organisms that originated directly as hyperthermophiles, e.g., *Pyrococcus furiosus*.[166]

that showed 11.6ºC and 12.5ºC increase in T_m were produced with a sequence alignment of a mesophilic and a psychrophilic AK.[167]

2.7 Constraint Network Analysis

In this work, I used and further developed a structure-based computational approach termed CNA introduced by S. Radestock and H. Gohlke.[63,64] Note that one should distinguish between the "CNA approach", which is originally developed by S. Radestock and H. Gohlke[63,64] and improved by us (**Publication II**), and the "CNA software", which is implemented by us during this thesis work (**Publication III**). CNA is a graph-theory based rigidity analysis approach wherein a biomolecule is modeled as a network (graph) of atoms, represented as sites (vertices), and interactions between them, represented as constraints (edges). Loosely put, in CNA, thermal unfolding of biomolecular networks of atoms and interactions is performed by successively removing non-covalent interactions from the network in increasing order of their strengths. During the thermal unfolding simulation, local (residue level) and global rigidity indices are calculated. Using global rigidity indices, phase transition points T_p (melting points) and unfolding nuclei (structural weak spots that should improve thermostability upon mutation) are identified. At a distinction with the computational approaches described in section 2.6.1, effect of a mutation on a protein's stability is not estimated in terms of the change in ΔG_{FU} in CNA approach. Rather, in CNA approach, weak spots are predicted as the residues from where unfolding of a protein begins at the phase transition point. Development of the CNA approach was inspired by the works of D. J. Jacobs *et al*. (introduction of the graph theory-based method for protein flexibility prediction), and B. M. Hespenheide *et al*.[169] and A. J. Rader *et al*.[68] (thermal unfolding simulation of protein constraint networks).

The CNA approach has been successfully applied for discriminating mesophilic and thermophilic homologous protein pairs, identifying weak spot residues and linking flexibility and function of proteins.[63,64] Despite these successful applications of the CNA approach, advancements both in technical and methodological domains were required. To the former, software packages for carrying out thermal unfolding simulations of proteins and interactive analysis of the results were required. To the latter, improved modeling of hydrophobic interactions and an ensemble-based version of the CNA to reduce sensitivity to the input structure were needed. Finally, the CNA approach needed to be validated by prospective prediction of thermostable mutants and subsequent experimental testing.

The CNA approach has been described in detail by us elsewhere (**Publication 0**).[170] In the following, the theory and methods behind the CNA approach are briefly described; the text comes mainly from the ref.[170]

2.7.1 Introduction to rigidity theory

The quest to identify rigid and flexible regions in networks (graphs) of sites (vertices) and constraints (edges) dates back long. In 1864, Maxwell proposed an approximate method to calculate the number of floppy modes F in a d-dimensional *generic* network, *i.e.*, a network without any symmetries like collinear constraints.[171] The term "floppy modes" denotes (independent) internal degrees of freedom in which the sites of the network can move without violating any of the constraints. For a network with N sites lacking any constraint, $F = dN - d(d+1)/2$, with the subtrahend denoting the global degrees of freedom (overall translation and rotation) of the d-dimensional network. Each added constraint, if independent of all other constraints, removes one floppy mode. Thus, if all constraints in the network were independent, as assumed by Maxwell, the number of floppy modes (F_{mxw}) in a network with N_c constraints can be calculated by eq. (3).

$$F_{MXW} = dN - N_c - \frac{d(d+1)}{2} \tag{3}$$

Usually, this underestimates F because in reality not all constraints are independent: if a constraint is placed between two already mutually rigid sites, it does not decrease the number of floppy modes any further and, thus, is a redundant constraint. Taking into account the number of redundant constraints N_r then leads to eq. (4).

$$F = dN - (N_c - N_r) - \frac{d(d+1)}{2} \tag{4}$$

Incorporating a redundant constraint introduces stress in the network; network regions with such constraints are thus called *over-constrained* or *stressed*. In contrast, a region with fewer constraints than internal degrees of freedom is called *under-constrained.* Finally, in a region with as many independent constraints as internal degrees of freedom, $F = 0$; this region is called *isostatically rigid*.

In 1970, a theorem by G. Laman[172] had a major impact in that it allowed to precisely determine the degrees of freedom in a two-dimensional network, even in the presence of redundant constraints, by applying constraint counting to all subgraphs within the network.

As such, a generic two-dimensional network does not have a redundant constraint if and only if for all subgraphs of size $n \geq 2$, the number of constraints in the subgraph $N_{cs} \leq 2n - 3$. By applying Laman's theorem, a network can be decomposed into rigid regions and flexible links in between. This constraint counting can be extended to a certain subtype of three-dimensional networks with a molecule-like character, so-called "bond-bending networks" or "molecular frameworks".[173,174] In these networks, bond angles (distances between second-nearest neighbor sites) are constrained in addition to bond lengths (distances between first-nearest neighbor sites), which makes them particularly applicable to biomolecules.

For both the two-dimensional and three-dimensional bond-bending networks, combinatorial algorithms called *pebble games* were devised that allow determining network flexibility and rigidity according to eq. (4).[66,67,175] These algorithms have been implemented in ProFlex (http://www.bch.msu.edu/~kuhn/software/proflex) and in early versions of the FIRST (http://flexweb.asu.edu) software package. As an example, bond-bending networks of two molecules are depicted in Figure 1. In both networks, fixed bond lengths and angles are modeled as distance constraints between nearest and next-nearest neighbor atoms. Free rotation about the bond between atom 1 and atom 2 in molecule M1 results in one floppy mode and two rigid clusters of three atoms each (Figure 1 a-c). A double bond is modeled by placing an additional distance constraint between third-nearest neighbors (Figure 1e), which results in molecule M2 being a single rigid cluster (Figure 1d-f)

A more recent implementation of FIRST uses a *body-and-bar* representation of three-dimensional networks where every atom is considered as a rigid body having six degrees of freedom.[176] Any number of bars between one and six can be placed between two such atoms, and every such bar removes one degree of freedom. The number of floppy modes is then computed according to eq. (5).

$$F = 6N - N_{ibar} - 6 \tag{5}$$

Here, N_{ibar} represents the total number of independent bars in the network. In the *body-and-bar* network representation, covalent single bonds are modeled as five bars between two atoms leaving one degree of freedom, the dihedral rotation (Figure 1c). Double bonds are modeled with six bars locking the rotation (Figure 1f). Apart from algorithmic advantages over the bond-bending representation, the body-and-bar representation also has a methodological advantage that lies in the fact that constraints can be modeled semi-

quantitatively: strong bonds are modeled with more bars, whereas weaker bonds are modeled with fewer bars.[176]

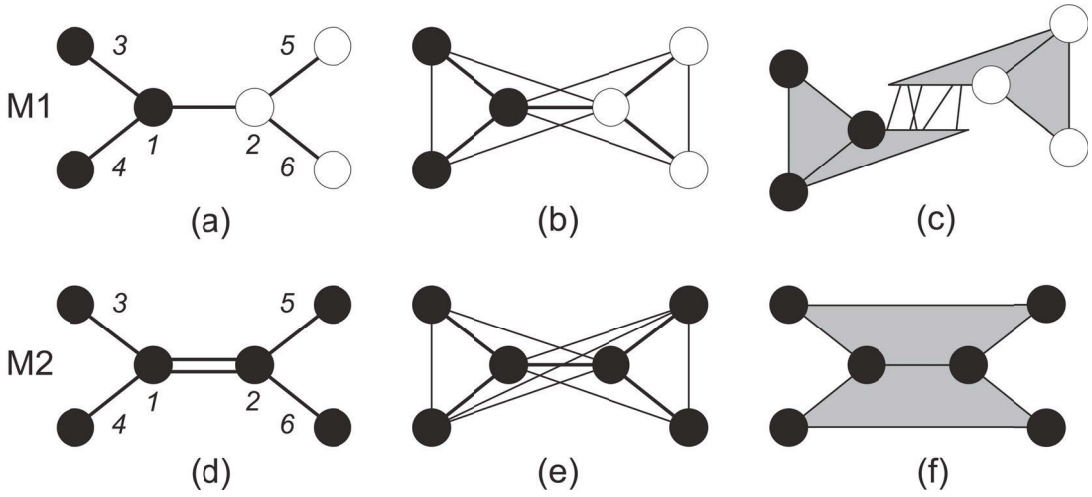


Figure 1. Network representations of molecules M1 (a) and M2 (d). In the *bond-bending* networks (b, e), the double bond in M2 is modeled by placing an additional constraint between atom 4 and 5. In the *body-and-bar* networks (c, f), the bond between atom 1 and 2 in M1 is modeled by 5 bars, whereas 6 bars are used in M2 for locking the rotation. The atom colors represent the rigid clusters they belong to: M1 has two rigid clusters and one flexible joint whereas all atoms of M2 belong to a single rigid cluster. Figure adapted from ref.[177]

2.7.2 Modeling biomolecules as constraint networks

Biomolecules can be effectively represented either as *bond-bending* or *body-and-bar* networks. Here, we describe how biomolecules can be represented by the latter representation that was used for all CNA calculations in the present thesis. The atoms of the biomolecules are modeled as bodies, while covalent and non-covalent bonds are modeled as bars. A covalent bond is generally modeled as five bars allowing for the dihedral rotation about it. Peptide and double bonds are modeled with six bars, disallowing any bond rotation. Considering that the mechanical rigidity of a biomolecule is largely determined by non-covalent interactions, one also needs to include hydrogen bonds, salt bridges, and hydrophobic interactions as constraints in the network. Stronger interactions such as hydrogen bonds (and salt bridges) and hydrophobic interactions are modeled as five bars and two bars, respectively.[176] Weaker interactions such as van der Waals (vdW) interactions are not modeled as constraints. **Figure 2** shows a network representation for hen egg white lysozyme (HEWL; PDB code: 2VB1), which is decomposed into rigid clusters and flexible joints (rigid cluster decomposition) by rigidity analysis using the FIRST software.

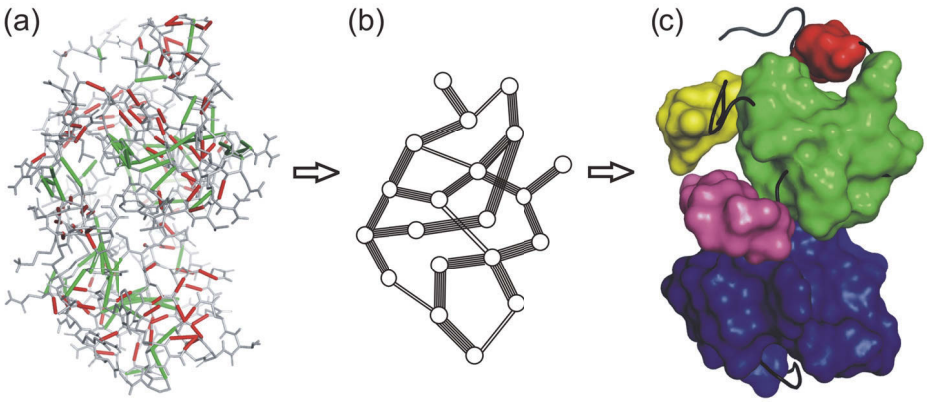


Figure 2. Workflow of rigidity analysis of biomolecules showing HEWL as an example. A PDB structure with added hydrogen atoms is used as input (a) from which a *body-and-bar* network is modeled (b). Covalent bonds are depicted in gray, hydrogen bonds in red, and hydrophobic tethers in green (a). Each bond is identified either as a part of rigid region or a flexible joint, resulting in a rigid cluster decomposition (c) where each rigid cluster has a unique color. Figure taken from ref.[170]

2.7.3 Simulating folded-unfolded transitions in biomolecules

By consecutively removing constraints from a network, one can simulate the melting of the network and identify a phase transition where the network switches from an overall rigid state to a floppy one. Phrased differently, at this so-called rigidity percolation threshold, the network loses its ability to transmit stress, *i.e.*, rigidity ceases to percolate through the network. Cross-linked covalently bonded three-dimensional network glasses have been thoroughly studied in that sense both computationally and experimentally.[178-180] It has been observed that the phase transition for network glasses is continuous or of second order.

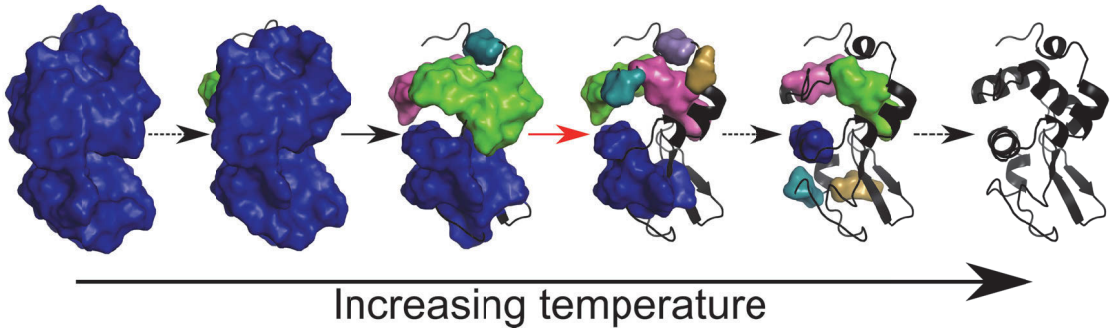


Figure 3. Rigid cluster decompositions of HEWL networks during the thermal unfolding simulation. Atoms belonging to a rigid cluster are depicted as uniquely colored body with the largest cluster being shown in blue. The giant rigid cluster at a low temperature breaks down into smaller clusters and flexible regions, finally leading to a completely flexible network at the highest end of the temperature scale. The biologically relevant phase transition point when a largely rigid network switches to a largely floppy one is marked with a red arrow. Dashed arrows indicate the presence of intermediate networks between the two network states, whereas a continuous arrow indicates a direct transition from one network state to the other during the thermal unfolding simulation.

However, it has also been found that the phase transition can become first order for self-organized networks where locally stressed regions or small rings of bonds are suppressed.[179] Biomolecular networks can be considered similar to network glasses, and the melting of the network can be realized by consecutively removing non-covalent bonds, which is equivalent to a thermal unfolding of the biomolecule. However, the percolation behavior of protein networks is usually more complex, and multiple transitions can be observed (Figure 3). This is due to the fact that protein structures are modular because they are assembled from secondary structure elements, subdomains, and domains. These modules often spontaneously break away from the largest rigid cluster as a whole giving rise to multiple transitions.

In the CNA approach, thermal unfolding simulations are carried out by sequentially removing non-covalent constraints from the initial network representation generating new network states, and subsequently rigidity analyses on all such newly generated networks are performed.[63,64,68,169] That is, for a given network state s = f(T) with hydrogen bond energy cutoff $E_{cut,s}$, hydrogen bonds (including salt bridges) with an energy $E_{HB} > E_{cut,s}$ are removed from the network.[181] This follows the idea that stronger hydrogen bonds will break at higher temperatures than weaker ones. To convert the original, geometry-based hydrogen bond energy scale E_{HB}[181] into a temperature scale T, S. Radestock and H. Gohlke proposed a simple linear fit (eq. (6)) by comparing computed phase transition temperatures for pairs of homologous mesophilic and thermophilic proteins with experimental melting temperatures.[63,64] The temperatures should be considered relative values only because the absolute values may depend on the size and architecture of the analyzed protein. In the original version of the CNA approach,[63,64] hydrophobic interactions were not modeled in a temperature-dependent manner; rather, the number of hydrophobic contacts was kept constant during the thermal unfolding. As such, a hydrophobic constraint was added between carbon or sulfur atoms, if the distance between these atoms is less than the sum of their vdW radii (C: 1.7 Å and S: 1.8 Å), plus a distance cut off D_{cut}. This was done because the strength of hydrophobic interactions remains constant or even increases with increasing temperature.[182] This presented an opportunity to improve CNA by modeling the increase in the strength of hydrophobic interactions with a rise in temperature;[183,184] this was taken up by us in the present thesis (**Publication II**).

$$T = \frac{-20K}{kcal * mol^{-1}} E_{cut,s} + 300K \qquad (6)$$

2.7.4 CNA on ensembles of network topologies

In principle, CNA can be performed on a single three-dimensional structure of a biomolecule. However, different conformations of a protein structure can lead to different results of the rigidity analysis as observed in previous studies.[185,186] This sensitivity arises from the facts that I) proteins are generally marginally stable[187] and II) different protein conformations can lead to different numbers of constraints being included based on geometric criteria. Consequently, as the protein network is already close to the rigidity percolation threshold (due to I), a few constraints more or less (due to II) can result in the network being largely rigid or already floppy. Apart from pointing out that the results from rigidity analyses are highly sensitive to the input structure used, these studies[185,186] also presented ways to tackle this by using information from a structural ensemble in rigidity analyses. We extended the concept of applying molecular dynamics (MD) simulation-based ensembles in rigidity analyses by H. Gohlke *et al.*[185] to CNA. In ensemble-based CNA developed by us, multiple conformations of a protein extracted from a MD simulation-derived trajectory (or any other source) are used as input, and CNA is run on each conformation in the ensemble (**Publication II**). Results obtained from different conformations are then averaged over the entire ensemble. This approach has the advantage that CNA is based on a thermodynamic *ensemble of conformations*. However, a computationally expensive MD simulation is required to generate the input. As an alternative, C. Pfleger and H. Gohlke have developed an approach termed ENT^{FNC} in which an ensemble of network topologies is generated from a single input structure by modulating noncovalent constraints using fuzzy constraint definitions.[188] As such, in the original implementation of ENT^{FNC}, flickering of hydrogen bonds was achieved by deriving probabilities from MD simulations with which a hydrogen bond persists in an ensemble of network topologies. Furthermore, energies of hydrogen bonds E_{HB} in the new network topologies were reset by adding Gaussian white noise to their original energies. As to hydrophobic contacts, a probability distribution $p(d_{ij})$ for including a hydrophobic constraint between a pair of carbon or sulfur atoms in a network topology is derived using a Gaussian function with a squared distance dependency (eq. (7)).

$$p(d_{ij}) = e^{-\frac{1}{2}\left(\frac{\left(d_{ij}-d_{vdW}\right)^2}{D_{cut}^2}\right)^2} \tag{7}$$

Here, d_{ij} is the distance between the hydrophobic atoms i and j, d_{vdw} is the sum of vdW radii of the two atoms, and D_{cut} determines the full width at half maximum of the Gaussian. All

parameters for modulating noncovalent constraints in the network were derived from MD simulations of HEWL structures.[188]

2.7.5 Indices to characterize flexibility and rigidity

The CNA software package developed by us (**Publication III**) calculates several local and global flexibility/rigidity indices developed and formalized mainly by C. Pfleger and S. Radestock.[189] All of these indices share the common feature that they are derived by analyzing a thermal unfolding simulation of a constraint network. A detailed description of these indices can be found in the original publication by the authors.[189]

To describe global rigidity percolation of a network, the *microstructure* of the network, *i.e.*, properties of the set of clusters generated during the thermal unfolding simulation can be analyzed. CNA calculates the floppy mode density Φ, the rigidity order parameter P_∞, the cluster configuration entropy H (eq. (8)), and the mean rigid cluster size S.[189] Among these global indices, H has been used to predict protein thermostability before.[63,64]

H has been introduced by C. Andraud *et al*. as a morphological descriptor for heterogeneous materials,[190] and has been adapted from Shannon's information theory; thus, it is a measure of the degree of disorder in the realization of a given state.

$$H = -\sum_s w_s \ln w_s \tag{8}$$

It is defined as a function of the probability (w_s, eq. (9)) that an atom is part of a rigid cluster of size s (s-cluster). Where

$$w_s = \frac{s^k n_s}{\sum_s s^k n_s} \tag{9}$$

with n_s being

$$n_s = \frac{\text{Number of clusters of size } s}{N} \tag{10}$$

and N being the total number of atoms. For H_{type1}, which corresponds to the original definition by C. Andraud *et al.*, a linear cluster size (k = 1) is used. By the modified version H_{type2} that considers a quadratic cluster size (k = 2), later phase transition points during the thermal unfolding simulation are potentiated and, hence, are preferentially identified. These

later transitions have been found to be related to protein thermostability.[63] As long as the giant cluster dominates the network, *H* is low because of the limited number of possible ways to configure a network with a very large cluster. At the rigidity percolation threshold, *H* jumps as the network is now in a partially flexible state with many ways to configure a network consisting of (many) small clusters. Biologically relevant phase transition points on *H vs. T* curves are identified as the maximum of the differences in the asymptote pairs of a fitted double sigmoid function (Figure 4a). In the case of an ensemble-based CNA run, the median phase transition point is calculated (Figure 4b). Using the phase transition point, unfolding nuclei or weak spots are identified as the residues that belong to the giant rigid cluster until the phase transition point and become flexible afterwards (Figure 4c). From the point of view of protein stability, unfolding of the giant rigid cluster at the phase transition point begins from these residues. Similarly, for an ensemble-based CNA, the probability of each residue being a weak spot over the entire ensemble is calculated (Figure 4d).

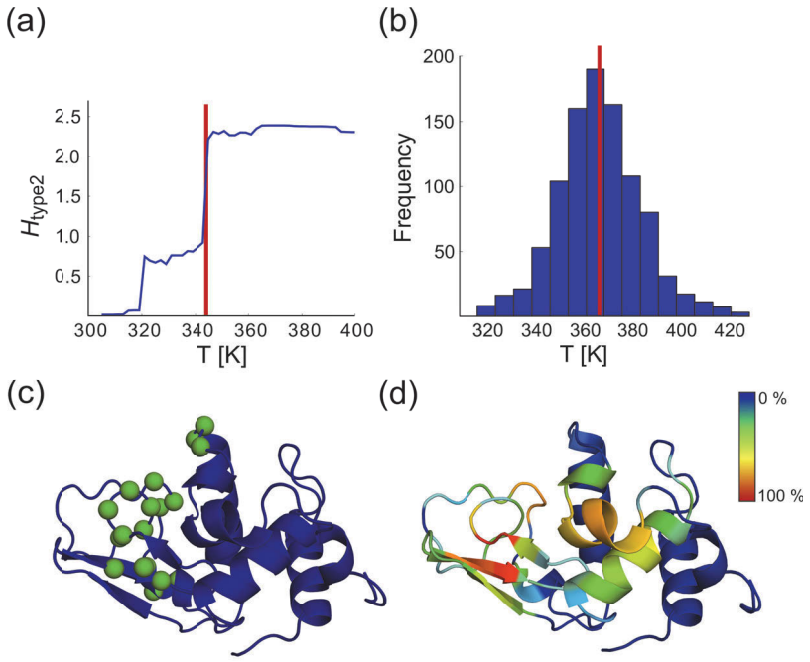


Figure 4. Results from CNA on HEWL (PDB-ID: 3LZT) Cluster configuration entropy H_{type2} derived from the single network topology (a). The entropy is plotted as a function of the temperature. The phase transition automatically identified by CNA is marked by the red vertical line. Frequency distribution of phase transition points identified from analyzing an ensemble of network topologies (b). The median is marked with a red vertical line. The green spheres highlight the identified weak spots in the single network topology (c). For depicting the probability of being a weak spot over the ensemble of network topologies, each residue is colored according to a color scale ranging from blue (low probability) to red (high probability) (d). Figure adapted from ref.[191]

Local rigidity indices characterize the network rigidity down to the bond level.[189] The *rigidity index* r_i is derived for each covalent bond in the network by monitoring the hydrogen bond energy cutoff $E_{cut,s}$ during a thermal unfolding simulation at which the bond switches from being rigid to flexible (Figure 5a and b). Phrased differently, r_i monitors when a bond segregates from *any* rigid cluster. For a C_α atom-based representation of a protein structure, the average of the r_i values of the two backbone bonds is taken. Another local rigidity index, the percolation index p_i monitors the percolation behavior (*i.e.*, the loss of rigidity when diluting the constraint network) of a biomolecule allowing identification of the hierarchical break-down of the giant percolating cluster during a thermal unfolding simulation. The giant percolating cluster is the largest rigid cluster present at the highest E_{cut} value (*i.e.*, at the lowest temperature at the beginning of a thermal unfolding simulation) with all constraints in place. More technically, p_i monitors the E_{cut} at which a bond segregates from the giant percolating cluster during a thermal unfolding simulation. For a C_α atom-based representation, the lower of the p_i values of the two backbone bonds is considered.

A *Stability map* rc_{ij} is a two-dimensional generalization of the rigidity index (Figure 5c and d). To derive a stability map, "rigid contacts" between two residues, represented by their Cα atoms, are identified. A rigid contact exists if two residues belong to the same rigid cluster. During a thermal unfolding simulation, stability maps are then constructed in that, for each residue pair, $E_{cut,s}$ is identified at which a rigid contact between the two residues is lost. That way, a contact's stability relates to the microscopic stability in the network and, combined, the microscopic stabilities of all residue-residue contacts result in a stability map. Thus, stability maps denote the distribution of rigidity and flexibility within the system, they identify regions that are flexibly or rigidly correlated across the structure, and they provide information on *how these properties change with temperature*. A filtered "neighbor stability map", where stability values of residue pairs separated by more than 5 Å are masked, provides useful information about the stability of short range contacts in a biomolecule (lower triangles in Figure 5c and d). For an ensemble based CNA, average local indices values are calculated (Figure 5b and d).

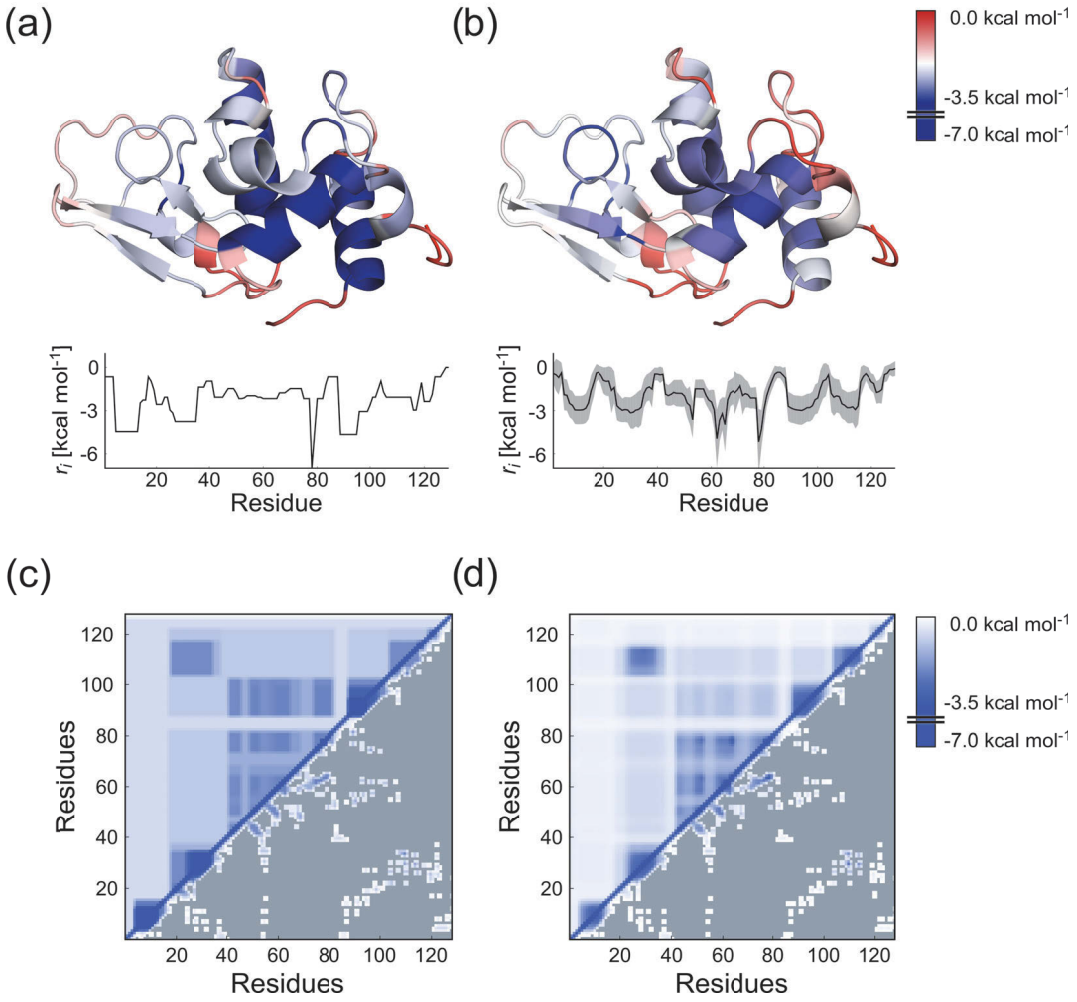


Figure 5. The rigidity index *ri* determined by analyzing the single network topology (a) and the ensemble of network topologies (b) is plotted against a residue identifier and color-coded onto the structure (range of the color code: red (flexible) to blue (rigid)). In addition, the plot in (b) shows the standard deviation as a gray area. Stability maps (upper triangle) and neighbor stability maps (lower triangle) determined by analyzing the single network topology (c) and an ensemble of network topologies (d). The color depicts how stably two residues are connected and ranges from light white (low stability) to blue (high stability). Gray areas in the neighbor stability map are displayed when residues are more than 5 Å away from each other. All results are obtained for HEWL (3LZT). Figure adapted from ref.[191]

2.8 Approaches related to CNA

As mentioned before, CNA uses FIRST[65] as its core engine for calculating bond-level network rigidity. Although a bond dilution procedure for simulating temperature rise is implemented in FIRST, it does neither provide a detailed thermal unfolding simulation nor does it calculate local and global rigidity indices that CNA provides. In a similar way as FIRST, KINARI (KINematic And RIgidity analysis of proteins) analyzes the rigidity and flexibility of a protein by applying the pebble game algorithm.[192] KINARI uses HBPLUS[193]

or Bndlst (http://kinemage.biochem.duke.edu/software/utilities.php) for identifying hydrogen bonds and similar to FIRST a geometry-based empirical function[181] for calculating hydrogen bond energies. The strength of hydrophobic interactions is measured in terms of either the distance between hydrophobic atoms (C or S), as in FIRST,[65] or pairwise vdW energies calculated using a Lennard-Jones potential with parameters taken from the AMBER parm99 forcefield.[194,195] . KINARI calculates the rigidity of a protein in a state with specified hydrogen bond energy cutoff and either the distance or energy cutoff for hydrophobic contacts. However, unlike the CNA, KINARI does not allow thermal unfolding simulations and calculations of rigidity indices. Using KINARI-based rigidity analysis, F. Jagodzinski *et al.* developed a web service termed KINARI-Mutagen for predicting changes in rigidity of a protein upon mutation.[196] KINARI-mutagen models the mutation of a residue to Gly (termed as *excision*) by removing hydrogen bonds and hydrophobic interactions due to side-chain of the residue from the protein's molecular model. Rigidity analyses of the excised structure and the WT structure are subsequently carried out and compared. The residues that affect the rigidity of the protein upon excision mutation are then inferred to be critical for the stability of the protein. As such, the authors predict an excision mutation as destabilizing if the rigidity analysis reveals that the largest rigid cluster in the mutated structure is smaller than the same for the WT. Out of 48 destabilizing single-point substitutions to Gly among 14 proteins extracted from the ProTherm database,[197] 22 mutations were correctly identified as such. However, the KINARI-Mutagen fails to predict destabilization in cases where no hydrogen bonds or hydrophobic contacts are detected for the mutated residue (13 cases), mutations occurring on solvent-exposed residues (8 cases), and mutations on hydrophobic residues (Val, Leu, Met, and Phe) for which too few hydrophobic contacts were detected by KINARI than expected for these residues in most protein cores (4 cases). Apart from these limitation, we note that the approach only predicts destabilizing mutations and that too when a residue is substituted by Gly.

Another rigidity analysis approach, the Distance Constraint Model (DCM), has been developed by D. J. Jacobs *et al.*[198] in which an ensemble of constraint topologies is generated by considering mean-field probabilities of hydrogen bonds and torsion constraints in a Monte Carlo sampling. Using the Pebble game algorithm,[66,67] flexibility of each network in the ensemble is then characterized. Finally, the equilibrium properties are characterized by averaging over the thermodynamic ensemble. An accurate estimation of the free energy is required for the ensemble averaging, which is done using a free energy decomposition

scheme[199,200] considering nonadditivity of conformational entropy components.[201,202] Here, the conformational entropy of a protein is estimated by considering the independence or redundancy of each constraint. As such, placement of a constraint in an already rigid region (i. e., a redundant constraint) does not lead to an entropy cost. Accordingly, a lowest upper bound estimate for the conformational entropy of the system is given as the sum over entropy contributions from all non-redundant constraints. Constraints are added in the network in decreasing order of their strengths (a stronger constraint that reduces entropy the most is added before a weaker constraint), and rigidity analysis is recursively applied to determine whether or not the constraint being added is redundant. Protein thermodynamics is then described using the free energy landscape defined by two order parameters given by the number of native hydrogen bonds, and the number of native backbone and side-chain torsion angles within a given macrostate.[200] Using mean field theory and Monte Carlo sampling for each macrostate, the partition function is then calculated from which all thermodynamic properties of interest, including C_p, can be calculated numerically. Because empirical parameters with the free energy function accounting for protein size, architecture, and solvent effect are not known, DCM requires knowledge of experimentally determined C_p curves for a protein-specific parameterization of the model.[200,203] Owing to the fact that several approximations were made, the approach is termed *minimal* DCM (mDCM).[200]

Using the mDCM, C_p curves were accurately reproduced for ubiquitin at five different pH conditions and for histidine binding protein in the apo and holo forms.[200] In another study, using the mDCM, in agreement with known experimental data, the authors

identified residues of an orthologous Ribonuclease H pair that are important for stability and function.[203] Furthermore, different enthalpy-entropy compensation mechanisms with respect to the unfolding process are determined that lead to globally similar stability and flexibility profiles between the pair at their respective T_m.[203] Recently, the mDCM was applied to study allostery in three bacterial CheY orthologs.[204] The authors demonstrated that residues likely to be involved in the transmission of allosteric information are both conserved and variable across the three CheY orthologs studied. Moreover, they predicted the strongest allosteric site to be located on the β4/α4 loop, which is known as a critical link in the intramolecular communication within CheY. In yet another study employing mDCM, D. Verma *et al.* demonstrated that mutation in human c-type lysozyme results in frequent, large, and sometimes long-ranged changes in the flexibility.[205] The authors observed that the frequency,

scale and complexity of the change in the flexibility are consistent with multiple NMR characterizations of mutant dynamics in a variety of proteins, including lysozyme. Despite these successful applications, the general applicability of the mDCM is limited by the requirement of C_p curves of the system being investigated.

Recently, L. C. González *et al.* have developed an ensemble-based rigidity analysis approach termed virtual pebble game (VPG), which similar to the ENT^{FNC} approach uses a single conformation of the system and does not require actual sampling of conformations.[206] Unlike ENT^{FNC}, an average constraint network over a Monte Carlo-derived ensemble is used in that the number of bars between all pair of bodies in the network (see section 2.7.2) is averaged over all network conformations in the ensemble. VPG then counts constraints and degrees of freedom to real numbers, allowing for fractional degrees of freedom. On a non-redundant dataset of 272 protein structures, rigidity characteristics computed by VPG were comparable with the ensemble-averaged characteristics computed by the regular pebble game.[206] We note that a drawback of the VPG approach is that it is less accurate at the rigidity percolation threshold rendering it less suitable for identifying T_p values during thermal unfolding simulations. This is because all hydrogen bonds are treated equally here disregarding their energies and, accordingly, the largest fluctuation in network topology occurs at rigidity percolation thresholds leading to the greatest differences between the regular pebble game and VPG results.

3 SCOPE OF THE THESIS

The main goal of this thesis was to develop knowledge-driven computational approaches for predicting thermostabilizing mutations on a protein. For a successful prediction of thermostabilizing mutations on a protein, computational approaches require knowledge of the factors that determine protein thermostability. From the literature survey, it became clear that a multitude of such factors are responsible for protein thermostability (see sections 2.3). However, it was not known which of these factors of protein thermostability are the most significant. Hence, I set out to identify the most significant determinants of protein thermostability; results of this are reported in Chapter 4 (**Publication I**). An approach for predicting weak spots, residues that are more likely to increase a protein's thermostability upon mutation, was developed based on the most significant determinant identified in this study. From Chapter 4, it became clear that hydrophobic interactions are the most significant determinants of protein thermostability; however, they were not modeled in a temperature-dependent manner in the Constraint Network Analysis (CNA) approach (see section 2.7.3). Furthermore, another limitation of CNA lies in that different conformations of a protein structure can lead to different results due to its sensitivity to minor changes in the input structure (see section 2.7.4). Therefore, I set out my next goal to improve CNA with respect to these two limitations by incorporating temperature-dependent hydrophobic interaction modeling as well developing an ensemble-based CNA. These results are reported in Chapter 5 (**Publication II**). Since a software for carrying out CNA along with automatic calculation of rigidity indices, phase transition points, and weak spots (see section 2.7) was not available, we set out our next goal to develop a command-line software packages for CNA. These results are reported in Chapter 6 (**Publication III**). In the backdrop of the fact that the results from CNA are highly information-rich in that it computes several local and global rigidity indices, rigid cluster decompositions of protein structures, phase transition points, and weak spots (see section 2.7.5), our next goal was to develop a web service and a GUI, which allow running CNA and extensive analysis of results in a user-friendly manner. The outcomes of this goal are reported in Chapter 6 (**Publication IV and V**). Next, I used CNA to study the relation between structural rigidity and protein thermostability. Results of this study on thermodynamically and kinetically thermostabilized variants of *Bs*LipA are reported in Chapter 7 (**Publication VI**). Finally, with the finding from **Publication VI** that CNA is able to correctly predict thermodynamic thermostability of closely related variants, I set out to validate CNA by a prospective application aimed at improving thermostability. Results of this on lipase A from *Bacillus subtilis* (*Bs*LipA) are reported in Chapter 8 (**Publication VII**).

4 PRIORITIZING FACTORS INFLUENCING PROTEIN THERMOSTABILITY (PUBLICATION I)*

4.1 Background

Identification of (the) dominant determinant(s) of protein thermostability is an important aspect of research in biotechnology because knowledge of this will facilitate rational design and data-driven approaches aimed at improving protein thermostability. Comparisons of pairs of meso- and (hyper)thermophilic proteins have revealed several such determinants, including improved hydrogen bonding,[40] ion pair and salt bridge networks,[39] better hydrophobic packing,[41] shortening of loops,[42] higher secondary structure content,[43] and increased rigidity of a protein (see section 2.3).[63,64,69,111] This indicates that a multitude of factors makes a protein more stable. This raises a pertinent question: which of these determinants of protein thermostability are the most significant. The answer to this question would help focusing on (the) most important determinant(s) when predicting protein thermostability and weak spots (as done in Chapter 5).

4.2 Methods

In **Publication I**, rather than analyzing thermostability in terms of structural or geometric properties, we focused on energetic factors with the aim to identify (the) most significant determinant(s) of protein thermostability. For a large dataset of 132 pairs of mesophilic/thermophilic and 149 pairs of mesophilic/hyperthermophilic homologous protomers, we calculated several residue-wise interaction energy components including electrostatic, vdW, hydrogen bond, and hydrophobic interaction energies. Initially, probability density functions (PDFs) of these energy components were compared on a global level to identify which of the interactions show on average a more favorable residue-wise energy in (hyper)thermophilic proteins. Next, we investigated (differences in) the spatial distribution of residue-wise interaction energies in pairs of mesophilic/(hyper)thermophilic protomers. Our hypothesis is that a larger cluster of residues with lower energies than a given energy cutoff exists in (hyper)thermophilic proteins than their mesophilic homologs. To test our hypothesis, we performed a hierarchical clustering of residues with respect to different interaction energy components, such that all neighboring residues with an energy component lower than a cutoff E_C for the respective clustering level are grouped in the same cluster.

* Part of this work was done at Department of Modeling and Formulation Research, BASF, Ludwigshafen, Germany during an internship under supervision of Dr. H. Wolfgang Höffken.

Thus, clusters grow in size as E_C increases (*i.e.*, the energy component becomes less favorable). For each E_C, the fraction of residues that is part of the largest cluster (F_{LC}) was calculated. According to our hypothesis, a correct distinction of a (hyper)thermophilic protein from a mesophilic protein would be the presence of a larger cluster of residues in the (hyper)thermophilic protein compared to its mesophilic counterpart at identical E_C values. Furthermore, based on this clustering approach, weak spots on a protein were identified as residues that have high energies and are spatially close to a large cluster of residues.

4.3 Results

Comparison of the PDFs of residue-wise electrostatic, vdW, hydrogen bond, and hydrophobic energies differ between mesophilic and (hyper)thermophilic protomers. (Hyper)thermophilic protomers showed a higher probability densities at more negative (*i.e.*, more favorable) energies except in the case of vdW energies for mesophilic/hyperthermophilic pairs. The observed differences are statistically significant ($p < 0.05$ for the hypothesis of equality) except for hydrogen bond energies in the case of mesophilic/hyperthermophilic protomers. According to the *p*-values, the most significant difference between PDFs of mesophilic/thermophilic and mesophilic/hyperthermophilic protomers is found in the case of residue-wise hydrophobic energies ($p < 0.0001$ for both cases). This was also reflected in the magnitudes of the median differences in the hydrophobic energies: On average, a residue in a thermophilic (hyperthermophilic) protomer has a hydrophobic interaction energy that is more favorable by 0.82 (1.27) kcal mol^{-1} than that of a residue in a mesophilic protomer. Next, clustering based on residue-wise hydrophobic interaction energies correctly discriminated 83% of the pairs of mesophilic/thermophilic protomers and 76% of the pairs of mesophilic/hyperthermophilic protomers (Figure 6). These discrimination accuracies are significantly ($p < 0.001$) different from the one of a random discrimination (50% correct discrimination). In contrast, when we used distance-based clustering of hydrophobic residues (Ala, Cys, Ile, Leu, Met, Phe, Trp, and Val) as a discriminator, only 53% and 62% pairs of mesophilic/thermophilic and mesophilic/hyperthermophilic protomers, respectively, could be successfully discriminated.

Finally, we used this information for prediction of weak spots that should improve a protein's thermostability upon mutation. As such, residues that have unfavorable (high) hydrophobic energies and are spatially close to a large cluster of residues comprising 50% of residues ($F_{LC} = 0.5$) were considered weak spots. We chose $F_{LC} = 0.5$ because we visually observed

that the cluster at this point represents the "hydrophobic core", and residues forming this should not be mutated. The weak spot prediction was validated using *E. coli* dihydrofolate reductase (DHFR) for which three out of the eight thermostabilizing residues were correctly predicted as weak spots by our approach. In turn, twelve out of the fourteen destabilizing residues were correctly predicted as being part of the "hydrophobic core" and, hence, not predicted as weak spots. This yields a classification accuracy of almost 70%, with our approach being more accurate in identifying non-weak spots (specificity = 85%) than weak spots (sensitivity = 38%).

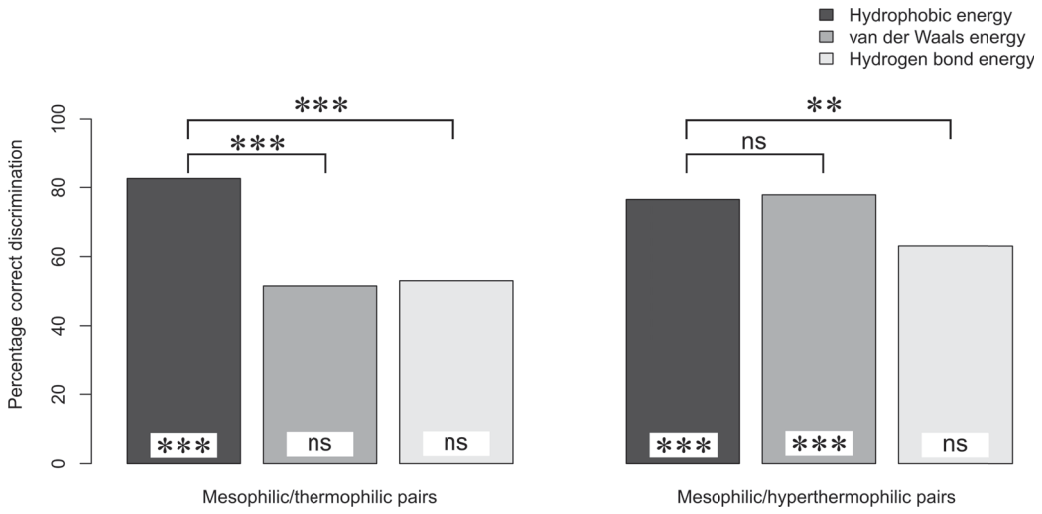


Figure 6. Discrimination accuracy between mesophilic and (hyper)thermophilic protomers based on clusters of residues with good residue-wise energy components. The statistical significance of the differences in discrimination accuracies is computed by a bootstrap hypothesis test of equality generating 10000 bootstrap samples; the significance levels are marked by *** ($p < 0.001$), ** ($p < 0.01$), and ns ($p > 0.05$).). A significance mark at the bottom for a column represents difference in the discrimination accuracy with a random discrimination (50%) whereas a mark pointing to two bars represents difference in the discrimination accuracies by clustering using the two interaction energies. Figure taken from ref.[207]

4.4 Conclusions and significance

- When considering energetic factors, we identify hydrophobic interactions to be the most significant factor for protein thermostability. In contrast to previous findings, we observe that not the *size* but rather the *quality* (energy) of hydrophobic interaction in a cluster determines the thermostability.
- The finding that hydrophobic interactions play the most significant role in proteins' thermostability prompted us to improve CNA such that, like hydrogen bonds, hydrophobic interactions are also treated in a temperature-dependent manner (see Chapter 5)

- Using clustering based on residue-wise hydrophobic energy components, we correctly discriminated 83% (76%) of the pairs of mesophilic/thermophilic (mesophilic/hyperthermophilic) protomers.
- For a prospective application, a method of identifying weak spots was developed that yielded a classification accuracy of almost 70% on the test set of 22 mutants of *E. coli* DHFR.
- All in all, our approach highlights the importance of quality (energy) of hydrophobic interactions in protein thermostability and provides a way to identify weak spots that should be preferentially mutated for protein engineering aimed at improving thermostability. The results and the computational efficiency of our approach position it as a valuable complement to existing approaches for knowledge-driven protein engineering for improving thermostability.

5 IMPROVING THE CNA APPROACH USING CITRATE SYNTHASE AS A TEST CASE (PUBLICATION II)

5.1 Background

Finding that the energy/quality of hydrophobic interactions is the most important discriminator of protein thermostability (see Chapter 4) prompted us to improve the CNA approach with respect to modeling hydrophobic interaction. Earlier, only hydrogen bonds were removed from the network during the thermal unfolding simulation to simulate the temperature rise, while the number of hydrophobic contacts remained constant (see section 2.7.3).[63,64] We used citrate synthase protein as a test case for validating the improved CNA here.

Citrate synthase (CS) is a homodimeric protein found in nearly all living cells, which catalyzes the first step of the Krebs cycle that synthesizes citrate using acetyl CoA and oxaloacetate. CS is one of the rare proteins for which X-ray crystal structures from several organisms living at temperatures from 0ºC to 100ºC are available in the PDB. This makes CS a highly valuable test case for validating CNA by thermostability prediction and weak spot identification. CS profoundly exists in two different conformations in which the active site is either open or closed depending on whether or not substrates are bound. CS structures in an open conformation from five different organisms living at temperatures from 37ºC to 100ºC were used in this study; they are referred hereinafter by abbreviations with their optimal growth temperatures (T_{og}): *Sus scrofa*: PigCS_37, *Thermoplasma acidophilum*: TaCS_59, *Thermus thermophilus HB8*: TtCS_75, *Sulfolobus solfataricus*: SsCS_87, and *Pyrobaculum aerophilum IM2*: PaCS_100.

5.2 Results

CNA calculations on single input structures did not correctly predict thermostabilities of five CSs. In fact, no correlation between experimental thermostabilities (T_{og}) and predicted thermostabilities (T_p) was obtained ($R^2 < 0.01$). Next, ensemble-based CNA (see section 2.7.4) developed in this study* was performed on 200 conformations of each CS structure extracted from a trajectory of an MD simulation of 10 ns length. This resulted in an improvement in thermostability prediction when compared to the one using single structures ($R^2 = 0.27$, $p = 0.374$, Figure 7); however, the prediction was still far from being satisfactory.

* Ensemble-based CNA was developed by S. Radestock.

Anticipating that the misprediction may arise from neglecting the temperature-dependence of hydrophobic contacts[183,184] with their energy/quality being the most significant determinant of protein thermostability (see Chapter 4), we treated hydrophobic interaction also in a temperature-dependent manner. More hydrophobic constraints were added in the network as the temperature was increased during the thermal unfolding simulation by linearly increasing the cutoff for including hydrophobic constraints D_{cut} (see section 2.7.3). This refined modeling of thermal unfolding simulations significantly improved the thermostability prediction ($R^2 = 0.88$, $p = 0.017$, Figure 7).

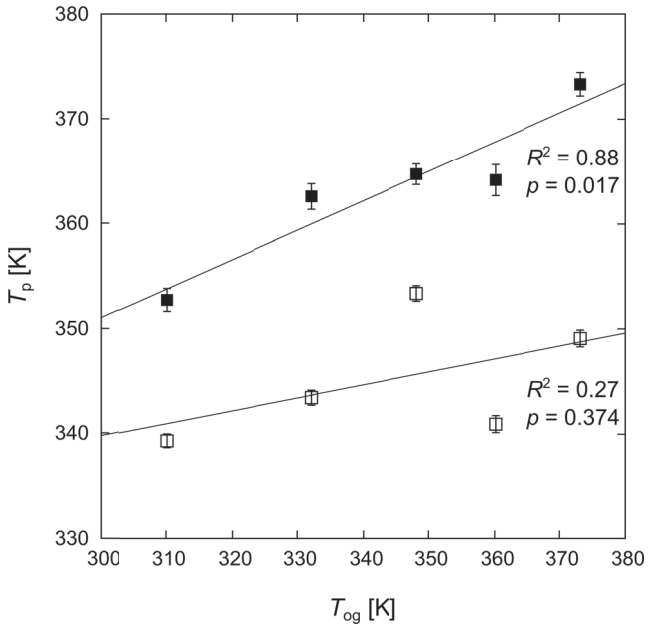


Figure 7. Correlation between T_p and T_{og}. T_p values obtained with and without considering temperature-dependence of hydrophobic interaction are marked by filled and empty squares respectively. Error bars represent the standard error in the mean. Least squares fit lines have been drawn for both the correlations. Figure adapted from ref.[70]

Next, we focused on a microscopic analysis, *i.e.*, weak spot prediction from the CNA. In an ensemble-based CNA developed here, frequency for each residue for being a weak spot across the entire ensemble is counted, and then weak spots are ranked according to these frequencies. Weak spots on CS structures from different organisms were predominantly found in the helices within the monomers, except in the case of TtCS_75 for which they mainly lie on the helices at the dimer interface. These interfacial weak spots in TtCS_75 were found to be reinforced in ScCS_87 by incorporating more hydrophobic contacts. To validate the weak spot prediction and to trace the stepwise thermal adaptation in the series of CS structures, the sequence of a less stable CS was compared to that of the next more

thermostable CS. We observed that weak spot residues predicted by us are preferentially mutated in the next more thermostable CS except in the case of SsCS_87 *vs.* PaCS_100 comparison. However, PaCS_100 adopts a rare mechanism of thermostabilization where a disulfide bond within each monomer results in a topological cross-link of the two chains; such a mechanism is not expected to be considered by our weak spot prediction approach. Furthermore, we observed that the amino acids substituted at the weak spot positions in a more thermostable CS were not generally conserved across a multiple sequence alignment (MSA) of 549 CS sequences. This makes weak spot prediction by CNA all the more significant, indicating that the weak spot positions were not substituted merely by chance. Finally, we explained the mechanism of weak spot reinforcement in atomic detail when going from a less thermostable to a more thermostable CS. As such, a better hydrogen bonding network, formation of aromatic clusters, and improved hydrophobic contacts at dimer interface were identified as mechanisms of weak spot reinforcement in the stepwise thermal adaptation of CSs.

5.3 Conclusions and significance

- The present study for the first time applies CNA to compare a series of five proteins from organisms with T_{og} in the range of 37–100ºC; so far, CNA was only applied to pairs of mesophilic/thermophilic proteins. Not only did this provide a thorough validation of the CNA approach but also allowed deciphering to what extent nature applies different mechanisms for achieving protein thermostability at different temperature ranges.
- The CNA approach was improved in that for the first time, I) Extending the concept of ensemble-based rigidity analysis introduced by H. Gohlke *et al.*,[185] a structural ensemble-based CNA was introduced that alleviates the sensitivity of the CNA results on the input structures (see section 2.7.4) and II) based on our finding that the energy/quality of hydrophobic interactions is the key determinant of thermostability (see Chapter 4), we refined the model underlying thermal unfolding simulations in that now hydrophobic contacts are also modeled in a temperature-dependent manner.
- Using the improved CNA, thermostabilities of a series of five homologous CS structures from organisms living at temperatures between 37ºC to 100ºC were correctly predicted: A very good correlation between experimental (T_m) and predicted thermostabilities (T_p) was obtained ($R^2 = 0.88$, $p = 0.017$).

- It was observed that high ranking weak spots predicted by CNA are more often mutated in a more thermostable CS than low ranking weak spots. This has an important implication for applying CNA in data-driven protein engineering projects aimed at improving protein thermostability: Mutations at high ranking weak spots are expected to more likely improve a protein's thermostability.
- Apart from the methodological advances, the mechanism of stepwise thermal adaptation CS in an atomic detail was analyzed. It was found that different mechanisms are in play during the stepwise thermal adaption of CSs: As such, a better hydrogen bonding network, formation of aromatic clusters, and improved hydrophobic contacts at the dimer interface were identified as ways of weak spot reinforcement in CSs.

6 DEVELOPMENT OF SOFTWARE PACKAGES AND A WEB SERVICE FOR THE CNA APPROACH (PUBLICATIONS III − V)

6.1 Background

A protein's flexibility (and its opposite, rigidity) plays a central role in its stability as well as function. Consequently, in order to relate a protein's structure to its activity and stability, one needs to characterize its flexibility at a great detail. Owing to their time- and resource-intensive nature, experimental methods including X-ray crystallography, cryo-electron microscopy, single-molecule fluorescence, and NMR cannot be routinely applied for biomolecular flexibility characterization. Hence, computational flexibility prediction methods come handy (see section 2.7and 2.8). One of such computational methods is implemented in the FIRST program[65] that characterizes the flexibility of a biomolecule by modeling it as a network (graph) wherein atoms and interactions between them are considered sites (vertices) and constraints (edges), respectively. The CNA approach was developed by S. Radestock and H. Gohlke[63,64] and improved by us (see Chapter 5) carries out thermal unfolding simulations of proteins for linking their flexibility with their function and stability. However, for application of the CNA in a user-friendly manner, a software package was required that, apart from carrying out thermal unfolding simulations, automatically predicts phase transition points and weak spots residue, and computes local and global rigidity indices. To this end we developed a command line CNA software package.[191]

The results from CNA are highly information-rich and require intuitive, synchronized, and interactive visualization for a comprehensive analysis. For instance, the data needs to be visualized as plots showing global and local flexibility indices and weak spots as well as 3D graphics representations of the biomolecule, the constraint network, and the decomposition into rigid and flexible regions. Furthermore, the speed of CNA allows performing real-time rigidity analyses on biomolecules. Thus, a workflow of interactive mutation of a protein and calculation of thermostability of the mutant on-the-fly is possible. Such a workflow would be very useful in protein engineering projects aimed at improving thermostability. With this in mind, we developed VisualCNA, an intuitive, easy-to-use graphical user interface (GUI) for CNA for synchronized and interactive visualization of the CNA results and protein engineering.

Finally, to allow a wider scientific community to use CNA we developed the CNA web service that allows setup of CNA calculations in the web browser in a user-friendly manner

and then assists in comprehending results easily by presenting them in an interactive, graphical manner.[208]

6.2 The CNA software package (Publication III)

6.2.1 Implementation of the CNA software*

The **Publication III** describes the development of the CNA software package for linking biomolecular structure, flexibility, (thermo-)stability and function. CNA takes a single or multiple PDB structures as input (ensemble-based CNA developed by us) and then performs a thermal unfolding simulation on each input structure followed by the calculation of local and global flexibility indices. CNA also allows running ENT^{FNC} calculations developed by C. Pfleger and H. Gohlke[188] in which ensemble of network topologies are derived from the single structure applying fuzzy non-covalent constraint definitions.[188] Thermal unfolding simulations can be run with or without considering temperature dependence of hydrophobic interaction (see Chapter 5). Finally phase transition points and weak spots are identified automatically (Figure 8).

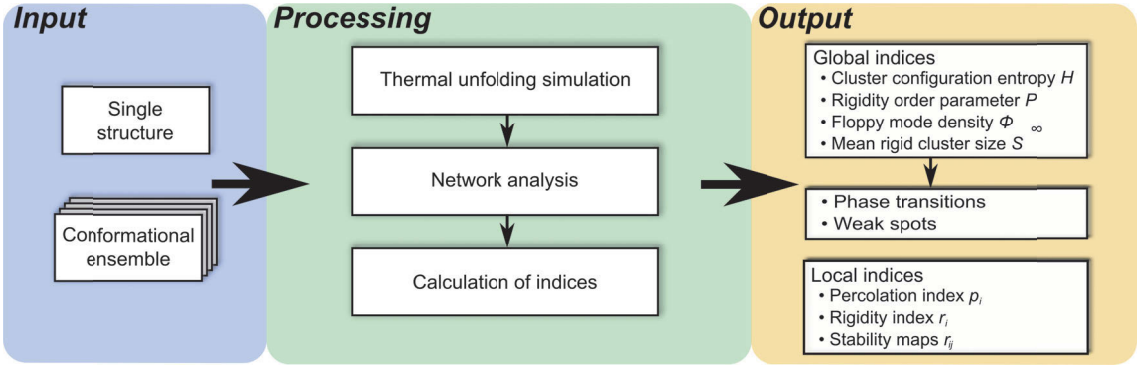


Figure 8. Schematic workflow of the CNA software. The figure adapted from ref.[191]

CNA is a command line program written in the Python programming language. An interface module *pyFIRST* was developed using the Simplified Wrapper and Interface Generator (SWIG) tool.[209] The *pyFIRST* module allows communication between Python (CNA) and C++ (FIRST), which avoids huge input/output overhead caused by reading output generated by the FIRST program. CNA has been written in an object-oriented design that makes extension of the program effortless. Important modules and their roles are described in Table

* The programming code for the CNA software package was written equally by C Pfleger and me, partly using a previous CNA implementation from S. Radestock. The test cases for the package were compiled and implemented by D. L. Klein.

1. The CNA is computationally very efficient; the analysis of a single protein structure by CNA usually takes only a few seconds for systems of several hundred residues on a single core. The runtime for analyses of ensembles of network topologies linearly increases depending on the number of network topologies.

Table 1. Important modules of the CNA program and their functions

Module	Description
CNAnalyis	Parses input options and running requested calculations
pdb	Parses PDB files
dilution	Prepares different network states during the thermal unfolding simulations
ensemble	Performs CNA on ensembles of PDB structures
fnc	Performs CNA on ensembles of network topologies generated using fuzzy non-covalent constraint definitions
network_analyisis	Computes local and global flexibility indices
transitions	Identifies folded-unfolded phase transition points
unfolding_nuclei	Identifies weak spots
output_results	Writes out the results

6.2.2 Showcase example: Flexibility characteristics of hen egg white lysozyme[†]

The usage of the CNA software package has been shown on HEWL as an example. The results from CNA runs of HEWL agree with the experimental findings. Rigid cluster decompositions during the thermal unfolding simulation of HEWL agree, in reverse order, with the *fast track* folding pathway described in refs.[210,211] Unfolding nuclei identified in helix B of HEWL are in agreement with the view that this helix plays a crucial role in stabilizing the tertiary structure of HEWL.[212] As for the local flexibility characteristics, the stable regions identified for residues 53 and 62-65 are in very good agreement with those identified by high protection factors in HDX exchange experiments for the native and denatured states of HEWL.[213] Furthermore, the hinge region predicted by CNA coincides with that suggested by J. A. McCammon *et al.*[214]. From stability maps, very weak contacts were identified for residues 81-87 that partially form a 3_{10} helix; this is in agreement with results from NMR experiments that showed a disordered structure in this region.[215]

[†] The calculation and analysis of flexibility of HEWL using the CNA software was done by C. Pfleger.

6.3 The CNA web service (Publication IV)

6.3.1 Design and implementation of the web service[‡]

The CNA web service was implemented in the Python programming language; it provides a layer of user-friendly input and output interfaces around the CNA software. As input, the web server only requires a PDB code or user-provided PDB file(s) of the input structure(s), and choosing the thermal unfolding simulation type on the submission page of the web service (Figure 9a). The web service supports analysis on a single structure as well on an ensemble of structures using ensemble-based CNA developed by us (see Chapter 5). Furthermore, it allows running CNA on an ensemble of network topologies generated using a single input structure using ENT^{FNC} approach developed by C. Pfleger and H. Gohlke[188]. Results are presented in the browser in an interactive manner using plots (Figure 9 b-d) and the 3D structure using the JmolApplet (Figure 9e). The first part of the results page contains a summary of input parameters used during the run. The second part contains a table that provides information about identified phase transition points. In the case of single-network analysis, plots for six global indices (see section 2.7.5) including H_{type2} (Figure 9d) are presented with the identified phase transition points. The third part contains plots of two local (residue-wise) indices: the rigidity index r_i (Figure 9b) and the percolation index p_i (see section 2.7.5). In the case of an ensemble-based analysis, the standard errors are also depicted in addition to the mean values of the local indices. Furthermore, both the local indices are mapped onto the input structure in a color-coded fashion using JmolApplets. Finally, the fourth part presents a stability map rc_{ij} (Figure 9c) and information about weak spots identified on the protein (see section 2.7.5). Information about weak spots is mapped onto the 3D structure of the input protein in a JmolApplet: in the case of a single-network analysis, identified weak spots are marked by red spheres (Figure 9e); in the case of an ensemble-based analysis, the frequency of a residue for being a weak spot across the entire ensemble is depicted using color coding. The CNA web service is accessible at http://cpclab.uni-duesseldorf.de/cna.

[‡] The CNA web service was implemented by D. M. Krüger building on two previous web services, $Drugscore^{PPI}$ and NMSim (http://cpclab.uni-duesseldorf.de/webservices).

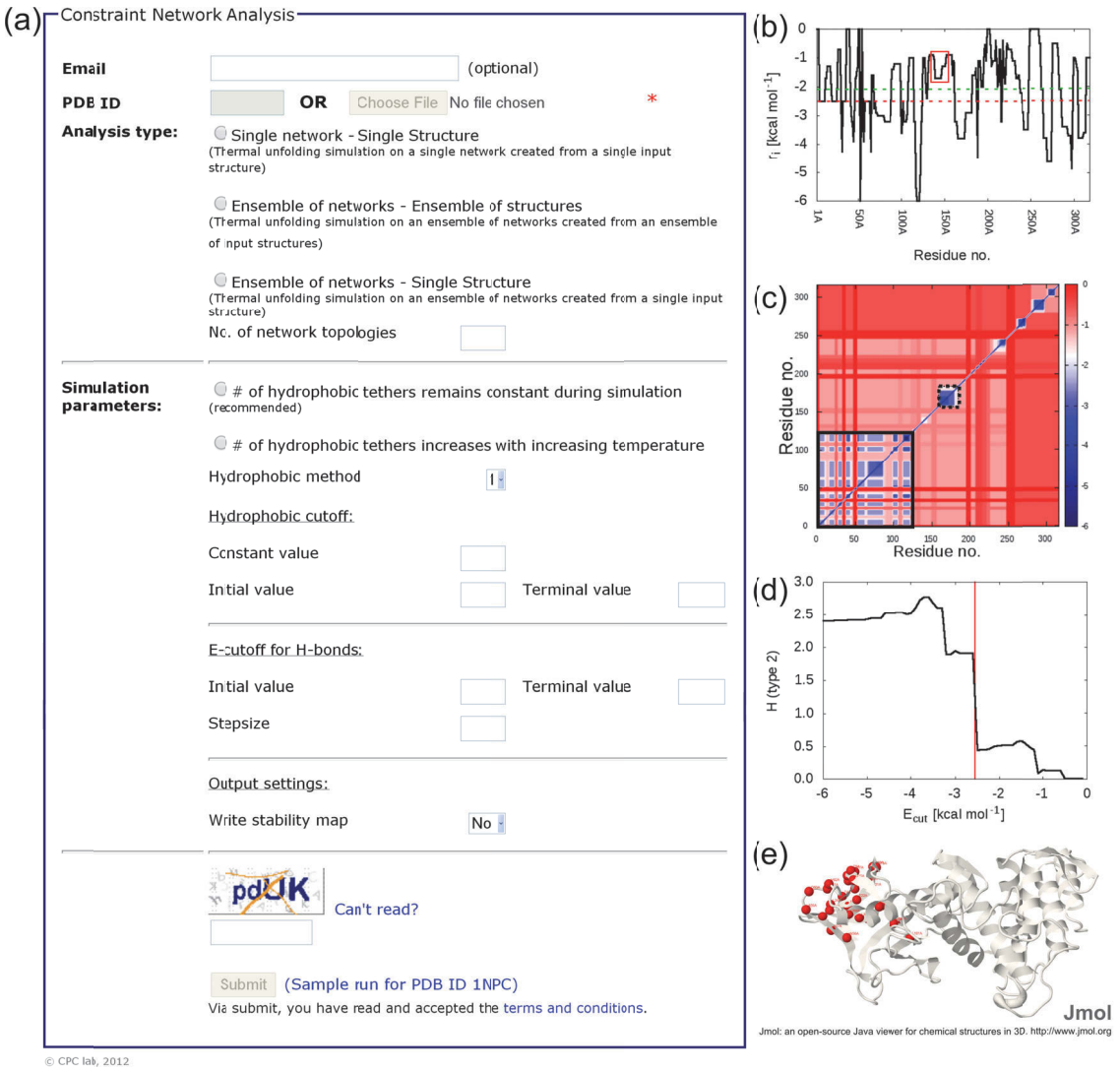


Figure 9. Screenshots of the CNA web service submission page (a) and output using thermolysin-like protease (TLP) as an example. Rigidity index r_i plot wherein red- and green-dashed horizontal lines represent the identified phase transition point and the working temperature of TLP, respectively. The central α-helix and two preceding Gly residues (residues 136–154) are enclosed in a red rectangle (b). Stability map rc_{ij} wherein red colors indicate pairs of residues where no or only a weak rigid contact exists. In contrast, blue colors indicate strong rigid contacts. The black box with a continuous line covers the N-terminal giant rigid cluster, whereas the box with the broken line indicates a rigid cluster in the C-terminal domain (c). Cluster configuration entropy H_{type2} during thermal unfolding of TLP as a function of the hydrogen bonding energy cutoff E_{cut} wherein the red vertical line indicates identified phase transition point T_p (d). Weak spots identified on the TLP structure are represented by red spheres (e). Figure adapted from ref.[208]

6.3.2 Showcase example: Predicting stability characteristics of thermolysin-like protease

In the **Publication IV**, we described rigidity analysis results using thermolysin-like protease (TLP) from *B. subtilis* as a test case. The decay of the giant rigid cluster occurs in a hierarchical fashion as reflected by the presence of multiple steps in the H_{type2} plot (Figure

9d). The reason for this percolation behavior is that the structure of TLP is composed of multiple sub-domains (N-terminal β-sheet domain and C-terminal α-helical domain connected by a central helix) that segregate from the giant cluster independently from each other. A phase transition point was identified at $E_{cut} = -2.55$ kcal mol^{-1} (equivalent to 351 K) on the H_{type2} curve, which is 20 K lower than its thermophilic homologue thermolysin from *Bacillus thermoproteolyticus*. For TLPs, the central α-helix (residues 139–154) and the preceding Gly136 and Gly137 are important with respect to a postulated hinge bending motion.[216,217] In line with this postulation, these residues were identified as being flexible at the working temperature of TLP ($E_{cut} = -2.1$ kcal mol^{-1}) equivalent to 342 K, on the rigidity index plot (Figure 9b). Furthermore, contacts of these residues with other residues of TLP are less stable than contacts between residues of the giant cluster (black box with continuous line) and contacts between residues of a large rigid cluster in the C-terminal domain (black box with broken line) as identified by stability map (Figure 9c). Finally, several of the weak spots identified by CNA in the N-terminal β-sheet domain of TLP (Figure 9e) have been shown to improve thermostability upon mutation in previous studies.[218-221]

6.4 VisualCNA: A GUI for interactive Constraint Network Analysis and protein engineering (Publication V)[§]

6.4.1 Description

VisualCNA is implemented in the Python programming language as a PyMOL plug-in for Linux operating systems. It uses the external modules NumPy, SciPy, Matplotlib, Biopython, tkintertable, and Open Babel. Apart from these external modules, VisualCNA requires CNA and FIRST for rigidity analysis, which are distributed independently.

The VisualCNA GUI consists of four panels: *Setup*, *Analyze*, *Modify*, and *Mutate*. The *Setup* panel presents a form for preparing and running single network[64] or ensemble-based[70,188] variants of thermal unfolding simulations in a user-friendly manner. The *Analyze* panel shows CNA results as interactive plots of global and local flexibility indices and weak spots; these are synchronously linked to the 3D structure visualized in PyMOL in that the state of the thermal unfolding shown and residues selected in PyMol are annotated on the plots. Constraints are visualized as cylinders of different colors according to their type and are grouped by their associated rigid cluster or flexible region to aid visualization and selection.

[§] The programming code for VisualCNA was equally written by D. Mulnaes and me.

The *Modify* panel allows modifying the constraint network of the protein by adding or deleting constraints in multiple ways including a table of constraints (check boxes to enable or disable constraints) and text fields (for entering atom ids to add constraint). The *Mutate* panel is the most important panel of VisualCNA from an application point of view. It allows interactive protein engineering for improving thermostability aided by multiple sequence alignment (MSA) of related proteins. Selecting a residue on the sequence conservation plot and a mutation in the substitution frequency plot (both plots are generated automatically from a user-provided MSA) constructs a model of the corresponding mutant using the PyMOL mutation tool, which allows the user to select an appropriate rotamer for the mutant. The mutants can be automatically submitted to CNA and compared to the WT facilitating interactive analysis of the effect of point mutations to optimize the protein structure towards increased thermostability.

6.4.2 Application scenarios

VisualCNA is an intuitive, easy-to-use graphical interface for CNA for an effortless rigidity analysis of biomolecules and interactive protein engineering even for non-bioinformaticians. With the richness of VisualCNA's functionality, it can be applied in a variety of scenarios from as basic as performing a simple CNA run to a very complex task of modifying the constraint network. A non-exhaustive list of tasks that can be performed using VisualCNA follows.

- Setting up and running CNA on a single structure or a structural ensemble.
- Interactive analysis of CNA results to identify critical constraints that break at transitions involving major loss of structural rigidity during thermal unfolding. The analysis helps identifying weak spots that should potentially be mutated to improve stability of the protein.
- Comparison of CNA results of two or more systems, e.g., a WT and a mutant or a mesophilic and thermophilic homologous protein pair to unravel mechanism of thermostabilization, identify weak spots on the mesophilic protein, and understanding functionally important residues by comparing active site flexibility.
- Modification of protein constraint network by deletion and addition of constraints when constraints are not correctly identified for ligands, ions, or non-standard residues automatically when studying the effect of ligand binding on the flexibility of proteins.

- Engineering an existing protein for improving its thermostability using an interactive workflow of mutation modeling → CNA run of the mutant → comparison with the WT.

6.5 Conclusions and significance

- A command-line based CNA software package was developed that efficiently characterizes a biomolecule's flexibility and links it to stability and function. The software allows setting up a variety of constraint network representations, processing the results obtained from FIRST, and calculating global and local indices for quantifying biomolecular stability.
- The CNA software incorporates several methodological advances in the field of rigidity analysis: (I) Analysis of ensembles of network topologies derived from a structural ensemble (see Chapter 5) or from a single structure using fuzzy non-covalent constraint definitions developed by C. Pfleger and H. Gohlke[188] giving more robust results that are not sensitive to the conformation of input structure; (II) temperature-dependent modeling of hydrophobic constraints that improves thermostability prediction (see Chapter 5), (III) automatic identification of small molecule constraints; (IV) automatic detection of phase transition points and weak spots for assisting protein engineering.
- Using the CNA software, we analyzed flexibility characteristics for the example HEWL that are in agreement with experimental findings including its unfolding pathway, residue flexibility inferred from HDX experiments, and disordered regions identified using NMR experiments.
- For a user-friendly application of CNA and interactive analysis of the results, a web service is developed. The CNA web service provides an input interface for easy setup of CNA calculations in the web browser. The results are graphically displayed in the form of plots and on the 3D structures allowing an in-depth analysis.
- CNA web service was validated in that flexibility characteristics of TLP predicted using the web service agree well with experimental findings.
- Finally, VisualCNA, a GUI for CNA was developed for allowing an easy setup of CNA calculations and interactive analysis of results using synchronized plots of rigidity indices (see section 2.7.5) and 3D structure showing rigid cluster decompositions and the constraint network (see section 2.7.2).

- The two most striking features of VisualCNA are: I) Facility for manual editing of constraint network, which is very useful when modeling non-standard residue, ligands, ions, etc. II) Workflow for interactive protein engineering involving computational mutagenesis followed by thermostability estimation using CNA; the workflow can be run iteratively until a mutant with desired thermostability is found.
- Given their speed and prediction accuracies, these software and the web service are valuable tools for protein engineering projects in general, particularly for the projects aimed at improving thermostability. Furthermore, VisualCNA and the web service allow application of CNA in a user-friendly manner making CNA studies amenable to non-bioinformaticians interested in rigidity analysis of proteins.

7 STRUCTURAL RIGIDITY AND THERMOSTABILITY OF *BACILLUS SUBTILIS* LIPASE A (PUBLICATION VI)

7.1 Background

Understanding the mechanism of elevated thermostability is of fundamental importance in protein science because it would allow engineering proteins for withstanding high temperatures.[28,29] Opposing views on increased or decreased structural rigidity of the folded state have been put forward in this context.[63,64,69-74] In part, they have been related to different mechanisms of thermostabilization[76] and the temporal resolution of the experimental technique or computational analysis used to measure protein flexibility.[99,100,102-104,109,110] In the **Publication VI**, we address the question of the relation between structural rigidity and protein thermostability by analyzing *directly* the static properties of a well-characterized set of 16 mutants of lipase A from *Bacillus subtilis* (*Bs*LipA). We do so by applying CNA on ensembles of network topologies (ENT^{FNC}),[188] thereby considering the *Bs*LipA variants to be in static equilibrium. Therefore, the rigidity and flexibility characteristics derived that way are time-independent. This is the first time that we apply CNA on several closely related variants of a protein; earlier only pairs of meso-/thermophilic homologs[63,64] or a series of homologous proteins from five different organisms were studies using CNA (see Chapter 5).

*Bs*LipA is an important member of the lipase class of enzymes and used in diverse biotechnological applications.[222,223] Owing to its importance, *Bs*LipA has been extensively studied with respect to structure[224-227] and thermostability[62,82,85-88,228]. As to the latter, M. T. Reetz *et al.* applied iterative saturation mutagenesis on the most flexible amino acids as identified by crystallographic B-factors, which resulted in *Bs*LipA mutants that were more thermostable than the WT showing an increase in T_{50}^{60} (the temperature required to reduce the initial enzymatic activity by 50% within 60 min) of ≤ 45 K.[62] Subsequent biophysical characterization of the three most thermostable mutants revealed that the improved activity retention resulted from a reduced rate of protein unfolding as well as a reduced precipitation of the unfolding intermediates, *i.e.*, due to kinetic reasons.[83] In contrast, N. M. Rao *et al.* sequentially developed several thermostable *Bs*LipA mutants using directed evolution assisted by structural information. These mutants were shown to be more thermostable than the WT due to predominantly thermodynamic reasons;[82,84-88] the most thermostable mutant displayed an increase in the melting temperature T_m of ~22 K. Apart from being a valuable test set for analyzing the relation between structural rigidity and thermostability, this also

forms a very good test set for validation of CNA to evaluate whether CNA can sense differences in thermostability between structures that are highly similar (pairwise sequence identity > 93% and pairwise RMSD < 0.38 Å).

7.2 Results

The loss of rigidity percolation of the WT *Bs*LipA during thermal unfolding as identified using CNA was in agreement with experimental findings on unfolding of proteins with an α/β hydrolase fold[229,230] in that it showed an early loss of α-helices during unfolding. The correct identification of the unfolding pathway of WT *Bs*LipA strongly indicates that side-chain-mediated interactions between amino acids are correctly represented by the ENT^{FNC}-based CNA and hence thermostability predictions can be relied upon. Next, using ENT^{FNC}[188] a significant (p = 0.002) correlation between predicted (H_{type2}-derived T_p) and experimental (T_m) thermostabilities with $R^2 = 0.58$ was obtained for thermodynamically thermostabilized mutants if the two structures with the lowest and highest T_m were considered outliers (Figure 10a). The reason for misprediction of the two outliers was traced to their different unfolding pathway distribution (derived by clustering of the percolation index (p_i) profiles of individual networks of the ensemble) compared to that of other variants. We propose the similarity (dissimilarity) in the unfolding pathways as an indicator for a reliable (unreliable) prediction of relative thermostabilities of proteins by CNA. As an alternative measure, which is less sensitive to the underlying unfolding pathways, we defined the *median stability of rigid contacts between residue-neighbors* $\tilde{rc}_{ij,neighbor}$ calculated from neighbor stability maps (see section 2.7.5) for predicting thermodynamic thermostability. A significant and fair correlation of $\tilde{rc}_{ij,neighbor}$ with T_m values of the thermodynamically stable mutants from N. M. Rao *et al.* was obtained (R^2 = 0.46, p = 0.004) with the two previous outliers being correctly ranked now (Figure 10b). Based on these findings, we recommend using H_{type2}-derived T_p values for comparing thermostabilities of proteins unless the underlying unfolding pathways of the proteins are dissimilar; in that case, we recommend using $\tilde{rc}_{ij,neighbor}$.

To probe the sensitivity of the ENT^{FNC} to the input structure used, we computed $\tilde{rc}_{ij,neighbor}$ using the ENT^{FNC} approach for five additional crystal structures of WT *Bs*LipA (Figure 10b). The standard error of the mean in $\tilde{rc}_{ij,neighbor}$ over all six WT *Bs*LipA structures is 0.57 K, which is likely within the experimental uncertainty, confirming previous results of robust rigidity analyses with ENT^{FNC}.[188] Kinetically stabilized mutants from M. T. Reetz *et al.* are found to have a lower thermodynamic thermostability than the WT both in T_p-based and

$\tilde{rc}_{ij,neighbor}$-based predictions, which is in very good agreement with experimental findings.[83]

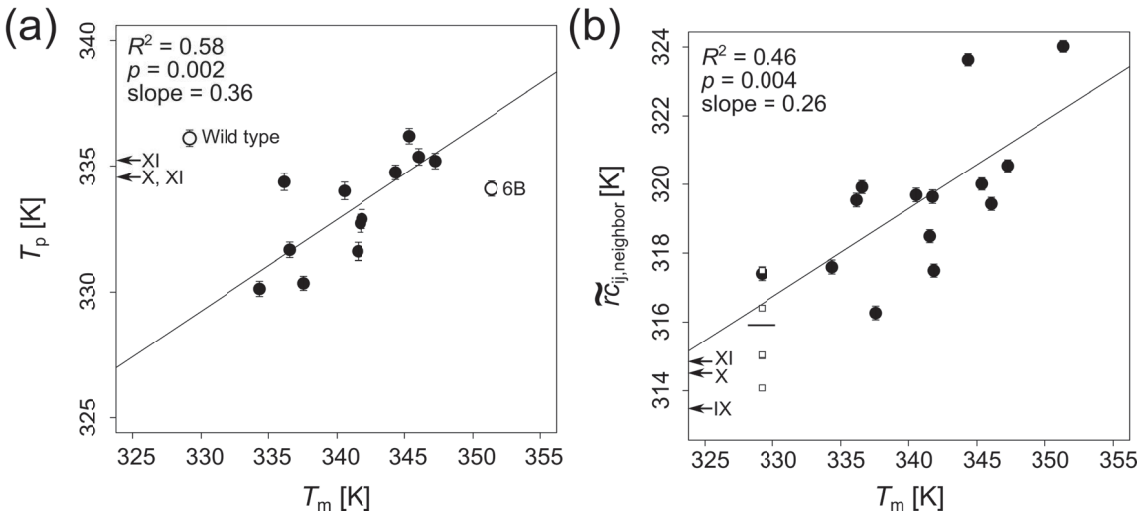


Figure 10. Correlation between predicted and experimental thermostabilities (T_m values) of *Bs*LipA variants; for the predictions, the ENT^{FNC} approach was used. (a): Correlation between T_p derived from the global index H_{type2} and T_m values for thermodynamically thermostabilized mutants from Rao *et al*. Data points shown as empty circles were considered outliers (see the main text for explanation) and excluded when calculating R^2 values and the correlation lines. (b): Correlation between $\tilde{rc}_{ij,neighbor}$ and T_m values for thermodynamically thermostabilized mutants from Rao *et al*. Data points shown as empty squares represent $\tilde{rc}_{ij,neighbor}$ values for five additional wild type crystal structures (see main text for details; two of the squares closely overlap; mean $\tilde{rc}_{ij,neighbor}$ over all six data points for wild type structures is shown as a small horizontal line: 315.9 ± 0.6 K). (a) and (b): Error bars represent the standard error in the mean. T_p and $\tilde{rc}_{ij,neighbor}$ values for kinetically thermostabilized mutants from Reetz *et al*. are marked by arrows on the corresponding ordinates.

Finally, we analyzed on a residue basis how changes in the thermostability relate to changes in local structural stability (rigidity) by comparing stability maps of the mutants against the WT. Thermodynamically stable mutants from N. M. Rao *et al.* showed increased strengths of certain inter-helical (αC/αD and αA/αF) and helix-central β-sheet (αA, αF) contacts as compared to the WT (Figure 11 a and b). As expected, the increase in the contact stability was the more pronounced the higher the thermodynamic thermostability is of the mutant (Figure 11 a and b). Surprisingly, inter-helical contacts for αB/αC are weaker in thermostable mutants than in the WT (Figure 11 a and b). This indicates that the strengthened stability between helices and the central β-sheet region is sufficient to keep the structure folded; this reduced rigidity of αB/αC helix pairs might even make the fold entropically more favorable.[74,231,232] In contrast, kinetically thermostabilized mutants from M. T. Reetz *et al.* showed a destabilization in rigid contacts between most residue neighbors (Figure 11c). These results are in line with experimental findings according to which these mutants are

thermodynamically less stable than the WT and with crystallographic B-factors observed for one of the variants.[83]

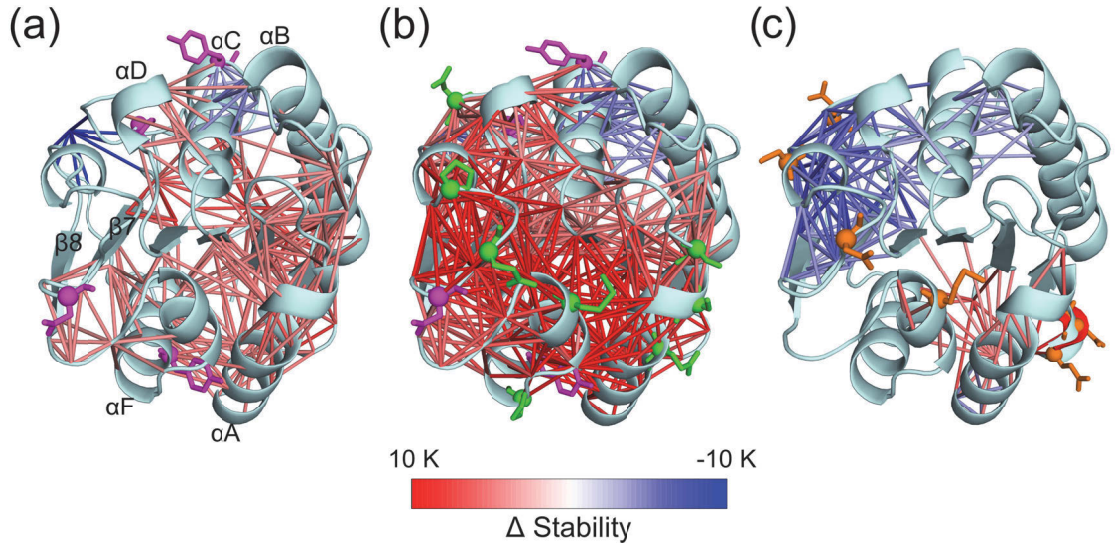


Figure 11. The differences in the stability of rigid contacts for residue neighbors displayed on the structures of the mutants by sticks connecting C_α atoms of residue pairs colored according to the color scale in the bottom for (a): 1-14F5 (the least thermostable variant from N. M. Rao *et al.*), (b): 6B (the most thermostable variant from N. M. Rao *et al.*), and (c): X (kinetically thermostabilized variant from M.T. Reetz *et al.*). Only those contacts that are stabilized by ≥ 4 K or destabilized by ≥ 3 K are shown for clarity; for the same reason, contacts between two residues of the same secondary structure element are not shown. Mutated residues are shown as sticks and a sphere at their C_α atoms: Common mutations in 1-14F5 (a) and 6B (b) are shown in magenta, unique mutations in 6B (b) are shown in green, and mutations in X (c) are shown in orange.

7.3 Conclusions and significance

- The main outcome of this work is the finding of a good correlation between the structural rigidity of all *Bs*LipA variants and their thermodynamic thermostability. This finding is highly relevant in the context of intense discussions if elevated protein thermostability is related to increased or decreased structural rigidity of the folded state.[69,75-80]
- For the first time, we showed that thermodynamic thermostability of mutants differing by as little as three to twelve mutations from the WT can be successfully predicted using the CNA.
- A good and statistically significant correlations between experimental melting temperatures (T_m) of mutants of Rao *et al.* and predicted thermodynamic thermostabilities have been found based on two independent measures (H_{type2} and $\widetilde{rc}_{ij,neighbor}$), as was correctly predicted that the thermodynamic thermostability of the mutants of Reetz *et al.* is lower than that of the wild type.

- We introduced the similarity/dissimilarity of unfolding pathways as a measure for judging thermostability predictions from CNA.
- A new measure $\widetilde{rc}_{ij,neighbor}$ was introduced for predicting thermodynamic thermostability. This measure is less sensitive to details of the unfolding pathway and, hence, CNA-predicted thermostabilities of the mutants that have dissimilar unfolding pathways can also be reliably compared.
- Finally, the mechanism of thermostabilization of thermodynamically thermostabilized mutants was rationalized based on the distribution of rigidity and flexibility on their structures: Thermodynamically stabilized mutants were unequivocally characterized by an overall increased structural rigidity; whereas, kinetically thermostabilized mutants showed a reduced rigidity.

8 PREDICTING THERMOSTABILIZING MUTATIONS ON *BACILLUS SUBTILIS* LIPASE A (PUBLICATION VII)

8.1 Background

High thermostability is a desired characteristic for proteins, more so when they are used in bio-technology industry because it allows bio-processes being carried out at high temperatures.[28,29] Engineering an existing protein for improving its thermostability via mutagenesis[25] is frequently attempted because not all enzymes in nature are optimized to withstand high temperature conditions.[30] A commonly used technique for improving a protein's thermostability, directed evolution, aims at simulating natural evolution of proteins in laboratory by performing cycles of random mutagenesis and selection of thermostable variants.[88,131-137] However, the approach is limited in that only a handful of all possible mutations of a protein can be experimentally tested.[37] To this end, data-driven approaches restrict the library size for screening by suggesting weak spots on a protein, *i.e.*, amino acid positions that are most likely to improve thermostability upon mutations. One such data-driven approach termed Constraint Network Analysis (CNA) for identifying weak spots has been developed by S. Radestock and H. Gohlke[63,64] and improved by us (see Chapter 5). The approach has been validated in retrospective studies[63,64] wherein the order of thermostabilities of mesophilic/thermophilic protein homologs was correctly predicted and so were the weak spots. Furthermore, I showed that CNA can successfully predict changes in thermodynamic thermostability arising out of a handful of mutations (see Chapter 7). As a logical next step to validate CNA prospectively, in the **Publication VII**, we developed a novel, unique, and highly efficient strategy employing weak spot prediction using structural ensemble-based CNA (see section 2.7.4 and Chapter 5). This strategy is assisted by information on sequence conservation in a multiple sequence alignment (MSA), and ANOLEA-based[233,234] quality assessment of the substituted amino acid at the weak spots. The present strategy, going one step further than merely predicting weak spot residues, also predicts optimum substitutions at these weak spots. The strategy was applied to develop thermostable variants of *Bs*LipA.

8.2 Strategy for predicting thermostabilizing mutations

Thermal unfolding simulation of a structural ensemble of *Bs*LipA obtained from MD simulations was carried out using ensemble-based CNA developed by us (see section 2.7.4 and Chapter 5; Figure 12 I and II). At variance with the previous way of weak spot identification only at the last major phase transition point related to the terminal loss of

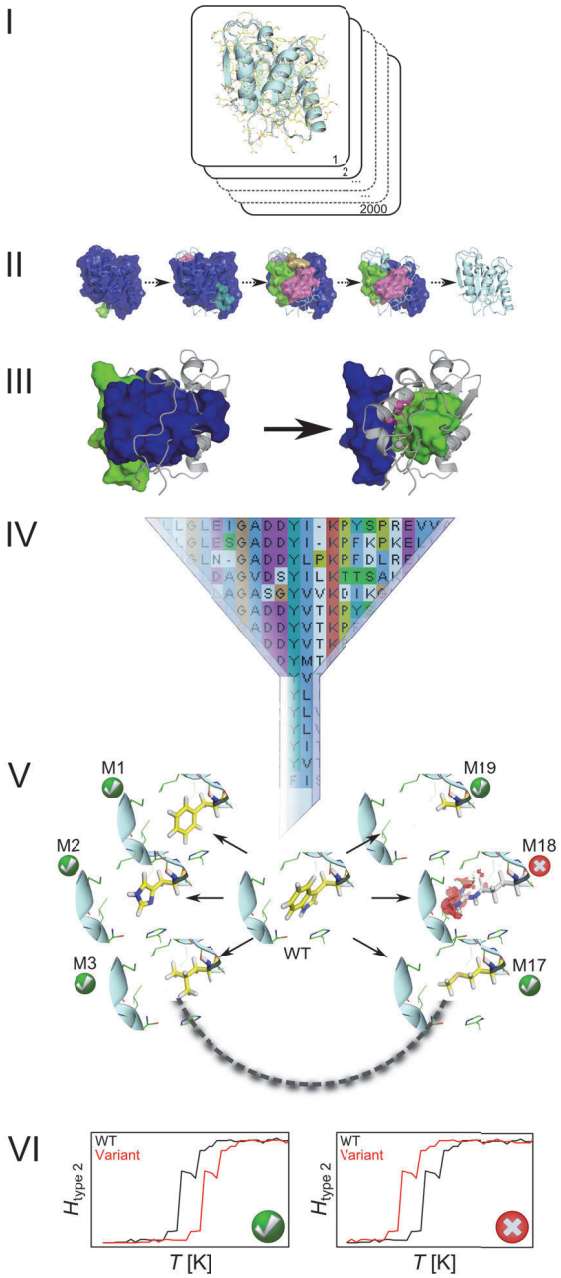


Figure 12. Strategy to rationally predict mutations that increase structural rigidity and thermostability. A structural ensemble of the respective protein is generated by MD simulations (I). The average thermal unfolding trajectory depicting a decomposition into rigid clusters (in the order of decreasing size colored in blue, green, magenta, cyan, orange, and violet) for each step of the unfolding simulation is created by subjecting the structural ensemble to CNA (II). For every major transition during the thermal unfolding, weak spot residues (depicted as a sphere for the C_α atom and sticks for the side-chain) are identified as residues that segregate from the largest rigid cluster and can potentially interact with the then largest rigid cluster upon mutation (III). Weak spot residues identified in step III that are highly conserved in a multiple sequence alignment of the protein family ($\geq$ 80% identity) are removed from the weak spot list (IV). For each remaining weak spot, structures of single-point variants involving mutations to all other 19 amino acids (termed M1-M19) are generated using the SCWRL program. Mutations that lead to energetically unfavorable structures (indicated by red discs around the mutated residue in the case of M18) as calculated by the ANOLEA server are not considered further (V). Finally, for each variant, the phase transition temperature T_p is computed using CNA; a higher T_p value than that of the WT protein indicates a thermostabilizing mutation (VI). All figures of *Bs*LipA structures in this publication were generated with PyMOL (http://www.pymol.org).

rigidity in the protein (see section 2.7.5),[63,64,70] in the present study we identified weak spots at all major transitions involving a substantial loss of rigidity during the thermal unfolding. The procedure followed here has the advantage that it allows evaluating whether strengthening residues that segregate from the largest rigid cluster in the early steps of thermal unfolding also stabilizes a protein. Such phase transitions were identified by visual inspection of the unfolding trajectory with the help of the VisualCNA software developed by us (see Chapter 6). On all such five major phase transitions, potential weak spots were identified as those residues that are spatially close to the largest rigid cluster from which they segregated (Figure 12 III). From this list of potential weak spot residues, highly conserved residues (≥ 80% sequence identity in a MSA of 296 lipase class 2 sequences obtained from the Pfam database[235]) were removed (Figure 12 IV). This was done because conserved residues are important for function and stability of a protein and, hence, should not be mutated.[50,51,236,237] Next, structures of all possible mutations at each weak spot residue were generated by the SCWRL program[238] using WT *Bs*LipA (PDB ID: 1ISP) as a template (Figure 12 V). Variant structures with mutated residues unfavorably embedded in the surroundings of the protein as judged by the ANOELA energy[233] were discarded (Figure 12 V). In such structures, the mutation apparently does not fit into the environment of the other residues. Then, the phase transition point T_p was predicted for each variant using ensemble-based ENT^{FNC} approach developed by C. Pfleger and H. Gohlke (see section 2.7.4).[188] Finally the variants were prioritized based on their ΔT_p (T_p (variant) − T_p (WT)) values resulting in twelve *Bs*LIpA variants for experimental evaluation (Figure 12 VI): For each weak spot residue and all mutations with $\Delta T_p > 1$ K, the mutation with the highest ΔT_p was chosen for experimental validation. The sole exception is G104 located in the active site, for which two mutations were chosen.

8.3 Results*

The thermostability of *Bs*LipA variants was quantified by T'_{50} values; these values report on the temperature at which the fraction of the activity to the initial activity (at 40ºC) is 50% after incubation for 30 min. This is different from the T_{50} values normally used for characterizing the thermostability of proteins[62,239,240] in that the activity here is measured at the temperature of incubation, not at room temperature after cooling. T'_{50} thus reports on the thermo-tolerance of an enzyme during operational bioprocesses carried out at elevated

* All non-computational experiments were performed by Alexander Fulton in the group of Prof. Dr. Karl-Erich Jaeger at the Institute of Molecular Enzymtechnology (IMET), Heinrich Heine University, Düsseldorf.

temperatures for a longer duration of time, e.g., as done in the lipid processing industry.[241] The three variants V54H, F58I, and V96S have T'_{50} values higher by 5.7, 6.6, and 3.6ºC, respectively, than WT *Bs*LipA. Furthermore, no significant impact on the Michaelis constants (K_M) was observed, and the turnover numbers (k_{cat}) were reduced by at most 25%. Thus, the thermostability of the variants has been increased without significantly influencing k_{cat} / K_M at 40°C. The three thermostable variants involve mutations at weak spots identified at the last two phase transitions during the thermal unfolding simulation. This finding supports the reasoning that the late phase transition(s) involving the final decay of the rigid core during thermal unfolding determine(s) the thermostability of a protein and, hence, strengthening residue contacts at these transitions should improve protein thermostability.[63,64] Accordingly, the three thermostable variants have in general stronger rigid contacts between residue neighbors both locally (close to the mutation site due to a better side-chain packing; Figure 13a) and globally (due to long-range stabilization; Figure 13b) than the WT as shown by neighbor stability maps (see section 2.7.5). As such, on an average, V54H, F58I, and V96S increased the strength of rigid contacts of residue neighbors (within 5 Å of each other) by 2.0, 1.2, and 0.4 K, respectively, compared to WT. As an apparent contradiction to our reasoning, five mutations at weak spots identified at the last two transitions resulted in lower T'_{50} values than that of WT *Bs*LipA. In each case, however, a small amino acid was substituted by a large amino acid, which likely could not be accommodated by the fold. This calls for improved modeling approaches for the variant construction in future studies, e.g., by applying comparative modeling rather than side-chain placement only as done in SCWRL.[238]

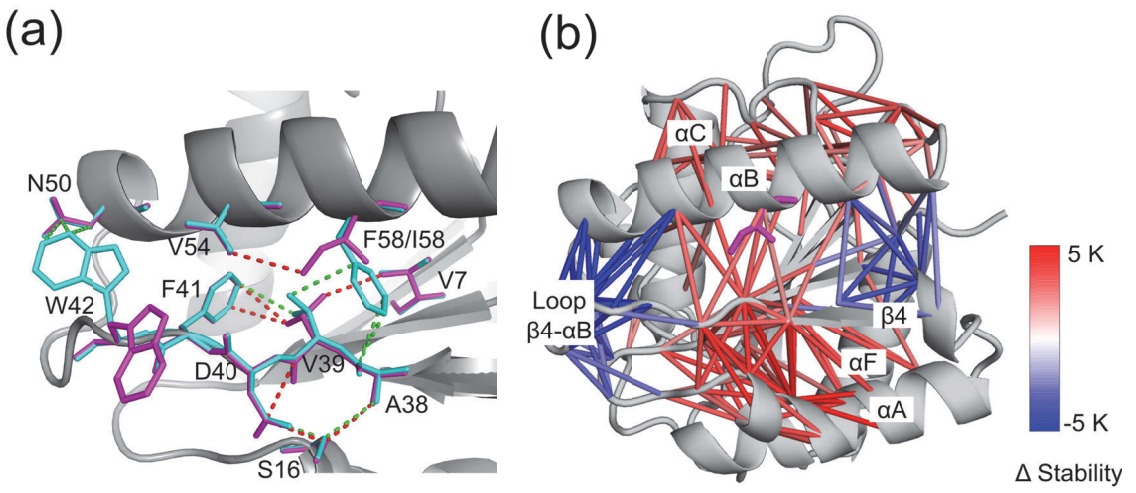


Figure 13. Hydrophobic contacts in the proximity of the mutation site between carbon atom pairs at most 3.8 Å apart are sown as green (WT) and red (F58I) dashed lines (a). Residues involved in making such contacts are shown as cyan (WT) and magenta (F58I) sticks. Differences in the stability of "rigid contacts" between variant

F58I and WT depicted on the variant structure (b). Two residues form a "rigid contact" if they belong to one rigid cluster. A red (blue) stick connecting C_α atoms of two residues indicates that a rigid contact in the variant is more (less) stable than in the WT (see color scale). Only those contacts of variant F58I that are stabilized or destabilized by ≥ 2 K are shown for clarity; for the same reason, contacts between two residues of the same secondary structure element are not shown. The mutated residue I58 is displayed by magenta sticks.

8.4 Conclusions and significance

- A novel, unique, and efficient strategy for predicting thermostabilizing mutations for a protein was developed, which involves ensemble-based CNA, information on residue conservation in an MSA, and ANOLEA[233,234] based quality assessment of the substituted amino acid at the weak spot residues.
- Of the twelve predicted variants harboring a single mutation, three variants, V54H, F58I, and V96S, increased T'_{50} by 5.7, 6.6 and 3.6ºC, respectively.
- This 25% enrichment is extraordinary considering the fact that only 3% of all possible single point mutants of *Bs*LipA showed an increased detergent stability in another study (A. Fulton, J. Frauenkron-Machedjou, P. Skoczinski, S. Wilhelm, U. Schwaneberg, and K.-E. Jaeger; unpublished results).
- All three thermostabilizing mutations were predicted on the weak spots identified at late transitions supporting the reasoning that the late phase transition(s) involving the final decay of the rigid core during thermal unfolding determine(s) the thermostability of a protein.[63,64] We recommend considering only residues segregating from the largest rigid cluster at such late transitions as weak spots in future studies.
- According to our study on how structural rigidity is related to protein thermostability (see Chapter 7), we expect our variants to be more thermostable due to thermodynamic reasons because these variants are more rigid than the WT.
- We showed for the first time that CNA can be successfully applied to assist protein engineering projects aimed at improving thermostability to reduce the time and efforts required in such projects.

9 SUMMARY AND PERSPECTIVES

In the present thesis, we developed software tools and computational approaches for predicting thermostabilizing mutations on proteins. One approach developed by us in this thesis is based on the clustering of residue according to their hydrophobic interaction energies, while the other approach that we improved is a graph-theory based Constraint Network Analysis (CNA) originally developed by S. Radestock and H. Gohlke.[63,64] Using the former approach, we correctly predicted thermostabilizing and -destabilizing residues with a classification accuracy of ~70% in a retrospective analysis. With a scheme based on the latter approach, we prospectively developed three thermostable variants of *Bs*LipA that showed an increase in T'_{50} values as high as 6.6ºC.

Based on our finding that the energy/quality of hydrophobic interaction is the most significant determinant of protein thermostability (see Chapter 4; **Publication I**), we improved the CNA approach[50,51] by modeling the temperature-dependence of hydrophobic interactions; only hydrogen bonds had been modeled in a temperature-dependent manner before.[183,184] Furthermore, by extending the concept of ensemble-based rigidity analysis from H. Gohlke *et al*,[185] we developed ensemble-based CNA. These improvements resulted in a promising prediction of thermostabilities of five citrate synthase structures yielding a significant ($p = 0.017$) correlation between predicted and experimental thermostabilities ($R^2 = 0.88$; see Chapter 5; **Publication II**). We envisage incorporating effects of hydrogen bond cooperativity[242] in constraint modeling and developing novel ways of weak spot prediction as future improvements for CNA.

Significance of the work carried out in this thesis is also evident from the development of two software tools and a web service for CNA. The CNA command line software offers calculation of local and global rigidity indices, identification of weak spot residues, and prediction of protein thermostability (see Chapter 6; **Publication III**); VisualCNA, a GUI to the CNA software, allows synchronized, interactive, and detailed analysis of CNA results and protein engineering for improving thermostability (see Chapter 6**; Publication V**). The CNA web service allows running CNA calculations and analyzing results graphically in a web browser (see Chapter 6; **Publication IV**). A future extension of VisualCNA by integrating a homology modeling module would increase the scope of CNA by allowing CNA on proteins for which no experimental structure is available.

Next, we ventured in to studying the relation between structural rigidity and protein thermostability. *Bs*LipA was an obvious test case for this since several thermostable mutants for this are reported in the literature.[62,82-88] The main outcome of this work is the finding of a good correlation between the structural rigidity of all *Bs*LipA variants and their thermodynamic thermostability. On the way, we carefully probed for the sensitivity of the results with respect to the input structures and developed an approach for detecting outliers based on differences in the pathways of thermal unfolding. We furthermore introduced a local stability measure for predicting thermodynamic thermostability, which complements the detection of the (global) phase transition point T_p (see Chapter 7; **Publication VI**). Although it remains to be shown how one can predict kinetic thermostabilization; a scheme involving generation of unfolded ensemble of a protein and subsequent prediction of its aggregation propensity should answer this.

With the motivation from the correct prediction of thermodynamic thermostabilities of sequentially closely related *Bs*LipA variants, we developed a scheme for predicting thermostabilizing mutations on a protein based on predictions by CNA assisted by information on residue conservation in a multiple sequence alignment. Applying this scheme on *Bs*LipA, three out of the twelve predicted single point mutants showed an improved thermostability (see Chapter 8; **Publication VII**). As a future step, we envisage to characterize the origin of the thermostability of these variants and construct multiple mutants combining single point mutations to evaluate how individual increases in thermostability add up. As a further advancement to our approach, a method to predict the effect of a mutation on the activity of a protein would be highly useful.

I believe that the findings of this work and the tools developed will prove valuable in protein engineering projects in general and specifically in projects aimed at improving thermodynamic thermostability.

10 ACKNOWLEDGEMENTS

First and foremost, I like to thank Prof. Dr. Holger Gohlke, my Ph.D. supervisor for his continual guidance, support, and motivation without which this work would not have seen this day. I also thank my co-supervisor Prof. Dr. Karl-Erich Jaeger for his valuable suggestions on this research work.

I thank all members of the Gohlke group for their thorough professionalism, kindness, and rigorous scientific discussions. In particular, I thank Christopher Pfleger, Daniel Mulnaes, Dr. Dennis M. Krüger, Doris L. Klein, Dr. Sebastian Radestock, and Dr. Simone Fulle for a successful scientific collaboration. I also thank Alexander Metz, Anuseema, Bartholomäus Daniel Ciupka, Christoph G. W. Gertzen, Jagmohan Saini, and Nadine Homeyer for helping me in numerous ways during this work. I thank Peter Sippel, Christopher Pfleger, and Christian Hanke for keeping computing machines always up. Finally, I thank Christoph G. W. Gertzen and Daniel Mulnaes for proofreading this thesis and Alexander Metz for translating the abstract into German.

I am grateful to Prof. Dr. Karl-Erich Jaeger and Alexander Fulton (Institute of Molecular Enzymtechnology (IMET), Heinrich Heine University, Düsseldorf) for the fruitful collaboration on part of this work. I also thank Dr. Hans Wolfgang Höffken and BASF, Ludwigshafen for providing me an opportunity for an internship at BASF.

I am grateful to Prof. Dr. P. V. Bharatam (NIPER, S. A. S. Nagar, India) and Dr. Prashant V. Desai (Eli Lilly, Indianapolis, IN, United States) for imparting basics of scientific research in me and convincing me to pursue a doctoral degree.

I am grateful to the Ministry of Innovation, Science, and Research of North Rhine-Westphalia and Heinrich-Heine-University, Düsseldorf, for a scholarship within the CLIB-Graduate Cluster Industrial Biotechnology. I acknowledge the “Zentrum fuer Informations- und Medientechnologie” (ZIM) at the Heinrich Heine University, Düsseldorf, for computational support.

I thank my flat mates, Rohit and Ajit, and my badminton-pals, Saan, Sanil, and Subash for making my stay at Düsseldorf enjoyable. I also thank my friends Hitesh and Rajendra for always encouraging me to achieve my goals.

I am falling short of words in expressing my gratitude for my family members, my mother, my brother, my sisters, and my better half Pramila Soni. Their love, care, and encouragement helped me concentrate on my work despite being far, far away from them.

11 APPENDIX

11.1 Figure creation

All 3D representations of molecules were created with PyMOL (http://www.pymol.org). All 2D data plots were generated with R program (http://www.r-project.org). All vector graphics were created using Inkscape program (http://www.inkscape.org)

11.2 Reprint permissions for publications

I am grateful to the American Chemical Society for the permission to reprint **Publication I**.
Reprinted (adapted) with permission from:
Prakash Chandra Rathi, Hans Wolfgang Höffken, and Holger Gohlke, Quality matters: Extension of clusters of residues with good hydrophobic contacts stabilize (hyper)thermophilic proteins. *Journal of Chemical Information and Modeling*, **2014**, 54: 355–361.
Copyright (2013) American Chemical Society.

I am grateful to Elsevier for the permission to reprint **Publication II.**
Reprinted (adapted) from Journal of Biotechnology, 159, Prakash Chandra Rathi, Sebastian Radestock, and Holger Gohlke, Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio, 135-144, Copyright (2012), with permission from Elsevier.

I am grateful to the American Chemical Society for the permission to reprint **Publication III**.
Reprinted (adapted) with permission from:
Christopher Pfleger, Prakash Chandra Rathi, Doris L. Klein, Sebastian Radestock, and Holger Gohlke, Constraint Network Analysis (CNA): A Python software package for efficiently linking biomacromolecular structure, flexibility, (thermo)stability, and function. *Journal of Chemical Information and Modeling*, **2013**, 53:1007-1015.
Copyright (2013) American Chemical Society.

I am grateful to the Oxford University Press for the permission to reprint **Publication IV**.
Reprinted (adapted) by permission of Oxford University Press:
Dennis M. Krüger, Prakash Chandra Rathi, Christopher Pfleger, and Holger Gohlke, CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo)stability, and function. *Nucleic Acids Research*, **2013**, 41:W340-348

I am grateful to the Wiley-VCH Verlag GmbH & Co. for the permission to reprint **Publication VIII**.
Reprinted (adapted) with permission from:
Prakash Chandra Rathi, Christopher Pfleger, Simone Fulle, Doris L. Klein, and Holger Gohlke, Statics of biomacromolecules. in: "Modeling of Molecular Properties", P. Comba (ed.), S. 281-299, Wiley-VCH, Weinheim, 2011.
Copyright (2011) Wiley-VCH Verlag GmbH & Co. KGaA.

12 REFERENCES

1. Schellman, J. A., Macromolecular binding. *Biopolymers* **1975**, 14, 999-1018.
2. Becktel, W. J.; Schellman, J. A., Protein Stability Curves. *Biopolymers* **1987**, 26, 1859-1877.
3. Dill, K. A.; Alonso, D. O. V.; Hutchinson, K., Thermal Stabilities of Globular-Proteins. *Biochemistry* **1989**, 28, 5439-5449.
4. Kumar, S.; Nussinov, R., How do thermophilic proteins deal with heat? *Cell. Mol. Life Sci.* **2001**, 58, 1216-1233.
5. Kumar, S.; Tsai, C. J.; Nussinov, R., Temperature range of thermodynamic stability for the native state of reversible two-state proteins. *Biochemistry* **2003**, 42, 4864-4873.
6. Ravindra, R.; Winter, R., On the temperature-Pressure free-energy landscape of proteins. *Chemphyschem* **2003**, 4, 359-365.
7. Rees, D. C.; Robertson, A. D., Some thermodynamic implications for the thermostability of proteins. *Protein Sci.* **2001**, 10, 1187-1194.
8. Schellman, J. A., The Thermodynamic Stability of Proteins. *Annu. Rev. Biophys. Biophys. Chem.* **1987**, 16, 115-137.
9. Takekiyo, T.; Yoshimura, Y., Relationship between the volume properties and pressure stability of helical-rich proteins. *High Pressure Res.* **2009**, 29, 671-675.
10. Kangur, L.; Timpmann, K.; Freiberg, A., Stability of integral membrane proteins under high hydrostatic pressure: The LH2 and LH3 antenna pigment-protein complexes from photosynthetic bacteria. *J. Phys. Chem. B* **2008**, 112, 7948-7955.
11. Kidman, G.; Park, H.; Northrop, D. B., Pressure stability of proteins at their isoelectric points. *Protein Pept. Lett.* **2004**, 11, 543-546.
12. Meersman, F.; Smeller, L.; Heremans, K., Pressure-assisted cold unfolding of proteins and its effects on the conformational stability compared to pressure and heat unfolding. *High Pressure Res.* **2000**, 19, 653-658.
13. Silva, J. L.; Weber, G., Pressure Stability of Proteins. *Annu. Rev. Phys. Chem.* **1993**, 44, 89-113.
14. Herskovits, T. T.; Jaillet, H.; DeSena, T., On the structural stability and solvent denaturation of proteins. 3. Denaturation by the amides. *J. Biol. Chem.* **1970**, 245, 6511-6517.
15. Herskovits, T. T.; Jaillet, H.; Gadegbeku, B., On the structural stability and solvent denaturation of proteins. II. Denaturation by the ureas. *J. Biol. Chem.* **1970**, 245, 4544-4550.
16. Herskovits, T. T.; Gadegbeku, B.; Jaillet, H., On the structural stability and solvent denaturation of proteins. I. Denaturation by the alcohols and glycols. *J. Biol. Chem.* **1970**, 245, 2588-2598.
17. Alonso, D. O.; Dill, K. A., Solvent denaturation and stabilization of globular proteins. *Biochemistry* **1991**, 30, 5974-5985.
18. Klibanov, A. M., Enzymatic catalysis in anhydrous organic solvents. *Trends Biochem. Sci.* **1989**, 14, 141-144.
19. Finney, J. L.; Gellatly, B. J.; Golton, I. C.; Goodfellow, J., Solvent effects and polar interactions in the structural stability and dynamics of globular proteins. *Biophys. J.* **1980**, 32, 17-33.
20. Pace, C. N.; Trevino, S.; Prabhakaran, E.; Scholtz, J. M., Protein structure, stability and solubility in water and other solvents. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2004**, 359, 1225-1234.
21. Scharnagl, C.; Reif, M.; Friedrich, J., Stability of proteins: Temperature, pressure and the role of the solvent. *BBA-Proteins Proteom.* **2005**, 1749, 187-213.

22. Bruins, M. E.; Janssen, A. E. M.; Boom, R. M., Thermozymes and their applications - A review of recent literature and patents. *Appl. Biochem. Biotechnol.* **2001**, 90, 155-186.
23. Egorova, K.; Antranikian, G., Industrial relevance of thermophilic Archaea. *Curr. Opin. Microbiol.* **2005**, 8, 649-655.
24. Eichler, J., Biotechnological uses of archaeal extremozymes. *Biotechnol. Adv.* **2001**, 19, 261-278.
25. Haki, G. D.; Rakshit, S. K., Developments in industrially important thermostable enzymes: a review. *Bioresour. Technol.* **2003**, 89, 17-34.
26. Kirk, O.; Borchert, T. V.; Fuglsang, C. C., Industrial enzyme applications. *Curr. Opin. Biotechnol.* **2002**, 13, 345-351.
27. Zamost, B. L.; Nielsen, H. K.; Starnes, R. L., Thermostable Enzymes for Industrial Applications. *J. Ind. Microbiol.* **1991**, 8, 71-81.
28. Demirjian, D. C.; Moris-Varas, F.; Cassidy, C. S., Enzymes from extremophiles. *Curr. Opin. Chem. Biol.* **2001**, 5, 144-151.
29. Van den Burg, B., Extremophiles as a source for novel enzymes. *Curr. Opin. Microbiol.* **2003**, 6, 213-218.
30. Polizzi, K. M.; Bommarius, A. S.; Broering, J. M.; Chaparro-Riggers, J. F., Stability of biocatalysts. *Curr. Opin. Chem. Biol.* **2007**, 11, 220-225.
31. Vieille, C.; Zeikus, G. J., Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* **2001**, 65, 1-43.
32. Lorenz, P.; Schleper, C., Metagenome - a challenging source of enzyme discovery. *J. Mol. Catal. B-Enzym.* **2002**, 19, 13-19.
33. Leisola, M.; Turunen, O., Protein engineering: opportunities and challenges. *Appl. Microbiol. Biotechnol.* **2007**, 75, 1225-1232.
34. Eijsink, V. G. H.; Gaseidnes, S.; Borchert, T. V.; van den Burg, B., Directed evolution of enzyme stability. *Biomol. Eng.* **2005**, 22, 21-30.
35. Eijsink, V. G. H.; Bjørk, A.; Gåseidnes, S.; Sirevåg, R.; Synstad, B.; van den Burg, B.; Vriend, G., Rational engineering of enzyme stability. *J. Biotechnol.* **2004**, 113, 105-120.
36. Chaparro Riggers, J. F.; Polizzi, K. M.; Bommarius, A. S., Better library design: data driven protein engineering. *Biotechnol. J.* **2007**, 2, 180-191.
37. Lehmann, M.; Wyss, M., Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Curr. Opin. Biotechnol.* **2001**, 12, 371-375.
38. Razvi, A.; Scholtz, J. M., Lessons in stability from thermophilic proteins. *Protein Sci.* **2006**, 15, 1569-1578.
39. Kumar, S.; Tsai, C. J.; Nussinov, R., Factors enhancing protein thermostability. *Protein Eng.* **2000**, 13, 179-191.
40. Vogt, G.; Woell, S.; Argos, P., Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.* **1997**, 269, 631-643.
41. Gromiha, M. M.; Pathak, M. C.; Saraboji, K.; Ortlund, E. A.; Gaucher, E. A., Hydrophobic environment is a key factor for the stability of thermophilic proteins. *Proteins* **2013**, 81, 715-721.
42. Russell, R. J.; Hough, D. W.; Danson, M. J.; Taylor, G. L., The crystal structure of citrate synthase from the thermophilic archaeon, Thermoplasma acidophilum. *Structure* **1994**, 2, 1157-1167.
43. Querol, E.; PerezPons, J. A.; MozoVillarias, A., Analysis of protein conformational characteristics related to thermostability. *Protein Eng.* **1996**, 9, 265-271.

44. Daniel, R. M.; Cowan, D. A., Biomolecular stability and life at high temperatures. *Cell. Mol. Life Sci.* **2000**, 57, 250-264.
45. Ladenstein, R.; Antranikian, G., Proteins from hyperthermophiles: stability and enzymatic catalysis close to the boiling point of water. *Adv. Biochem. Eng. Biotechnol.* **1998**, 61, 37-85.
46. Sterner, R.; Liebl, W., Thermophilic adaptation of proteins. *Crit. Rev. Biochem. Mol. Biol.* **2001**, 36, 39-106.
47. Vieille, C.; Burdette, D. S.; Zeikus, J. G., Thermozymes. *Biotechnol. Annu. Rev.* **1996**, 2, 1-83.
48. Haney, P. J.; Badger, J. H.; Buldak, G. L.; Reich, C. I.; Woese, C. R.; Olsen, G. J., Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, 96, 3578-3583.
49. Perl, D.; Mueller, U.; Heinemann, U.; Schmid, F. X., Two exposed amino acid residues confer thermostability on a cold shock protein. *Nat. Struct. Biol.* **2000**, 7, 380-383.
50. Lehmann, M.; Loch, C.; Middendorf, A.; Studer, D.; Lassen, S. F.; Pasamontes, L.; van Loon, A. P. G. M.; Wyss, M., The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng.* **2002**, 15, 403-411.
51. Lehmann, M.; Pasamontes, L.; Lassen, S. F.; Wyss, M., The consensus concept for thermostability engineering of proteins. *Bba-Protein Struct. Mol. Enzym.* **2000**, 1543, 408-415.
52. Lehmann, M.; Kostrewa, D.; Wyss, M.; Brugger, R.; D'Arcy, A.; Pasamontes, L.; van Loon, A. P. G. M., From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. *Protein Eng.* **2000**, 13, 49-57.
53. Hoppe, C.; Schomburg, D., Prediction of protein thermostability with a direction- and distance-dependent knowledge-based potential. *Protein Sci.* **2005**, 14, 2682-2692.
54. Bordner, A. J.; Abagyan, R. A., Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins* **2004**, 57, 400-413.
55. Tian, J. A.; Wu, N. F.; Chu, X. Y.; Fan, Y. L., Predicting changes in protein thermostability brought about by single- or multi-site mutations. *BMC Bioinformatics* **2010**, 11.
56. Gilis, D.; Rooman, M., Predicting protein stability changes upon mutation using database-derived potentials: Solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol.* **1997**, 272, 276-290.
57. Korkegian, A.; Black, M. E.; Baker, D.; Stoddard, B. L., Computational thermostabilization of an enzyme. *Science* **2005**, 308, 857-860.
58. Parthiban, V.; Gromiha, M. M.; Schomburg, D., CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* **2006**, 34, W239-W242.
59. Masso, M.; Vaisman, I. I., Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics* **2008**, 24, 2002-2009.
60. Cheng, J. L.; Randall, A.; Baldi, P., Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* **2006**, 62, 1125-1132.
61. Seeliger, D.; de Groot, B. L., Protein Thermostability Calculations Using Alchemical Free Energy Simulations. *Biophys. J.* **2010**, 98, 2309-2316.
62. Reetz, M. T.; Carballeira, J. D.; Vogel, A., Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew. Chem. Int. Ed. Engl.* **2006**, 45, 7745-7751.

63. Radestock, S.; Gohlke, H., Protein rigidity and thermophilic adaptation. *Proteins* **2011**, 79, 1089-1108.
64. Radestock, S.; Gohlke, H., Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.* **2008**, 8, 507-522.
65. Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F., Protein flexibility predictions using graph theory. *Proteins* **2001**, 44, 150-165.
66. Jacobs, D. J.; Thorpe, M. F., Generic rigidity percolation: the pebble game. *Phys. Rev. Lett.* **1995**, 75, 4051-4054.
67. Jacobs, D. J.; Hendrickson, B., An algorithm for two-dimensional rigidity percolation: the pebble game. *J. Comp. Phys.* **1997**, 137, 346-365.
68. Rader, A. J.; Hespenheide, B. M.; Kuhn, L. A.; Thorpe, M. F., Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, 99, 3540-3545.
69. Vihinen, M., Relationship of protein flexibility to thermostability. *Protein Eng.* **1987**, 1, 477-480.
70. Rathi, P. C.; Radestock, S.; Gohlke, H., Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.* **2012**, 159, 135-144.
71. Hernandez, G.; LeMaster, D. M., Reduced temperature dependence of collective conformational opening in a hyperthermophile rubredoxin. *Biochemistry* **2001**, 40, 14384-14391.
72. Hernandez, G.; Jenney, F. E.; Adams, M. W. W.; LeMaster, D. M., Millisecond time scale conformational flexibility in a hyperthermophile protein at ambient temperature. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, 97, 3166-3170.
73. Fitter, J.; Heberle, J., Structural equilibrium fluctuations in mesophilic and thermophilic alpha-amylase. *Biophys. J.* **2000**, 79, 1629-1636.
74. Danciulescu, C.; Ladenstein, R.; Nilsson, L., Dynamic arrangement of ion pairs and individual contributions to the thermal stability of the cofactor-binding domain of glutamate dehydrogenase from Thermotoga maritima. *Biochemistry* **2007**, 46, 8537-8549.
75. Jaenicke, R., Do ultrastable proteins from hyperthermophiles have high or low conformational rigidity? *Proc. Natl. Acad. Sci. U. S. A.* **2000**, 97, 2962-2964.
76. Jaenicke, R.; Böhm, G., The stability of proteins in extreme environments. *Curr. Opin. Struct. Biol.* **1998**, 8, 738-748.
77. Kalimeri, M.; Rahaman, O.; Melchionna, S.; Sterpone, F., How Conformational Flexibility Stabilizes the Hyperthermophilic Elongation Factor G-Domain. *J. Phys. Chem. B.* **2013**, 117, 13775-13785.
78. Basu, S.; Sen, S., Do Homologous Thermophilic-Mesophilic Proteins Exhibit Similar Structures and Dynamics at Optimal Growth Temperatures? A Molecular Dynamics Simulation Study. *J. Chem. Inf. Model.* **2013**, 53, 423-434.
79. Oyeyemi, O. A.; Sours, K. M.; Lee, T.; Kohen, A.; Resing, K. A.; Ahn, N. G.; Klinman, J. P., Comparative Hydrogen-Deuterium Exchange for a Mesophilic vs Thermophilic Dihydrofolate Reductase at 25 degrees C: Identification of a Single Active Site Region with Enhanced Flexibility in the Mesophilic Protein. *Biochemistry* **2011**, 50, 8251-8260.
80. Marcos, E.; Jimenez, A.; Crehuet, R., Dynamic Fingerprints of Protein Thermostability Revealed by Long Molecular Dynamics. *J. Chem. Theory Comput.* **2012**, 8, 1129-1142.
81. Ó'Fágáin, C. Engineering protein stability. In *Protein Chromatography*, Walls, D.; Loughran, S. T., Eds.; Springer: 2011, pp 103-136.
82. Kamal, M. Z.; Ahmad, S.; Molugu, T. R.; Vijayalakshmi, A.; Deshmukh, M. V.; Sankaranarayanan, R.; Rao, N. M., In Vitro Evolved Non-Aggregating and

Thermostable Lipase: Structural and Thermodynamic Investigation. *J. Mol. Biol.* **2011**, 413, 726-741.
83. Augustyniak, W.; Brzezinska, A. A.; Pijning, T.; Wienk, H.; Boelens, R.; Dijkstra, B. W.; Reetz, M. T., Biophysical characterization of mutants of Bacillus subtilis lipase evolved for thermostability: factors contributing to increased activity retention. *Protein Sci.* **2012**, 21, 487-497.
84. Kamal, M. Z.; Mohammad, T. A. S.; Krishnamoorthy, G.; Rao, N. M., Role of Active Site Rigidity in Activity: MD Simulation and Fluorescence Study on a Lipase Mutant. *PLoS ONE* **2012**, 7, e35188.
85. Kamal, M. Z.; Ahmad, S.; Yedavalli, P.; Rao, N. M., Stability curves of laboratory evolved thermostable mutants of a Bacillus subtilis lipase. *BBA-Proteins Proteom.* **2010**, 1804, 1850-1856.
86. Acharya, P.; Rajakumara, E.; Sankaranarayanan, R.; Rao, N. M., Structural basis of selection and thermostability of laboratory evolved Bacillus subtilis lipase. *J. Mol. Biol.* **2004**, 341, 1271-1281.
87. Ahmad, S.; Rao, N. M., Thermally denatured state determines refolding in lipase: Mutational analysis. *Protein Sci.* **2009**, 18, 1183-1196.
88. Ahmad, S.; Kamal, M. Z.; Sankaranarayanan, R.; Rao, N. M., Thermostable Bacillus subtilis lipases: In vitro evolution and structural insight. *J. Mol. Biol.* **2008**, 381, 324-340.
89. Booth, D. R.; Sunde, M.; Bellotti, V.; Robinson, C. V.; Hutchinson, W. L.; Fraser, P. E.; Hawkins, P. N.; Dobson, C. M.; Radford, S. E.; Blake, C. C. F.; Pepys, M. B., Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature* **1997**, 385, 787-793.
90. Verheul, M.; Roefs, S. P. F. M.; de Kruif, K. G., Kinetics of heat-induced aggregation of beta-lactoglobulin. *J. Agric. Food Chem.* **1998**, 46, 896-903.
91. Wang, W.; Nema, S.; Teagarden, D., Protein aggregation-Pathways and influencing factors. *Int. J. Pharm.* **2010**, 390, 89-99.
92. Vermeer, A. W. P.; Norde, W., The thermal stability of immunoglobulin: Unfolding and aggregation of a multi-domain protein. *Biophys. J.* **2000**, 78, 394-404.
93. Duy, C.; Fitter, J., Thermostability of irreversible unfolding alpha-amylases analyzed by unfolding kinetics. *J. Biol. Chem.* **2005**, 280, 37360-37365.
94. Volkin, D. B.; Klibanov, A. M., Alterations in the structure of proteins that cause their irreversible inactivation. *Dev. Biol. Stand.* **1992**, 74, 73-80.
95. Parsell, D. A.; Sauer, R. T., The Structural Stability of a Protein Is an Important Determinant of Its Proteolytic Susceptibility in Escherichia-Coli. *J. Biol. Chem.* **1989**, 264, 7590-7595.
96. Kubbutat, M. H. G.; Vousden, K. H., Proteolytic cleavage of human p53 by calpain: A potential regulator of protein stability. *Mol. Cell. Biol.* **1997**, 17, 460-468.
97. Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A., Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* **2008**, 51, 6237-6255.
98. Daniel, R. M.; Dunn, R. V.; Finney, J. L.; Smith, J. C., The role of dynamics in enzyme activity. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, 32, 69-92.
99. Ishima, R.; Torchia, D. A., Protein dynamics from NMR. *Nat. Struct. Mol. Biol.* **2000**, 7, 740-743.
100. Englander, S. W.; Kallenbach, N. R., Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Q. Rev. Biophys.* **1983**, 16, 521-655.

101. Scholtz, J. M.; Robertson, A. D., Hydrogen exchange techniques. *Methods Mol. Biol.* **1995**, 40, 291-311.
102. Weiss, S., Fluorescence spectroscopy of single biomolecules. *Science* **1999**, 283, 1676-1683.
103. Zhang, X. J.; Wozniak, J. A.; Matthews, B. W., Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozymes. *J. Mol. Biol.* **1995**, 250, 527-552.
104. Frank, J.; Agrawal, R. K., A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature* **2000**, 406, 318-322.
105. Feller, G.; d'Amico, D.; Gerday, C., Thermodynamic stability of a cold-active alpha-amylase from the Antarctic bacterium Alteromonas haloplanctis. *Biochemistry* **1999**, 38, 4613-4619.
106. Fontana, A.; Filippis, V. D.; Laureto, P. P. d.; Scaramella, E.; Zambonin, M. Rigidity of thermophilic enzymes. In *Progress in Biotechnology*, A. Ballesteros, F. J. P. J. L. I.; Halling, P. J., Eds.; Elsevier: 1998; Vol. 15, pp 277-294.
107. Hollien, J.; Marqusee, S., Structural distribution of stability in a thermophilic enzyme. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, 96, 13674-13678.
108. Zavodszky, P.; Kardos, J.; Svingor, A.; Petsko, G. A., Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, 95, 7406-7411.
109. Karplus, M.; McCammon, J. A., Molecular dynamics simulations of biomolecules. *Nat. Struct. Mol. Biol.* **2002**, 9, 646-652.
110. Case, D. A., Normal mode analysis of protein dynamics. *Curr. Opin. Struct. Biol.* **1994**, 4, 285-290.
111. Taylor, T. J.; Vaisman, I. I., Discrimination of thermophilic and mesophilic proteins. *BMC Struct. Biol.* **2010**, 10(Suppl 1), S5.
112. Závodszky, P.; Kardos, J.; Svingor, Á.; Petsko, G. A., Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, 95, 7406-7411.
113. Machius, M.; Declerck, N.; Huber, R.; Wiegand, G., Kinetic stabilization of Bacillus licheniformis α-amylase through introduction of hydrophobic residues at the surface. *J. Biol. Chem.* **2003**, 278, 11546-11553.
114. Whitmore, L.; Wallace, B. A., Protein secondary structure analyses from circular dichroism spectroscopy: Methods and reference databases. *Biopolymers* **2008**, 89, 392-400.
115. Lasch, J.; Bessmertnaya, L.; Kozlov, L. V.; Antonov, V. K., Thermal-Stability of Immobilized Enzymes Circular-Dichroism, Fluorescence and Kinetic Measurements of Alpha-Chymotrypsin Attached to Soluble Carriers. *Eur. J. Biochem.* **1976**, 63, 591-598.
116. Zheng, L.; Brennan, J. D., Measurement of intrinsic fluorescence to probe the conformational flexibility and thermodynamic stability of a single tryptophan protein entrapped in a sol-gel derived glass matrix. *Analyst* **1998**, 123, 1735-1744.
117. Meeker, A. K.; GarciaMoreno, B.; Shortle, D., Contributions of the ionizable amino acids to the stability of staphylococcal nuclease. *Biochemistry* **1996**, 35, 6443-6449.
118. Uniewicz, K. A.; Ori, A.; Xu, R. Y.; Ahmed, Y.; Wilkinson, M. C.; Fernig, D. G.; Yates, E. A., Differential Scanning Fluorimetry Measurement of Protein Stability Changes upon Binding to Glycosaminoglycans: A Screening Test for Binding Specificity. *Anal. Chem.* **2010**, 82, 3796-3802.
119. Senisterra, G. A.; Finerty, P. J., High throughput methods of assessing protein stability and aggregation. *Mol. Biosyst.* **2009**, 5, 217-223.

120. Boeckler, F. M.; Joerger, A. C.; Jaggi, G.; Rutherford, T. J.; Veprintsev, D. B.; Fersht, A. R., Targeted rescue of a destabilized mutant of p53 by an in silico screened drug. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, 105, 10360-10365.
121. Niesen, F. H.; Berglund, H.; Vedadi, M., The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat. Protoc.* **2007**, 2, 2212-2221.
122. Lavinder, J. J.; Hari, S. B.; Sullivan, B. J.; Magliery, T. J., High-throughput thermal scanning: a general, rapid dye-binding thermal shift screen for protein engineering. *J. Am. Chem. Soc.* **2009**, 131, 3794-3795.
123. Graziano, G.; Catanzano, F.; Giancola, C.; Barone, G., DSC study of the thermal stability of S-protein and S-peptide/S-protein complexes. *Biochemistry* **1996**, 35, 13386-13392.
124. Bruylants, G.; Wouters, J.; Michaux, C., Differential scanning calorimetry in life science: Thermodynamics, stability, molecular recognition and application in drug design. *Curr. Med. Chem.* **2005**, 12, 2011-2020.
125. Bernal, V.; Jelen, P., Thermal-Stability of Whey Proteins - a Calorimetric Study. *J. Dairy Sci.* **1985**, 68, 2847-2852.
126. Pace, C. N., Measuring and Increasing Protein Stability. *Trends Biotechnol.* **1990**, 8, 93-98.
127. Mozhaev, V. V., Mechanism-Based Strategies for Protein Thermostabilization. *Trends Biotechnol.* **1993**, 11, 88-95.
128. Anbar, M.; Gul, O.; Lamed, R.; Sezerman, U. O.; Bayer, E. A., Improved Thermostability of Clostridium thermocellum Endoglucanase Cel8A by Using Consensus-Guided Mutagenesis. *Appl. Environ. Microbiol.* **2012**, 78, 3458-3464.
129. Wakarchuk, W. W.; Sung, W. L.; Campbell, R. L.; Cunningham, A.; Watson, D. C.; Yaguchi, M., Thermostabilization of the Bacillus circulans xylanase by the introduction of disulfide bonds. *Protein Eng.* **1994**, 7, 1379-1386.
130. Bommarius, A. S.; Paye, M. F., Stabilizing biocatalysts. *Chem. Soc. Rev.* **2013**, 42, 6534-6565.
131. Jochens, H.; Aerts, D.; Bornscheuer, U. T., Thermostabilization of an esterase by alignment-guided focussed directed evolution. *Protein Eng. Des. Sel.* **2010**, 23, 903-909.
132. Zhang, Z. G.; Yi, Z. L.; Pei, X. Q.; Wu, Z. L., Improving the thermostability of Geobacillus stearothermophilus xylanase XT6 by directed evolution and site-directed mutagenesis. *Bioresour. Technol.* **2010**, 101, 9272-9278.
133. Chow, J. Y.; Xue, B.; Lee, K. H.; Tung, A.; Wu, L.; Robinson, R. C.; Yew, W. S., Directed Evolution of a Thermostable Quorum-quenching Lactonase from the Amidohydrolase Superfamily. *J. Biol. Chem.* **2010**, 285, 40911-40920.
134. Kotzia, G. A.; Labrou, N. E., Engineering thermal stability of L-asparaginase by in vitro directed evolution. *FEBS J.* **2009**, 276, 1750-1761.
135. Akbulut, N.; Ozturk, M. T.; Pijning, T.; Ozturk, S. I.; Gumusel, F., Improved activity and thermostability of Bacillus pumilus lipase by directed evolution. *J. Biotechnol.* **2013**, 164, 123-129.
136. Nakazawa, H.; Okada, K.; Onodera, T.; Ogasawara, W.; Okada, H.; Morikawa, Y., Directed evolution of endoglucanase III (Cel12A) from Trichoderma reesei. *Appl. Microbiol. Biotechnol.* **2009**, 83, 649-657.
137. Steffler, F.; Guterl, J. K.; Sieber, V., Improvement of thermostable aldehyde dehydrogenase by directed evolution for application in Synthetic Cascade Biomanufacturing. *Enzyme Microb. Technol.* **2013**, 53, 307-314.

138. Vazquez-Figueroa, E.; Chaparro-Riggers, J.; Bommarius, A. S., Development of a thermostable glucose dehydrogenase by a structure-guided consensus concept. *Chembiochem* **2007**, 8, 2295-2301.
139. Tian, J. A.; Wang, P.; Gao, S.; Chu, X. Y.; Wu, N. F.; Fan, Y. L., Enhanced thermostability of methyl parathion hydrolase from Ochrobactrum sp. M231 by rational engineering of a glycine to proline mutation. *FEBS J.* **2010**, 277, 4901-4908.
140. Leemhuis, H.; Rozeboom, H. J.; Dijkstra, B. W.; Dijkhuizen, L., Improved thermostability of Bacillus circulans cyclodextrin glycosyltransferase by the introduction of a salt bridge. *Proteins* **2004**, 54, 128-134.
141. Kaneko, H.; Minagawa, H.; Shimada, J., Rational design of thermostable lactate oxidase by analyzing quaternary structure and prevention of deamidation. *Biotechnol. Lett.* **2005**, 27, 1777-1784.
142. Hirose, S.; Kawamura, Y.; Mori, M.; Yokota, K.; Noguchi, T.; Goshima, N., Development and evaluation of data-driven designed tags (DDTs) for controlling protein solubility. *New Biotechnol.* **2011**, 28, 225-231.
143. Synstad, B.; Gaseidnes, S.; van Aalten, D. M. F.; Vriend, G.; Nielsen, J. E.; Eijsink, V. G. H., Mutational and computational analysis of the role of conserved residues in the active site of a family 18 chitinase. *Eur. J. Biochem.* **2004**, 271, 253-262.
144. Benedix, A.; Becker, C. M.; de Groot, B. L.; Caflisch, A.; Bockmann, R. A., Predicting free energy changes using structural ensembles. *Nat. Methods* **2009**, 6, 3-4.
145. Pokala, N.; Handel, T. M., Energy functions for protein design: Adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J. Mol. Biol.* **2005**, 347, 203-227.
146. Guerois, R.; Nielsen, J. E.; Serrano, L., Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.* **2002**, 320, 369-387.
147. Capriotti, E.; Fariselli, P.; Casadio, R., I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **2005**, 33, W306-W310.
148. Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D., Protein structure prediction using Rosetta. *Methods Enzymol.* **2004**, 383, 66-93.
149. Potapov, V.; Cohen, M.; Inbar, Y.; Schreiber, G., Protein structure modelling and evaluation based on a 4-distance description of side-chain interactions. *BMC Bioinformatics* **2010**, 11.
150. Potapov, V.; Cohen, M.; Schreiber, G., Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng. Des. Sel.* **2009**, 22, 553-560.
151. Lazaridis, T.; Karplus, M., Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **2000**, 10, 139-145.
152. Russ, W. P.; Ranganathan, R., Knowledge-based potential functions in protein design. *Curr. Opin. Struct. Biol.* **2002**, 12, 447-452.
153. Smola, A. J.; Schölkopf, B., A tutorial on support vector regression. *Stat. Comput.* **2004**, 14, 199-222.
154. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, 118, 11225-11236.
155. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, 112, 6127-6129.

156. Van Gunsteren, W.; Billeter, S.; Eising, A.; Hünenberger, P.; Krüger, P.; Mark, A.; Scott, W.; Tironi, I., Biomolecular simulations: the GROMOS96 manual and user guide. 1996. *Zürich: VdF Hochschulverlag ETHZ*.
157. Honig, B.; Sharp, K.; Yang, A. S., Macroscopic models of aqueous solutions: biological and chemical applications. *J. Phys. Chem.* **1993**, 97, 1101-1109.
158. Karplus, M.; Kushick, J. N., Method for estimating the configurational entropy of macromolecules. *Macromolecules* **1981**, 14, 325-332.
159. de Groot, B. L.; van Aalten, D. M.; Scheek, R. M.; Amadei, A.; Vriend, G.; Berendsen, H. J., Prediction of protein conformational freedom from distance constraints. *Proteins* **1997**, 29, 240-251.
160. Fischer, A.; Seitz, T.; Lochner, A.; Sterner, R.; Merkl, R.; Bocola, M., A Fast and Precise Approach for Computational Saturation Mutagenesis and its Experimental Validation by Using an Artificial (beta alpha)(8)-Barrel Protein. *Chembiochem* **2011**, 12, 1544-1550.
161. Gerber, P. R.; Muller, K., Mab, a Generally Applicable Molecular-Force Field for Structure Modeling in Medicinal Chemistry. *J. Comput. Aided Mol. Des.* **1995**, 9, 251-268.
162. Gribenko, A. V.; Patel, M. M.; Liu, J.; McCallum, S. A.; Wang, C. Y.; Makhatadze, G. I., Rational stabilization of enzymes by computational redesign of surface charge-charge interactions. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, 106, 2601-2606.
163. Matthew, J. B.; Gurd, F. R. N., Calculation of Electrostatic Interactions in Proteins. *Methods Enzymol.* **1986**, 130, 413-436.
164. Ibarra-Molero, B.; Sanchez-Ruiz, J. M., Genetic algorithm to design stabilizing surface-charge distributions in proteins. *J. Phys. Chem. B* **2002**, 106, 6609-6613.
165. Tanford, C.; Kirkwood, J. G., Theory of Protein Titration Curves. I. General Equations for Impenetrable Spheres. *J. Am. Chem. Soc.* **1957**, 79, 5333-5339.
166. Berezovsky, I. N.; Shakhnovich, E. I., Physics and evolution of thermophilic adaptation. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, 102, 12742-12747.
167. Bae, E.; Bannen, R. M.; Phillips, G. N., Jr., Bioinformatic method for protein thermal stabilization by structural entropy optimization. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, 105, 9594-9597.
168. Chan, C. H.; Liang, H. K.; Hsiao, N. W.; Ko, M. T.; Lyu, P. C.; Hwang, J. K., Relationship between local structural entropy and protein thermostability. *Proteins* **2004**, 57, 684-691.
169. Hespenheide, B. M.; Rader, A. J.; Thorpe, M. F.; Kuhn, L. A., Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graph. Model.* **2002**, 21, 195-207.
170. Rathi, P. C.; Pfleger, C.; Fulle, S.; Klein, D. L.; Gohlke, H. Statics of biomacromolecules. In *Molecular Modeling*, Comba, P., Ed.; Wiley-VCH: Weinheim, 2011, pp 281-299.
171. Maxwell, J. C., On the calculation of the equilibrium and stiffness of frames. *Philos. Mag. Series 4* **1864**, 27, 294-299.
172. Laman, G., On graphs and rigidity of plane skeletal structures. *J. Eng. Math.* **1970**, 4, 331-340.
173. Tay, T. S.; Whiteley, W., Recent advances in the generic ridigity of structures. *Struct. Topol.* **1984**, 9, 31-38.
174. Katoh, N.; Tanigawa, S. A proof of the molecular conjecture. In Proceedings of the 25th annual symposium on Computational geometry, Aarhus, Denmark, Jun 8-10, 2009, 2009; ACM: Aarhus, Denmark, 2009; pp 296-305.

175. Jacobs, D. J., Generic rigidity in three-dimensional bond-bending networks. *J. Phys. A Math. Gen.* **1998**, 31, 6653-6668.
176. Hespenheide, B. M.; Jacobs, D. J.; Thorpe, M. F., Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. *J. Phys. Condens. Matter* **2004**, 16, S5055-S5064.
177. Whiteley, W., Counting out to the flexibility of molecules. *Phys. Biol.* **2005**, 2, S116-S126.
178. Stevens, M.; Boolchand, P.; Hernandez, J. G., Universal structural phase transition in network glasses. *Phys. Rev. B Condens. Matter Mater. Phys.* **1985**, 31, 981-991.
179. Thorpe, M. F.; Jacobs, D. J.; Chubynsky, M. V.; Phillips, J. C., Self-organization in network glasses. *J. Non-Cryst. Solids* **2000**, 266-269, 859-866.
180. Wang, Y.; Wells, J.; Georgiev, D. G.; Boolchand, P.; Jackson, K.; Micoulaut, M., Sharp rigid to floppy phase transition induced by dangling ends in a network glass. *Phys. Rev. Lett.* **2001**, 87, 185503.
181. Dahiyat, B. I.; Gordon, D. B.; Mayo, S. L., Automated design of the surface positions of protein helices. *Protein Sci.* **1997**, 6, 1333-1337.
182. Makhatadze, G. I.; Privalov, P. L., Energetics of protein structure. *Adv. Protein Chem.* **1995**, 47, 307-425.
183. Privalov, P. L.; Gill, S. J., Stability of protein structure and hydrophobic interaction. *Adv. Protein Chem.* **1988**, 39, 191-234.
184. Schellman, J. A., Temperature, stability, and the hydrophobic interaction. *Biophys. J.* **1997**, 73, 2960-2964.
185. Gohlke, H.; Kuhn, L. A.; Case, D. A., Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins* **2004**, 56, 322-337.
186. Mamonova, T.; Hespenheide, B.; Straub, R.; Thorpe, M. F.; Kurnikova, M., Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.* **2005**, 2, S137-S147.
187. Taverna, D. M.; Goldstein, R. A., Why are proteins marginally stable? *Proteins* **2002**, 46, 105-109.
188. Pfleger, C.; Gohke, H., Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. *Structure* **2013**, 21, 1725-1734.
189. Pfleger, C.; Radestock, S.; Schmidt, E.; Gohlke, H., Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comp. Chem.* **2013**, 34, 220-233.
190. Andraud, C.; Beghdadi, A.; Lafait, J., Entropic analysis of random morphologies. *Physica A* **1994**, 207, 208-212.
191. Pfleger, C.; Rathi, P. C.; Klein, D. L.; Radestock, S.; Gohlke, H., Constraint Network Analysis (CNA): A Python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. *J. Chem. Inf. Model.* **2013**, 53, 1007-1015.
192. Fox, N.; Jagodzinski, F.; Li, Y.; Streinu, I., KINARI-Web: a server for protein rigidity analysis. *Nucleic Acids Res.* **2011**, 39, W177-W183.
193. McDonald, I. K.; Thornton, J. M., Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **1994**, 238, 777-793.
194. Cheatham 3rd, T. E.; Cieplak, P.; Kollman, P. A., A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* **1999**, 16, 845-862.
195. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force

field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, 117, 5179-5197.
196. Jagodzinski, F.; Hardy, J.; Streinu, I., Using Rigidity Analysis to Probe Mutation-Induced Structural Changes in Proteins. *J. Bioinform. Comput. Biol.* **2012**, 10, 1242010.
197. Kumar, M. D.; Bava, K. A.; Gromiha, M. M.; Prabakaran, P.; Kitajima, K.; Uedaira, H.; Sarai, A., ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* **2006**, 34, D204-D206.
198. Jacobs, D. J.; Dallakyan, S.; Wood, G. G.; Heckathorne, A., Network rigidity at finite temperature: relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2003**, 68, 061109.
199. Livesay, D. R.; Dallakyan, S.; Wood, G. G.; Jacobs, D. J., A flexible approach for understanding protein stability. *FEBS Lett.* **2004**, 576, 468-476.
200. Jacobs, D. J.; Dallakyan, S., Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys. J.* **2005**, 88, 903-915.
201. Vorov, O. K.; Livesay, D. R.; Jacobs, D. J., Helix/coil nucleation: a local response to global demands. *Biophys. J.* **2009**, 97, 3000-3009.
202. Vorov, O. K.; Livesay, D. R.; Jacobs, D. J., Nonadditivity in conformational entropy upon molecular rigidification reveals a universal mechanism affecting folding cooperativity. *Biophys. J.* **2011**, 100, 1129-1138.
203. Livesay, D. R.; Jacobs, D. J., Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins* **2006**, 62, 130-143.
204. Mottonen, J. M.; Jacobs, D. J.; Livesay, D. R., Allosteric response is both conserved and variable across three CheY orthologs. *Biophys. J.* **2010**, 99, 2245-2254.
205. Verma, D.; Jacobs, D. J.; Livesay, D. R., Changes in Lysozyme Flexibility upon Mutation Are Frequent, Large and Long-Ranged. *PLoS Comput. Biol.* **2012**, 8, e1002409.
206. González, L. C.; Wang, H.; Livesay, D. R.; Jacobs, D. J., Calculating Ensemble Averaged Descriptions of Protein Rigidity without Sampling. *PLoS ONE* **2012**, 7, e29176.
207. Rathi, P. C.; Höffken, H. W.; Gohlke, H., Quality matters: extension of clusters of residues with good hydrophobic contacts stabilize (hyper)thermophilic proteins. *J. Chem. Inf. Model.* **2014**, 54, 355-361.
208. Kruger, D. M.; Rathi, P. C.; Pfleger, C.; Gohlke, H., CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo-)stability, and function. *Nucleic Acids Res.* **2013**, 41, W340-W348.
209. Beazley, D. M., Automated scientific software scripting with SWIG. *Future Gener. Comp. Sy.* **2003**, 19, 599-609.
210. Radford, S. E.; Dobson, C. M.; Evans, P. A., The Folding of Hen Lysozyme Involves Partially Structured Intermediates and Multiple Pathways. *Nature* **1992**, 358, 302-307.
211. Matagne, A.; Radford, S. E.; Dobson, C. M., Fast and slow tracks in lysozyme folding: Insight into the role of domains in the folding process. *J. Mol. Biol.* **1997**, 267, 1068-1074.
212. Haliloglu, T.; Bahar, I., Structure-based analysis of protein dynamics: Comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data. *Proteins* **1999**, 37, 654-667.
213. Radford, S. E.; Buck, M.; Topping, K. D.; Dobson, C. M.; Evans, P. A., Hydrogen-Exchange in Native and Denatured States of Hen Egg-White Lysozyme. *Proteins* **1992**, 14, 237-248.

214. McCammon, J. A.; Gelin, B. R.; Karplus, M.; Wolynes, P. G., The hinge-bending mode in lysozyme. *Nature* **1976**, 262, 325-326.
215. Smith, L. J.; Sutcliffe, M. J.; Redfield, C.; Dobson, C. M., Structure of Hen Lysozyme in Solution. *J. Mol. Biol.* **1993**, 229, 930-944.
216. van Aalten, D. M.; Amadei, A.; Linssen, A. B.; Eijsink, V. G.; Vriend, G.; Berendsen, H. J., The essential dynamics of thermolysin: confirmation of the hinge-bending motion and comparison of simulations in vacuum and water. *Proteins* **1995**, 22, 45-54.
217. Veltman, O. R.; Eijsink, V. G.; Vriend, G.; de Kreij, A.; Venema, G.; van den Burg, B., Probing catalytic hinge bending motions in thermolysin-like proteases by glycine-alanine mutations. *Biochemistry* **1998**, 37, 5305-5311.
218. Imanaka, T.; Shibazaki, M.; Takagi, M., A new way of enhancing the thermostability of proteases. *Nature* **1986**, 324, 695-697.
219. Van den Burg, B.; Dijkstra, B. W.; Vriend, G.; Vandervinne, B.; Venema, G.; Eijsink, V. G. H., Protein stabilization by hydrophobic interactions at the surface. *Eur. J. Biochem.* **1994**, 220, 981-985.
220. Van den Burg, B.; Vriend, G.; Veltman, O. R.; Venema, G.; Eijsink, V. G., Engineering an enzyme to resist boiling. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, 95, 2056-2060.
221. Veltman, O. R.; Vriend, G.; Middelhoven, P. J.; van den Burg, B.; Venema, G.; Eijsink, V. G., Analysis of structural determinants of the stability of thermolysin-like proteases by molecular modelling and site-directed mutagenesis. *Protein Eng.* **1996**, 9, 1181-1189.
222. Jaeger, K. E.; Eggert, T., Lipases for biotechnology. *Curr. Opin. Biotechnol.* **2002**, 13, 390-397.
223. Jaeger, K. E.; Ransac, S.; Dijkstra, B. W.; Colson, C.; Vanheuvel, M.; Misset, O., Bacterial Lipases. *FEMS Microbiol. Rev.* **1994**, 15, 29-63.
224. Droge, M. J.; Boersma, Y. L.; van Pouderoyen, G.; Vrenken, T. E.; Ruggeberg, C. J.; Reetz, M. T.; Dijkstra, B. W.; Quax, W. J., Directed evolution of Bacillus subtilis lipase A by use of enantiomeric phosphonate inhibitors: crystal structures and phage display selection. *Chembiochem* **2006**, 7, 149-157.
225. Kawasaki, K.; Kondo, H.; Suzuki, M.; Ohgiya, S.; Tsuda, S., Alternate conformations observed in catalytic serine of Bacillus subtilis lipase determined at 1.3 A resolution. *Acta Crystallogr. D Biol. Crystallogr.* **2002**, 58, 1168-1174.
226. van Pouderoyen, G.; Eggert, T.; Jaeger, K. E.; Dijkstra, B. W., The crystal structure of Bacillus subtilis lipase: a minimal alpha/beta hydrolase fold enzyme. *J. Mol. Biol.* **2001**, 309, 215-226.
227. Rajakumara, E.; Acharya, P.; Ahmad, S.; Sankaranaryanan, R.; Rao, N. M., Structural basis for the remarkable stability of Bacillus subtilis lipase (Lip A) at low pH. *BBA-Proteins Proteom.* **2008**, 1784, 302-311.
228. Abraham, T.; Pack, S. P.; Yoo, Y. J., Stabilization of Bacillus subtilis Lipase A by increasing the residual packing. *Biocatal. Biotransfor.* **2005**, 23, 217-224.
229. Beermann, B.; Guddorf, J.; Boehm, K.; Albers, A.; Kolkenbrock, S.; Fetzner, S.; Hinz, H. J., Stability, unfolding, and structural changes of cofactor-free 1H-3-hydroxy-4-oxoquinaldine 2,4-dioxygenase. *Biochemistry* **2007**, 46, 4241-4249.
230. Hung, H. C.; Chang, G. G., Multiple unfolding intermediates of human placental alkaline phosphatase in equilibrium urea denaturation. *Biophys. J.* **2001**, 81, 3456-3471.
231. Seewald, M. J.; Pichumani, K.; Stowell, C.; Tibbals, B. V.; Regan, L.; Stone, M. J., The role of backbone conformational heat capacity in protein stability: Temperature dependent dynamics of the B1 domain of Streptococcal protein G. *Protein Sci.* **2000**, 9, 1177-1193.

232. Stone, M. J.; Gupta, S.; Snyder, N.; Regan, L., Comparison of protein backbone entropy and beta-sheet stability: NMR-derived dynamics of protein G B1 domain mutants. *J. Am. Chem. Soc.* **2001**, 123, 185-186.
233. Melo, F.; Devos, D.; Depiereux, E.; Feytmans, E., ANOLEA: a www server to assess protein structures. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1997**, 5, 187-190.
234. Melo, F.; Feytmans, E., Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.* **1997**, 267, 207-222.
235. Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; Heger, A.; Holm, L.; Sonnhammer, E. L. L.; Eddy, S. R.; Bateman, A.; Finn, R. D., The Pfam protein families database. *Nucleic Acids Res.* **2012**, 40, D290-D301.
236. Steipe, B.; Schiller, B.; Pluckthun, A.; Steinbacher, S., Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. *J. Mol. Biol.* **1994**, 240, 188-192.
237. Vihinen, M.; Ollikka, P.; Niskanen, J.; Meyer, P.; Suominen, I.; Karp, M.; Holm, L.; Knowles, J.; Mantsala, P., Site-Directed Mutagenesis of a Thermostable Alpha-Amylase from Bacillus-Stearothermophilus - Putative Role of 3 Conserved Residues. *J. Biochem. (Tokyo)* **1990**, 107, 267-272.
238. Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L., Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **2009**, 77, 778-795.
239. Tielen, P.; Kuhn, H.; Rosenau, F.; Jaeger, K. E.; Flemming, H. C.; Wingender, J., Interaction between extracellular lipase LipA and the polysaccharide alginate of Pseudomonas aeruginosa. *BMC Microbiol.* **2013**, 13.
240. Eijsink, V. G. H.; Vriend, G.; Vandervinne, B.; Hazes, B.; Vandenburg, B.; Venema, G., Effects of Changing the Interaction between Subdomains on the Thermostability of Bacillus Neutral Proteases. *Proteins* **1992**, 14, 224-236.
241. Cho, A. R.; Yoo, S. K.; Kim, E. J., Cloning, sequencing and expression in Escherichia coli of a thermophilic lipase from Bacillus thermoleovorans ID-1. *FEMS Microbiol. Lett.* **2000**, 186, 235-238.
242. Sheridan, R. P.; Lee, R. H.; Peters, N.; Allen, L. C., Hydrogen-bond cooperativity in protein secondary structure. *Biopolymers* **1979**, 18, 2451-2458.

13 CURRICULUM VITAE

Personal Information

Name	Prakash Chandra Rathi
Date of birth	13/09/1982 in Pokaran, India

Education and Professional Experience

04/2010 – to date	**PhD Student** at Heinrich-Heine-University Düsseldorf, Germany. Supervisor: Prof. Dr. Holger Gohlke Topic: Development of computational approaches for knowledge-driven protein engineering aimed at improving thermostability
08/2011 – 10/2011	**Intern** at Department of Modeling and Formulation Research, BASF, Ludwigshafen. Supervisor: Dr. Wolfgang Höffken Topic: Identifying the determinants of thermostability of proteins
07/2009 – 03/2010	**Research fellow** at Department of Medicinal Chemistry, National Institute of Pharmaceutical Education and Research, S.A.S. Nagar, India with Prof. Dr. P. V. Bharatam. Topic: 3D models of CYPs – Prediction of cytochrome P450-mediated metabolism of drugs
07/2007 – 06/2009	**Master of science (Pharm.)** in Pharmacoinformatics at National Institute of Pharmaceutical Education and Research, S.A.S. Nagar, India. Supervisor: Prof. Dr. P. V. Bharatam Topic: Cytochrome P450 mediated metabolism of drugs: Molecular docking and semi-empirical QM analyses
07/2003 – 06/2007	**Bachelor in Pharmacy** at Mohan Lal Sukhadia University, Udaipur, Rajasthan, India.

Publications

Rathi, PC, Hoeffken, HW, Gohlke, H, *Quality matters: Extension of clusters of residues with good hydrophobic contacts stabilize (hyper)thermophilic proteins*. J. Chem. Inf. Model. **2014**, 54, 355–361.

Krüger, DM*, Rathi, PC*, Pfleger, C, Gohlke, H, *CNA web server: Rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo)stability, and function*. Nucleic Acid Res. **2013**, 41, W340-348. [* Co-first authors]

Pfleger, C*, Rathi, PC*, Klein, D, Radestock, S, Gohlke, H, *Constraint Network Analysis (CNA): A Python software package for efficiently linking biomolecular structure, flexibility, (thermo-)stability, and function.* J. Chem. Inf. Model. **2013**, 53, 1007-1015. [* Co-first authors]

Rathi, PC, Radestock, S, Gohlke, H, *Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio*. J. Biotechnol. **2012**, 159, 135-144.

Rathi, PC, Pfleger, C, Fulle, S, Klein, DL, Gohlke, H, *Statics of biomacromolecules* in "Modeling of Molecular Properties", P. Comba (ed.), S. 281-299, Wiley-VCH, Weinheim, **2011**.

Dixit, VA, Rathi, PC, Bharatam, PV, *Intramolecular dihydrogen bond: A new perspective in Lewis acid catalyzed nucleophilic epoxide ring opening reaction.* J. Mol. Struct.: THEOCHEM. **2010,** 962, 97-100.

14 PUBLICATIONS

Publication I

Quality matters: Extension of clusters of residues with good hydrophobic contacts stabilize (hyper)thermophilic proteins

Prakash Chandra Rathi, Hans Wolfgang Höffken, and Holger Gohlke
Journal of Chemical Information and Modeling, **2014**, 54: 355-361.
2012 impact factor: 4.20; contribution: 70%

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Letter

pubs.acs.org/jcim

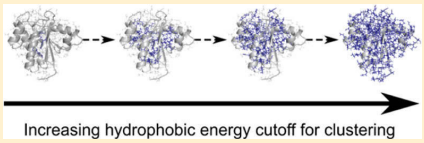
Quality Matters: Extension of Clusters of Residues with Good Hydrophobic Contacts Stabilize (Hyper)Thermophilic Proteins

Prakash Chandra Rathi,[†] Hans Wolfgang Höffken,[‡] and Holger Gohlke*[,†]

[†]Institute for Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich Heine University, Universitätsstr. 1, 40225 Düsseldorf, Germany

[‡]BASF SE, GVM/C - A030, 67056 Ludwigshafen, Germany

Supporting Information



ABSTRACT: Identifying determinant(s) of protein thermostability is key for rational and data-driven protein engineering. By analyzing more than 130 pairs of mesophilic/(hyper)thermophilic proteins, we identified the quality (residue-wise energy) of hydrophobic interactions as a key factor for protein thermostability. This distinguishes our study from previous ones that investigated predominantly structural determinants. Considering this key factor, we successfully discriminated between pairs of mesophilic/(hyper)thermophilic proteins (discrimination accuracy: ~80%) and searched for structural weak spots in *E. coli* dihydrofolate reductase (classification accuracy: 70%).

Thermostable enzymes are sought after in industrial biotechnology because they allow carrying out biocatalysis at elevated temperatures, leading to an increase in reaction rates and, thus, making industrial processes economically more favorable.[1,2] Proteins from thermophilic and hyperthermophilic organisms tend to be more thermostable than their counterparts from mesophilic organisms.[1,3] This makes identifying and using enzymes from (hyper)thermophilic organisms an obvious approach in industrial biotechnology.[4,5] Screening large metagenomic libraries in search of a protein with desired properties is cumbersome, however.[6] Engineering proteins to improve thermostability is a promising alternative.[7] Directed evolution,[8] rational design,[9] and data-driven approaches[10] have been successfully applied for this.

The latter two approaches require knowledge of the mechanisms of how a protein can be made more thermostable. Comparisons of pairs of meso- and (hyper)thermophilic proteins have revealed several such mechanisms,[11,12] including improved hydrogen bonding,[13] ion pair and salt bridge networks,[12] better hydrophobic packing,[14] shortening of loops,[15] higher secondary structure content,[16] and increased rigidity of a protein.[17−21] As this list indicates, the focus of these analyses has been on structural factors, which may be the reason why different determinants of thermostability have been revealed.

In the present study, we systematically analyze a large data set of 132 pairs of mesophilic/thermophilic and 149 pairs of mesophilic/hyperthermophilic homologous protomers with the aim to identify the dominant determinant(s) of protein thermostability. To do so, we compared residue-wise interaction energy components and developed a hierarchical 3-D clustering of residues in a protein structure based on the energy components for discriminating mesophilic and (hyper)-thermophilic proteins. The clustering reveals that (hyper)-thermophilic proteins have larger clusters of residues of *good hydrophobic contacts* than their mesophilic counterparts. Compared to previous studies,[12,14,22,23] our results thus emphasize the *quality* (energy) of hydrophobic interactions as a discriminating factor rather than the sheer size of a cluster of hydrophobic residues. Thereby, our approach also allows suggesting residues where mutations should be incorporated for improving thermostability, as we demonstrate below.

The data set used here is an updated version of the one described in a previous study by Taylor et al.[21] in that it does not have duplicate (hyper)thermophilic protomers. The protomers in this data set are characterized by a high crystallographic quality (crystallographic resolution $\leq$ 2.2 Å and R-factor $\leq$ 0.23) and a high sequence diversity (sequence identity < 30% between structures of two different pairs). Furthermore, structures of a pair in the data set (I) show root-mean-square deviations less than 4 Å, (II) lead to structural alignments that include greater than or equal to 80% of each structure, and (III) have identical or closely related EC numbers or functional annotations (see Tables S1 and S2, Supporting Information (SI), for PDB IDs and chain IDs of protomer pairs in the data set). Finally, the data set we used is highly diverse in that the structures come from a variety of structural classes and vary in size (67−732 residues) (Figure S1, SI).

Rather than analyzing thermostability in terms of structural or geometric properties, we focused on energetic factors with the aim to identify (the) most significant determinant(s) of protein thermostability. Initially, we calculated for all protomers in the data set several residue-wise components to the interaction energy, i.e., electrostatic, van der Waals (vdW), hydrophobic, and hydrogen bond parts (supplemental experimental procedures, SI). We chose these interaction

Published: January 19, 2014

ACS Publications © 2014 American Chemical Society

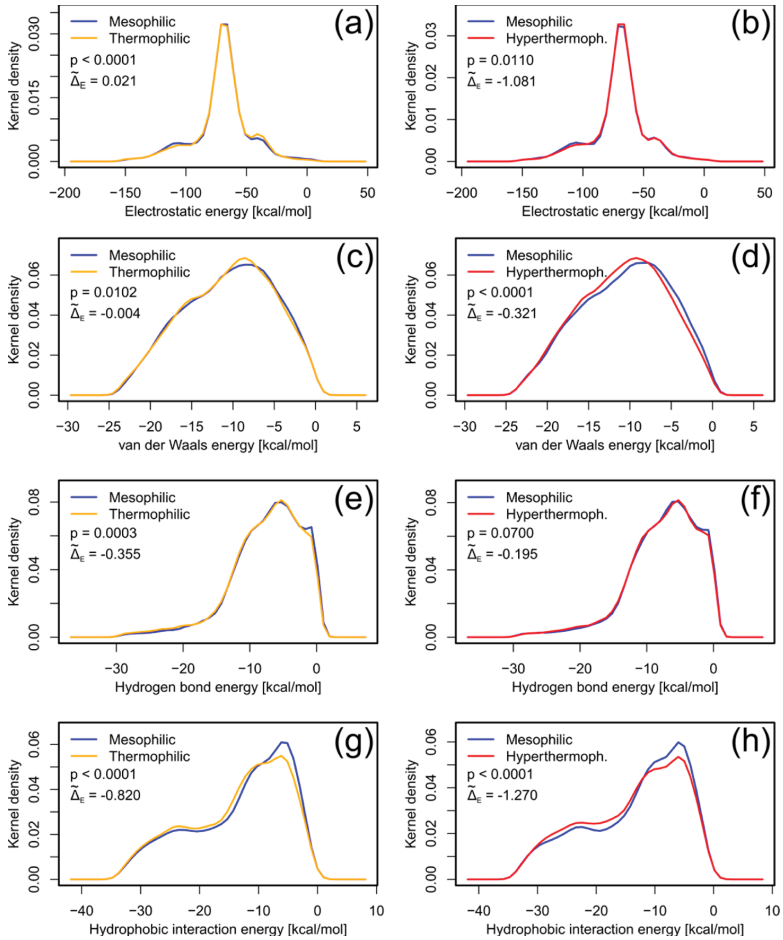


Figure 1. PDFs obtained by kernel density estimation of residue-wise energy components: electrostatic energy (a and b), van der Waals energy (c and d), hydrogen bond energy (e and f), and hydrophobic interaction energy (g and h) for pairs of mesophilic/thermophilic (a, c, e, g), as well as mesophilic/hyperthermophilic (b, d, f, h) protomers. A normal kernel function with an optimal smoothing parameter[45] at each data point was used for calculating the PDFs. The residue-wise energy values were trimmed to exclude values <1 percentile and >99 percentile. The statistical significance of the difference of two PDFs was calculated by a bootstrap hypothesis test of equality generating 10000 bootstrap samples as implemented in the "sm" package[46] of the R program (http://www.r-project.org). $\tilde{\Delta}_E$ indicates the difference between median residue-wise energies for (hyper)thermophilic and mesophilic protomers calculated from the kernel estimates.

energy components because these were identified as determinants for protein thermostability in previous studies using a small number of proteins.[13,14,24,25] All energy terms except the hydrogen bond energy were calculated using the Prime module version 3.0 of the Schrödinger software (Schrödinger, LLC, New York, NY, 2011).[26,27] The hydrogen bond energy (including charge-assisted hydrogen bonds) E_{HB} was calculated using a geometry-based energy function developed for protein design[28] as implemented in the FIRST software,[29] and then energies of all hydrogen bonds of a residue were summed.

In order to identify (the) dominant determinant(s) of protein thermostability, we initially compared distributions of residue-wise energy components at a global level, i.e., between all mesophilic and (hyper)thermophilic protomers. For this, probability density functions (PDFs) of these distributions were obtained from kernel density estimation,[30] which is a nonparametric way to estimate a PDF from a distribution based on a finite data sample. The PDFs of residue-wise electrostatic energies, vdW energies, hydrogen bond energies, and hydrophobic interaction energies differ between mesophilic and (hyper)thermophilic protomers with (hyper)thermophilic protomers showing higher probability densities at more negative (i.e., more favorable) energies (Figure 1); exceptions are the electrostatic (in the case of mesophilic/thermophilic pairs) and vdW energies (in the case of mesophilic/hyperthermophilic pairs) where the differences in the median energies of mesophilic and (hyper)thermophilic protomers ($\tilde{\Delta}_E$) are close to zero. A favorable difference in residue-wise electrostatic energies in the case of mesophilic/hyperthermophilic protomers but not in the case of mesophilic/thermophilic protomers is in line with results that ion pair interactions become preferentially stabilizing at higher temperatures because

of a reduced desolvation penalty.[31] The observed differences are statistically significant ($p < 0.05$ for the hypothesis of equality; Figure 1a−e, g, h) except for hydrogen bond energies in the case of mesophilic/hyperthermophilic protomers (Figure 1f). The statistical significance of the differences between two PDFs was calculated by a bootstrap hypothesis test of equality generating 10000 bootstrap samples. Here, during each bootstrap run, two new PDFs are generated by randomly choosing values from the combined set of values of the two data series. *P*-values are then calculated as the fraction of bootstrap samples that showed an equal or higher difference in the two new PDFs than the difference between the two original PDFs.

According to the *p*-values, the most significant difference between PDFs of mesophilic/thermophilic (Figure 1g) and mesophilic/hyperthermophilic (Figure 1h) protomers is found in the case of residue-wise hydrophobic energies ($p < 0.0001$ for both cases). This is also reflected in the magnitudes of the respective $\tilde{\Delta}_E$ values. On average, a residue in a thermophilic (hyperthermophilic) protomer has a hydrophobic energy that is more favorable by 0.82 (1.27) kcal mol^{-1} than that of a residue in a mesophilic protomer. The shoulder in the PDFs for hydrophobic interaction energies at around −24 kcal mol^{-1} is a result of the larger hydrophobic interaction energies of large hydrophobic and/or aromatic amino acids (Ile, Leu, Met, Phe, Trp, Tyr, and Val). These amino acids are not enriched in (hyper)thermophilic proteins (for our data set, we do not see a significant increase in the number of these amino acids in (hyper)thermophilic proteins compared to the mesophilic homologues; data not shown). Rather, the hydrophobic interaction energies of these residues are more favorable in the case of (hyper)thermophilic proteins. Overall, this demonstrates an energetically better hydrophobic packing in thermophilic proteins than in mesophilic proteins and an even better packing in hyperthermophilic proteins, which reflects that hydrophobic interactions become stronger with increasing temperature.[32,33] Note that, in contrast to previous studies[12,14,22,23] where the size of a cluster of hydrophobic residues was considered, our finding emphasizes the quality (energy) of residue-wise hydrophobic interactions as a discriminating factor.

Next, we investigated (differences in) the spatial distribution of residue-wise vdW, hydrogen bond, and hydrophobic interaction energies (i.e., where $\tilde{\Delta}_E < 0$ for both thermophilic and hyperthermophilic protomers compared to mesophilic protomers) in pairs of mesophilic/(hyper)thermophilic protomers. Following the idea of Protein Energy Networks introduced by Vijayabhaskar et al.,[34] our hypothesis is that a larger cluster of residues with lower energies than a given cutoff E_C exists in (hyper)thermophilic proteins than in their mesophilic homologues. However, in contrast to the study of Vijayabhaskar et al.,[34] we analyze residue-wise energy components rather than the total inter-residue interaction energy. This will allow us to identify, coupled to spatial resolution, which energy components are most determining for protein thermostability. To test our hypothesis, we performed a hierarchical clustering of residues with respect to vdW, hydrogen bond, and hydrophobic interaction energy components, respectively, such that all neighboring residues with an energy component lower than E_C for the respective clustering level are grouped in the same cluster (Figure 2). Thus, clusters grow in size as E_C increases (i.e., the energy component becomes less favorable). For each E_C, the fraction of residues

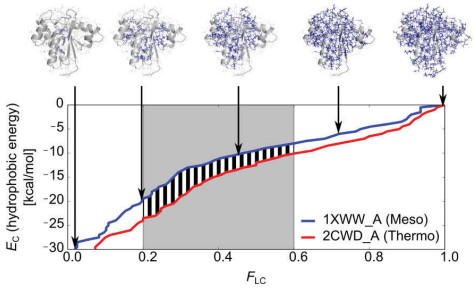


Figure 2. Discriminating mesophilic and (hyper)thermophilic proteins based on clusters of residues with good residue-wise energy components. Residues are clustered together if they are neighbors and if their values of the residue-wise energy components are below a cutoff E_C (largest clusters for selected E_C values are shown in the structures on the top as blue sticks). Residues are considered neighbors if the distance between the closest pair of atoms is less than or equal to 4 Å. E_C is increased in a stepwise manner, and the clustering is repeated. As a result, a hierarchical clustering is obtained where clusters become larger as E_C increases. For each E_C value, the fraction of residues that is part of the largest cluster with respect to all protein residues (F_{LC}) is calculated. As a descriptor for the discrimination, the area between the respective E_C vs F_{LC} curves for the (hyper)thermophilic and mesophilic proteins (black stripes) is then determined for the range of $F_{LC} \in [0.2, 0.6]$ (gray shading). If this value is negative, clusters of equal relative size have better residue-wise energy components in the case of the (hyper)thermophilic protein than in the case of the mesophilic protein. Preliminary tests showed that using other ranges of F_{LC} values for determining the area between the E_C vs F_{LC} curves does not result in significantly different discrimination accuracies than the best discrimination accuracies obtained with $F_{LC} \in [0.2, 0.6]$.

that is part of the largest cluster (F_{LC}) was calculated. With increasing E_C, F_{LC} increases from 0, when no residue is part of the largest cluster, to 1, when all residues belong to the largest cluster. If our hypothesis were true, the E_C vs F_{LC} curve of a (hyper)thermophilic protein should be shifted downward (toward lower E_C values) from the one of a mesophilic homologue; this is shown in Figure 2 for the case of phosphotyrosyl phosphatase (PDB IDs: 1XWW and 2CWD) considering the hydrophobic interaction energy (see caption of Figure 2 for more details). When analyzed across our data set, this finding holds for 83% of the pairs of mesophilic/thermophilic protomers and 76% of the pairs of mesophilic/hyperthermophilic protomers (Figure 3). These discrimination accuracies are significantly ($p < 0.001$) different from the one of a random discrimination (50%). This demonstrates that for the majority of (hyper)thermophilic proteins it is the size of clusters of residues *with good hydrophobic contacts* that is the dominant factor responsible for a high thermostability. Still, for approximately 20% of the pairs, this factor does not lead to a successful discrimination. Identifying other mechanisms of thermostabilization is not unequivocal, however. If residue-wise vdW energies are used for the clustering, a correct discrimination was obtained for 52% of mesophilic/thermophilic and 78% of mesophilic/hyperthermophilic pairs; the corresponding discrimination accuracies were 53% and 63% in the case of the hydrogen bond energy (Figure 3). Thus, only in the case of vdW energies applied to pairs of mesophilic/hyperthermophilic protomers, a discrimination accuracy similar to the one obtained with hydrophobic interaction energies was

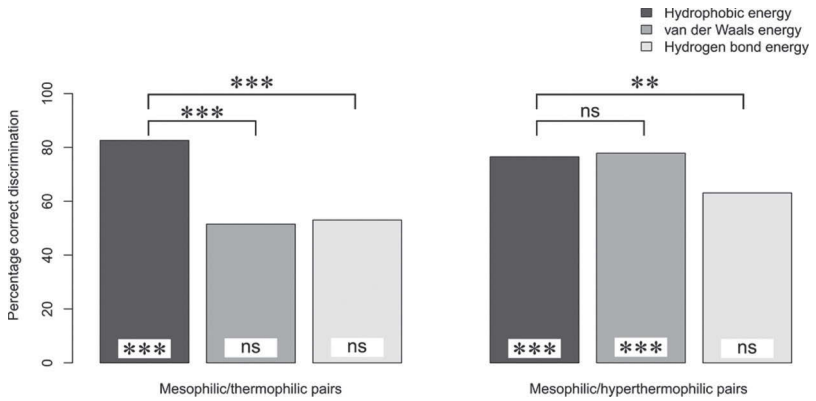


Figure 3. Discrimination accuracy between mesophilic and (hyper)thermophilic protomers based on clusters of residues with good residue-wise energy components. Lines connecting two bars indicate if the difference in discrimination accuracies for the two respective energy components is statistically significant. Marks at the bottom of a column indicate if the discrimination accuracy is significantly different from a random discrimination (50%). The statistical significance of the difference in discrimination accuracies is computed in both cases by a bootstrap hypothesis test of equality generating 10000 bootstrap samples. The significance levels are marked by ***: $p < 0.001$; **: $p < 0.01$; and ns: $p > 0.05$.

found (see Tables S3 and S4, SI, for *p*-values related to the significance of differences between all discrimination accuracies including random discrimination). In contrast, residue-wise hydrogen bond and vdW energies do not allow discriminating between pairs of mesophilic/thermophilic protomers.

We repeated the hierarchical clustering based only on inter-residue spatial distances; now all residues of the type "hydrophobic" (Ala, Cys, Ile, Leu, Met, Phe, Trp, and Val) were clustered together that are within a distance cutoff for each clustering level. This resulted in discrimination accuracies of 53% (62%) for pairs of mesophilic/thermophilic (mesophilic/hyperthermophilic) protomers, with only the discrimination accuracy for the mesophilic/hyperthermophilic protomers being significantly different from the one of a random discrimination ($p = 0.0369$) (see Table S5, SI, for discrimination accuracies and their statistical significances). This result is remarkable in that it demonstrates that it is the quality (energy) of hydrophobic interactions that discriminates mesophilic from (hyper)thermophilic proteins rather than the sheer size of the largest cluster of hydrophobic residues. One of the reasons is that with the criterion of hydrophobic interaction energy, residues that would usually not be classified as hydrophobic can also be considered part of the largest cluster: We observe that the largest cluster at $F_{LC} = 0.5$ also includes Arg (average fraction with respect to the number of residues in the cluster: 7.61%), Asn (0.54%), Asp (0.65%), Gln (2.06%), Glu (3.78%), His (2.09%), Pro (5.71%), Ser (1.63%), Thr (5.33%), and Tyr (6.31%) apart from residues of type "hydrophobic".

We further evaluated whether the state of a protein structure influences the outcome of the discrimination between pairs of mesophilic/(hyper)thermophilic proteins. So far, we had analyzed single chains of a protein. Now, we investigated protein pairs in terms of the biological assemblies such that interactions at protein interfaces are also considered. For this, only those pairs were used where both biological assemblies had the same oligomeric state and no residues were missing in the structures. This resulted in 67 mesophilic/thermophilic pairs as well as 67 mesophilic/hyperthermophilic pairs of biological assemblies. When performing the hierarchical clustering of residues based on the residue-wise hydrophobic energies, the discrimination accuracies are 87% (78%) for pairs of mesophilic/thermophilic (mesophilic/hyperthermophilic) biological assemblies. These results are not significantly different from the ones found for protomers ($p > 0.4$ for a hypothesis of equality using 10000 bootstrap samples) (see Figure S2, SI, for the accuracy of discrimination between mesophilic and (hyper)thermophilic biological assemblies). This means that for most (hyper)thermophilic proteins better hydrophobic packing within a protomer (rather than across the interface of a biological assembly) is the dominant factor responsible for a high thermostability.

In order to evaluate the robustness of our method with respect to the data set composition, we divided the data set in groups of protomer pairs based on sequence- or structure-related properties (sequence length, sequence identity, resolution, oligomeric state, presence of structural ions, SCOP class, and CATH class); then, we reanalyzed the results obtained from hierarchical clustering of residues based on the residue-wise hydrophobic interaction energies. We did not observe a pronounced influence of any of the properties on the discrimination accuracy except for the sequence length (Figure S1, SI). Longer protein chains result in higher discrimination accuracy. Likely, this is because larger proteins have larger hydrophobic clusters in which more residues with good hydrophobic interactions can be found in the case of (hyper)thermophilic proteins compared to mesophilic proteins. Overall, these results demonstrate that discriminating mesophilic and (hyper)thermophilic proteins based on clusters of residues with good hydrophobic interactions is highly robust with respect to the properties of the protein pairs considered.

Finally, we turned to investigating whether our finding that a larger cluster of residues with good hydrophobic interaction energies results in a more thermostable protein can be exploited *prospectively* for data-driven protein engineering by predicting structural weak spots, i.e., residues that when mutated would improve protein thermostability. As in a real-life scenario, we only used the structural information of the mesophilic protein for this. To predict such residues, E_C was set such that half of the protein's residues belong to the largest

Table 1. Experimental Validation of Predicted Weak Spots on *E. coli* DHFR

residue[a]	mutation(s)	weak spot rank[b]	reference[c]
Stabilizing mutations			
G15	A	35	36
W22	L	–[d]	36
D27	N	87	37
L28	R	71	37
L54	V	–[d]	36
P66	A	15	38
V88	I, A	–[d]	39
G95	A	32	40
Destabilizing mutations			
P21	L	–[d]	36
L24	V	–[d]	36
W30	M, Y, A, R, N, S, H, E	–[d]	41
F31	V, A	–[d]	36,37
T35	A	–[d]	36
P39	C	–[d]	42
V40	I, L, A, R, M, F, N, S, H	–[d]	41
G43	A	34	36
W74	F	–[d]	43
T113	V	–[d]	37
D122	A	29	36
E139	K, Q	64	44
S148	A, E, K, N, P, R, T, V	–[d]	41
I155	A, L, A, D, E, K, L, Q, R, S, T, V, W, Y	–[d]	36,41

[a]Residue IDs in bold indicate a true positive or a true negative weak spot prediction. [b]Weak spot rank based on the hydrophobic interaction energies; high ranks (low numbers) indicate weaker spots in comparison to residues with a low ranks. [c]Studies reporting thermostability evaluations of mutants. [d]Residue is not identified as a weak spot at any rank.

cluster (i.e., $F_{LC} = 0.5$). We chose $F_{LC} = 0.5$ because we visually observed that the cluster at this point represents the "hydrophobic core", and residues forming this should not be mutated. Residues in the immediate neighborhood of this cluster have a high (unfavorable) hydrophobic interaction energy, and mutating them should likely lead to a larger cluster of residues with good hydrophobic interaction energies. Hence, we consider these spatially close residues weak spot candidates. In order to prune the number of candidates, we ranked them by their hydrophobic interaction energies such that the weakest spot (highest energy) has the highest rank. After ranking, the top 25% with respect to the total number of residues of the protein are finally considered weak spots. In doing so, we use the ranking to enrich sites where a mutation should more likely improve thermostability. In addition to the site of mutagenesis, the actual outcome of a mutation on a protein's thermostability also depends on the types of residues exchanged. Thus, one cannot expect the weak spot rank alone to quantitatively correlate with the effect of a mutation on thermostability.

We evaluated this weak spot prediction first using *Escherichia coli* dihydrofolate reductase (DHFR) from our data set as an example. Several mutants of *E. coli* DHFR have been experimentally evaluated for their thermostability; the Protherm database (http://www.abren.net/protherm/)[35] lists eight (14) residues that stabilize (destabilize) the protein upon single-point mutation(s) (Table 1; see Table S6, SI, for individual mutants and their difference in thermostability from the wild-type). Three out of the eight thermostabilizing residues were correctly predicted as weak spots by our approach (Table 1; Figure 4). In turn, 12 out of the 14 destabilizing residues were correctly predicted as nonweak spots (Table 1; Figure 4). This yields a classification accuracy of almost 70%, with our approach being more accurate in identifying nonweak spots (specificity: 85%) than weak spots (sensitivity: 38%). Of the five weak spots missed, two (D27N, L28R) resulted in a more thermostable protein upon mutation to equally polar or even more polar residues. Thus, expecting to identify these residues as weak spots appears to be beyond the scope of our approach. In fact, these residues were assigned low weak spot ranks (87, 71), indicating that improving hydrophobic interactions at these spots might not lead to a more thermostable protein. Regarding two further weak spots missed

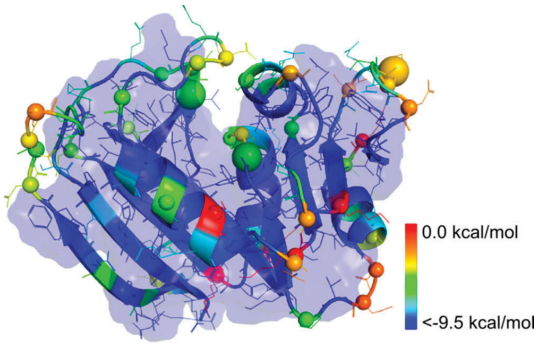


Figure 4. Predicted weak spots mapped onto the structure of *E. coli* DHFR. Residues are colored by a rainbow color ramp according to their hydrophobic interaction energies. The largest cluster with $F_{LC} = 0.5$ observed at a cutoff of the hydrophobic interaction energy $E_C = -9.5$ kcal mol^{-1} is enclosed by a transparent surface. C_α atoms of weak spot residues are represented as spheres. Weak spots that have been validated in the literature are marked by a large sphere.

dx.doi.org/10.1021/ci400568c | J. Chem. Inf. Model. 2014, 54, 355–361

(W22L, L54V), mutations to smaller hydrophobic residues there led to a more thermostable protein. Because our method is particularly suited for identifying weak spots that when mutated to residues with improved hydrophobic interactions should lead to improved thermostability, missing these two weak spots thus is not unexpected either. *E. coli* DHFR in our data set is a rare example for which comprehensive sets of single mutants leading to stabilization or destabilization have been recorded in the Protherm database. For further validation of our weak spot prediction, we thus resorted to two systems for which only stabilizing or only destabilizing mutants have been reported. For *Bacillus subtilis* adenylate kinase, two thermostabilizing multiple mutants have been reported in the Protherm database, incorporating in total 26 mutations. We correctly predicted nine out of 19 mutations (excluding mutations involving the exchange of one hydrophobic residue with another) as weak spots (sensitivity: 47%; see Table S7, SI, for details). As a counter example, we considered the *E. coli* maltose binding protein (MBP) for which all but one (Gly to Cys mutation at position 19) of the 16 destabilizing single-point mutations reported in the Protherm database were correctly predicted as nonweak spots (specificity: 93.75%; see Table S7, SI, for details). Note that this result is not trivial as one might be tempted to think considering that all but one (Tyr to Asp mutation at position 283) of the correct predictions involve mutations of larger hydrophobic residues to smaller ones. Rather, even without considering the actual outcome of a mutation on a protein's thermostability, our method suggests that for improving thermostability these nonweak spot residues should not be mutated because they are already part of the "hydrophobic core" with good hydrophobic interaction energies. Finally, considering the results for all three systems shows that our method is more accurate in identifying nonweak spots than weak spots. In our view, these results are encouraging given, first, the fact that we could reliably exclude the majority of nonweak spots and, second, the ease of computation with which this classification is obtained. The former would already result in a much reduced experimental effort when performing site saturation mutagenesis for identifying thermostable mutants; the latter suggests that our approach can be used as a prefilter for further rational design approaches where more rigorous (and costly) prediction methods are applied. In particular, as our approach focuses on identifying weak spots where improving hydrophobic interactions should lead to improved thermostability, we recommend combining it with other approaches for weak spot prediction that focus on different mechanisms of thermostabilization.

In summary, in the present study, we aimed at identifying dominant determinant(s) of protein thermostability. On the basis of one of the largest data sets investigated in this context and thorough statistical evaluation, our results substantiate the importance of the quality (energy) of hydrophobic interactions for protein thermostability. Considering residue-wise hydrophobic interaction energies at a global level, an energetically better hydrophobic packing in thermophilic proteins than in mesophilic proteins is detected, and an even better packing in hyperthermophilic proteins. Accordingly, by identifying clusters of residues with good hydrophobic interaction energies alone, we were able to successfully discriminate between pairs of mesophilic/(hyper)thermophilic proteins with an accuracy of ~80%. These results are robust with respect to the properties of protein pairs considered. Considering the size of clusters of hydrophobic residues instead resulted in at most a weak discriminatory power. Finally, we successfully applied the criterion of clusters of residues with good hydrophobic interaction energies to search for structural weak spots, which will allow guiding data-driven protein engineering. These results and the computational efficiency position our approach as a valuable complement to existing approaches for analyzing proteins with respect to thermostability and identifying structural weak spots.

■ ASSOCIATED CONTENT

Supporting Information

Detailed protocols of experimental procedures, additional tables showing data set composition (Tables S1–S2), *p*-values regarding equality in discrimination accuracies (Tables S3–S4), discrimination between mesophilic and (hyper)-thermophilic protomers when clustering residues of type "hydrophobic" by inter-residue spatial distances (Table S5), thermostability of *E. coli* DHFR mutants (Table S6), and further validation of the weak spot prediction (Table S7), as well as additional figures showing the discrimination accuracy between mesophilic and (hyper)thermophilic protomers (Figure S1) and biological assemblies (Figure S2). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: (+49) 211-81-13662. Fax: (+49) 211-81-13847. E-mail: gohlke@uni-duesseldorf.de.

Notes

The authors declare the following competing financial interest(s): Dr. Wolfgang Hoeffken is an employee of BASF SE.

■ ACKNOWLEDGMENTS

We thank Taylor Todd, National Cancer Institute, Bethesda, MD, for providing the dataset of mesophilic/(hyper) thermophilic protomer pairs. We are grateful to the Ministry of Innovation, Science, and Research of North Rhine-Westphalia and Heinrich Heine University, Düsseldorf, for a scholarship to P.C.R. within the CLIB-Graduate Cluster Industrial Biotechnology and BASF SE for giving P.C.R. the opportunity to work as an intern.

■ REFERENCES

(1) Demirjian, D. C.; Moris-Varas, F.; Cassidy, C. S. Enzymes from extremophiles. *Curr. Opin. Chem. Biol.* **2001**, *5*, 144–151.
(2) Van den Burg, B. Extremophiles as a source for novel enzymes. *Curr. Opin. Microbiol.* **2003**, *6*, 213–218.
(3) Vieille, C.; Zeikus, G. J. Hyperthermophilic enzymes: Sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* **2001**, *65*, 1–43.
(4) Egorova, K.; Antranikian, G. Industrial relevance of thermophilic Archaea. *Curr. Opin. Microbiol.* **2005**, *8*, 649–655.
(5) Niehaus, F.; Bertoldo, C.; Kahler, M.; Antranikian, G. Extremophiles as a source of novel enzymes for industrial application. *Appl. Microbiol. Biotechnol.* **1999**, *51*, 711–729.
(6) Lorenz, P.; Schleper, C. Metagenome—A challenging source of enzyme discovery. *J. Mol. Catal. B: Enzymatic* **2002**, *19*, 13–19.
(7) Leisola, M.; Turunen, O. Protein engineering: Opportunities and challenges. *Appl. Microbiol. Biotechnol.* **2007**, *75*, 1225–1232.
(8) Eijsink, V. G. H.; Gaseidnes, S.; Borchert, T. V.; van den Burg, B. Directed evolution of enzyme stability. *Biomol. Eng.* **2005**, *22*, 21–30.

(9) Eijsink, V. G. H.; Bjørk, A.; Gåseidnes, S.; Sirevåg, R.; Synstad, B.; van den Burg, B.; Vriend, G. Rational engineering of enzyme stability. *J. Biotechnol.* **2004**, *113*, 105−120.
(10) Chaparro Riggers, J. F.; Polizzi, K. M.; Bommarius, A. S. Better library design: Data driven protein engineering. *Biotechnol. J.* **2007**, *2*, 180−191.
(11) Razvi, A.; Scholtz, J. M. Lessons in stability from thermophilic proteins. *Protein Sci.* **2006**, *15*, 1569−1578.
(12) Kumar, S.; Tsai, C. J.; Nussinov, R. Factors enhancing protein thermostability. *Protein Eng.* **2000**, *13*, 179−191.
(13) Vogt, G.; Woell, S.; Argos, P. Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.* **1997**, *269*, 631−643.
(14) Gromiha, M. M.; Pathak, M. C.; Saraboji, K.; Ortlund, E. A.; Gaucher, E. A. Hydrophobic environment is a key factor for the stability of thermophilic proteins. *Proteins: Struct., Funct., Bioinf.* **2013**, *81*, 715−721.
(15) Russell, R. J.; Hough, D. W.; Danson, M. J.; Taylor, G. L. The crystal structure of citrate synthase from the thermophilic archaeon, *Thermoplasma acidophilum*. *Structure* **1994**, *2*, 1157−1167.
(16) Querol, E.; PerezPons, J. A.; MozoVillarias, A. Analysis of protein conformational characteristics related to thermostability. *Protein Eng.* **1996**, *9*, 265−271.
(17) Vihinen, M. Relationship of protein flexibility to thermostability. *Protein Eng.* **1987**, *1*, 477−480.
(18) Rathi, P. C.; Radestock, S.; Gohlke, H. Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.* **2012**, *159*, 135−144.
(19) Radestock, S.; Gohlke, H. Protein rigidity and thermophilic adaptation. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 1089−1108.
(20) Radestock, S.; Gohlke, H. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.* **2008**, *8*, 507−522.
(21) Taylor, T. J.; Vaisman, I. I. Discrimination of thermophilic and mesophilic proteins. *BMC Struct. Biol.* **2010**, *10* (Suppl1), S5.
(22) Kannan, N.; Vishveshwara, S. Aromatic clusters: A determinant of thermal stability of thermophilic proteins. *Protein Eng.* **2000**, *13*, 753−761.
(23) Kim, T.; Joo, J. C.; Yoo, Y. J. Hydrophobic interaction network analysis for thermostabilization of a mesophilic xylanase. *J. Biotechnol.* **2012**, *161*, 49−59.
(24) Dominy, B. N.; Minoux, H.; Brooks, C. L., III. An electrostatic basis for the stability of thermophilic proteins. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 128−141.
(25) Glyakina, A. V.; Garbuzynskiy, S. O.; Lobanov, M. Y.; Galzitskaya, O. V. Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. *Bioinformatics* **2007**, *23*, 2231−2238.
(26) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. A hierarchical approach to all-atom protein loop prediction. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 351−367.
(27) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **2002**, *320*, 597−608.
(28) Dahiyat, B. I.; Gordon, D. B.; Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci.* **1997**, *6*, 1333−1337.
(29) FIRST, a program for analysing flexibility of networks. http://flexweb.asu.edu/ (accessed January 17, 2014).
(30) Nadaraya, É. A. On non-parametric estimates of density functions and regression curves. *Theory Probab. Appl.* **1965**, *10*, 186−190.
(31) Elcock, A. H. The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. *J. Mol. Biol.* **1998**, *284*, 489−502.
(32) Privalov, P. L.; Gill, S. J. Stability of protein structure and hydrophobic interaction. *Adv. Protein Chem.* **1988**, *39*, 191−234.
(33) Schellman, J. A. Temperature, stability, and the hydrophobic interaction. *Biophys. J.* **1997**, *73*, 2960−2964.
(34) Vijayabaskar, M. S.; Vishveshwara, S. Comparative analysis of thermophilic and mesophilic proteins using Protein Energy Networks. *BMC Bioinf.* **2010**, *11* (Suppl 1), S49.
(35) Kumar, M. D.; Bava, K. A.; Gromiha, M. M.; Prabakaran, P.; Kitajima, K.; Uedaira, H.; Sarai, A. ProTherm and ProNIT: Thermodynamic databases for proteins and protein−nucleic acid interactions. *Nucleic Acids Res.* **2006**, *34*, D204−D206.
(36) Arai, M.; Maki, K.; Takahashi, H.; Iwakura, M. Testing the relationship between foldability and the early folding events of dihydrofolate reductase from *Escherichia coli*. *J. Mol. Biol.* **2003**, *328*, 273−288.
(37) Perry, K. M.; Onuffer, J. J.; Touchette, N. A.; Herndon, C. S.; Gittelman, M. S.; Matthews, C. R.; Chen, J. T.; Mayer, R. J.; Taira, K.; Benkovic, S. J.; Howell, E. E.; Kraut, J. Effect of single amino acid replacements on the folding and stability of dihydrofolate reductase from *Escherichia coli*. *Biochemistry* **1987**, *26*, 2674−2682.
(38) Texter, F. L.; Spencer, D. B.; Rosenstein, R.; Matthews, C. R. Intramolecular catalysis of a proline isomerization reaction in the folding of dihydrofolate reductase. *Biochemistry* **1992**, *31*, 5687−5691.
(39) Ahrweiler, P. M.; Frieden, C. Effects of point mutations in a hinge region on the stability, folding, and enzymatic activity of *Escherichia coli* dihydrofolate reductase. *Biochemistry* **1991**, *30*, 7801−7809.
(40) Svensson, A. K.; O'Neill, J. C., Jr.; Matthews, C. R. The coordination of the isomerization of a conserved non-prolyl cis peptide bond with the rate-limiting steps in the folding of dihydrofolate reductase. *J. Mol. Biol.* **2003**, *326*, 569−583.
(41) Arai, M.; Iwakura, M. Probing the interactions between the folding elements early in the folding of Escherichia coli dihydrofolate reductase by systematic sequence perturbation analysis. *J. Mol. Biol.* **2005**, *347*, 337−353.
(42) Villafranca, J. E.; Howell, E. E.; Oatley, S. J.; Xuong, N. H.; Kraut, J. An engineered disulfide bond in dihydrofolate reductase. *Biochemistry* **1987**, *26*, 2182−2189.
(43) Garvey, E. P.; Swank, J.; Matthews, C. R. A hydrophobic cluster forms early in the folding of dihydrofolate reductase. *Proteins* **1989**, *6*, 259−266.
(44) Perry, K. M.; Onuffer, J. J.; Gittelman, M. S.; Barmat, L.; Matthews, C. R. Long-range electrostatic interactions can influence the folding, stability, and cooperativity of dihydrofolate reductase. *Biochemistry* **1989**, *28*, 7961−7968.
(45) Silverman, B. W. *Density Estimation for Statistics and Data Analysis*; Chapman & Hall/CRC: London, 1998.
(46) Bowman, A. W.; Azzalini, A. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*; Oxford University Press: Oxford, 1997.

Publication I – Supplementary Information

Quality matters: Extension of clusters of residues with good hydrophobic contacts stabilize (hyper)thermophilic proteins

Prakash Chandra Rathi, Hans Wolfgang Höffken, and Holger Gohlke
Journal of Chemical Information and Modeling, **2014**, 54: 355-361.
2012 impact factor: 4.20; contribution: 70%

Supplemental experimental procedures

Preparation of protein structures

All proteins in the dataset were downloaded from the Protein Data Bank (PDB)[1], and if required, the desired chain was extracted from the PDB file. All water molecules, ligands, and ions were removed from the structures. The command line version of the protein preparation wizard[2] of the Schrödinger software (Schrödinger, LLC, New York, NY, 2011) was used to prepare the protomer structures in order to I) add hydrogen atoms, II) add missing side chain atoms, and III) build disulphide bridges. Protein bio-assemblies were downloaded from the PDB and prepared in an identical manner as the protomers.

Calculation of residue-wise energy components

Protomers (and bio-assemblies) were minimized using the Prime module[3,4] version 3.0 of the Schrödinger software (Schrödinger, LLC, New York, NY, 2011) using default settings. Then, residue-wise electrostatic, van der Waals, and hydrophobic interaction energy components were calculated for the minimized structures by Prime. The hydrogen bond (including charge assisted hydrogen bonds) energy was calculated using a geometry-based energy function developed for protein design[5] as implemented in the FIRST software[6]. The energies of all hydrogen bonds for a residue were summed for calculating residue-wise hydrogen bond energies.

Clustering of residues by residue-wise energy components

Residues in a protein structure are clustered together if they are neighbors and if their values of the residue-wise energy components are below a cutoff E_C. Residues are considered neighbors if the distance between the closest pair of atoms is ≤ 4 Å. E_C is increased in a stepwise manner, and the clustering is repeated for each new E_C. As a result, a hierarchical clustering is obtained where clusters become larger as E_C increases. For each E_C value, the fraction of residues that is part of the largest cluster with respect to all protein residues (F_{LC}) is calculated. E_C was increased from an initial value of −30 kcal mol^{-1} to a final value of 0 kcal mol^{-1} with a step size of 0.5 kcal mol^{-1}.

Clustering of hydrophobic residues by inter-residue distances

Residues in a protein structure are clustered together if they belong to the type “hydrophobic” (Ala, Cys, Ile, Leu, Met, Phe, Trp, and Val) and are within a distance cutoff D_C. The distance between the closest atoms of two residues was considered the distance between these residues. D_C is increased in a stepwise-manner, and the clustering is repeated for each new D_C. As a result, a hierarchical clustering is obtained where clusters become larger as D_C increases. For each D_C value, the fraction of residues that is part of the largest cluster with respect to all protein residues (F_{LC}) is calculated. D_C was increased from an initial value of 1 Å to a final value of 5 Å with a step size of 0.05 Å.

Supplemental tables

Table S1. PDB ID and chain identifier of pairs of mesophilic/thermophilic proteins.

Thermoph.	Mesoph.	Thermoph.	Mesoph.	Thermoph.	Mesoph.	Thermoph.	Mesoph.
1nw2_A	1fb6_A	1hbn_B	1e6y_B	1odk_B	1vhw_A	2cuy_A	1mla_A
1urd_A	1anf_A	2f2b_A	1z98_A	1uay_A	1e6w_D	2cwd_A	1xww_A
2hm7_A	1lzl_A	2q5b_A	2cj3_A	1ub3_A	1p1x_A	2d1y_C	2zat_A
2sqc_A	1w6j_A	2v08_A	1ls9_A	1ufy_A	1dbf_A	2d29_A	2vig_A
1c9o_A	2es2_A	1ugp_B	2cz1_B	1ug6_A	1e4i_A	2d4e_A	2ve5_D
2b5a_A	1y7y_A	1b06_A	1bsm_A	1ui0_A	2c2q_A	2d4p_A	1s3z_A
1b4b_A	2p5m_A	1mp9_A	1qna_A	1uir_B	2pt9_A	2d5b_A	1pfv_A
1g2w_A	1iye_A	1thm_A	2tec_E	1uj5_A	2f8m_A	2d5c_A	1nyt_A
1gtf_A	1wap_A	3tec_I	1cse_I	1ulr_A	1urr_A	2d5w_A	1zu0_A
1hvx_A	3bh4_A	1h1n_A	1a3h_A	1umd_A	2ozl_A	2d8d_A	1ecm_A
1lqy_A	2okl_A	1i1x_A	1ta3_B	1umd_B	2ozl_B	2dt9_A	2dtj_A
1r2z_A	1xc8_A	1mtp_A	1sek_A	1v37_A	2a6p_A	2e7u_A	2hp1_A
1tqh_A	3dlt_A	2fla_A	1b0y_A	1v6s_A	1hdi_A	2ebj_A	1aug_A
1whi_A	1vqo_K	1bqc_A	1a3h_A	1v8f_A	1n2e_A	2eg4_A	1h4k_X
1y51_A	1ptf_A	1tf4_A	1ga2_B	1v8m_A	3ees_A	2eiy_B	1iye_A
1zdr_A	3dau_A	1tml_A	1dys_A	1v98_A	1fb6_A	2ekp_A	2zat_A
1zin_A	2eu8_A	1tib_A	3tgl_A	1vbi_A	1wtj_A	2fk5_A	1e4c_P
2bkm_A	2qrw_B	1yna_A	2dfb_A	1vc3_B	1uhe_A	2is8_A	1ihc_A
2exi_A	1y7b_A	2dte_A	1uzn_A	1vcd_A	1ktg_A	2j07_A	1owl_A
2tlx_A	1bqb_A	1my6_A	1xre_A	1ve1_A	2pqm_A	2p5y_A	1ek6_A
2bd0_A	1oaa_A	2c41_A	2c2u_A	1vef_A	2oat_A	2prd_A	2bqx_A
1aoh_A	1g1k_A	1esw_A	1x1n_A	1vfj_A	2pii_A	2pwy_A	1i9g_A
1cem_A	1v5d_A	2ng1_A	2qy9_A	1wlu_A	1q4u_A	2qhs_A	1w66_A
1h6y_A	1gny_A	1b5p_A	1asd_A	1wmw_A	2f94_F	2yqu_A	3lad_A
1nbc_A	1g43_A	1bxb_A	1oad_A	1wo8_A	1b93_A	2yvp_A	1g0s_B
1xyz_A	1e0w_A	1gd7_A	2q2i_A	1wur_A	1a8r_A	2z1a_A	1hpu_A
2b59_A	1qzn_A	1iv3_A	1h47_A	1wz8_A	2zqq_A	2z1y_B	1j32_A
2olj_A	1b0u_A	1iz9_A	2hjr_A	1x1o_A	2jbm_C	2zc8_A	1sjd_A
2q8x_A	1uqz_A	1j33_A	1l5o_A	1yya_A	2dp3_A	2zdb_A	3d0s_A
3d60_A	1gyh_A	1j3n_A	1oxh_A	1z54_A	1s5u_A	2zdh_A	1iow_A
1eje_A	3bnk_A	1j3w_A	1vet_B	2b3f_A	1anf_A	3cm0_A	3adk_A
1ep0_A	2ixc_A	1n97_A	1bu7_A	2bhq_A	2o2r_A	3hrx_A	1dci_A
1g5c_A	1ylk_A	1nza_A	2zfh_A	2cuk_A	2gcg_A	3mds_A	1xre_A

Table S2. PDB ID and chain identifier of pairs of mesophilic/hyperthermophilic proteins.

Hyper-thermoph.	Mesoph.	Hyper-thermoph.	Mesoph.	Hyper-thermoph.	Mesoph.	Hyper-thermoph.	Mesoph.
1h2b_A	1n8k_A	3c7b_B	2v4j_B	1zjj_A	2c4n_A	2z30_A	1st3_A
1n7k_A	1p1x_A	3cnu_A	2qzt_B	2cun_A	1hdi_A	1eu8_A	1anf_A
1tyo_A	1pb1_A	3do8_B	1coz_A	2cwp_A	3ers_X	1wst_A	2r2n_A
2fc3_A	1zwz_A	1g6h_A	2ff7_A	2d69_A	1rbl_A	1uxt_A	1euh_A
2yvu_A	2pez_A	1l2t_A	2ff7_A	2dbb_A	2qz8_A	2r91_A	2v8z_A
1c3p_A	1t64_A	1pkh_A	2qxx_A	2dr1_A	1w23_A	1d1g_A	3fq0_A
1hqk_A	2c92_A	1snn_A	1k4i_A	2dxe_A	1npk_A	1inl_C	2o07_B
1mzh_A	1p1x_A	1twi_A	1ko0_A	2e5f_A	1moq_A	1kq3_A	1ta9_B
1tz7_A	1x1n_A	2eb0_B	1k20_A	2e5w_A	2pt9_A	1nf2_A	1rkq_A
1ulz_A	2w70_A	2j9d_C	2pii_A	2ekn_A	2eey_A	1o0x_A	1y1n_A
1wwr_D	2b3j_A	2pa6_A	2akz_A	2hun_B	1r66_A	1o0y_A	1p1x_A
2e55_A	1bd3_A	2yww_A	2fzc_B	3cg3_A	3cfx_A	1o4s_A	1asd_A
2e8e_A	1n2f_A	2z02_A	2gqs_A	1ais_A	1qna_A	1oh4_A	1pmj_X
2ebd_A	1zow_A	2z8u_B	1qna_A	1mxd_A	3bh4_A	1p1m_A	2i9u_A
2egj_A	1s5u_A	1ftr_A	1m5s_A	1b7g_O	1u8f_R	1tmy_A	3chy_A
2ehh_A	3di1_A	1vcv_A	1p1x_A	1io7_A	3bdz_A	1tzx_A	1eyv_A
2ehs_A	1l0i_A	1ml4_A	1ekx_A	1je1_C	1vhw_A	1vbu_A	1ta3_B
2eja_A	3gw0_A	1aj8_A	1csh_A	1nto_A	1n8k_A	1vc1_A	1h4y_A
2hk9_A	1nyt_A	1gtm_A	1bgv_A	1uwr_A	2e3z_A	1vj0_A	1n8k_A
2omd_A	2q5w_E	1jg1_A	1i1n_A	1vph_A	1vmh_A	1vl8_A	1gee_A
2pbq_A	2g4r_A	1nnh_A	12as_A	1xtt_A	1bd3_A	1vlc_A	1cnz_A
2pbr_A	1e9e_A	1pvv_A	1oth_A	2f5g_A	2vjv_A	1vlg_C	1eum_A
2pnf_A	1q7b_A	1vkc_A	2fe7_B	2i6j_A	1fpz_A	1vlh_B	1qjc_A
2r75_1	2vxy_A	1ybz_A	1ecm_A	2var_C	1rkd_A	1vlj_A	1ta9_B
2yvl_A	1i9g_A	2dsk_A	2uy3_A	3f8p_D	1trb_A	1vm7_A	2fv7_A
2yvw_A	1uae_A	1gde_A	2r5e_A	1vgm_A	2h12_B	1vma_A	2qy9_A
2yw2_A	3g8c_A	1iu8_A	1aug_A	1wlt_A	2ixc_A	1vmj_A	1vmh_A
2z1m_A	1rpn_A	1lk5_A	1m0s_A	1wrj_A	1sfe_A	1vp2_A	1ex2_A
1coj_A	1bsm_A	1ub9_A	1r1u_A	1x0u_A	1on3_A	1vq0_A	1vzy_A
1jji_A	1lzl_A	1udd_A	2qcx_B	1x25_A	1qd9_A	1w2t_A	1y4w_A
1lbv_A	2qfl_A	1uku_A	2zfh_A	2e0q_A	1fb6_A	1w3j_A	1e4i_A
1p1l_A	2nuh_A	1v96_B	2h1c_A	2e7x_A	2qz8_A	1wa3_A	1wbh_A
1txg_A	1n1e_A	1w2i_A	1urr_A	2ehg_A	1jl1_A	1wos_A	1wsr_A
1vi6_A	3bch_A	1wqa_A	1k2y_X	2ekl_A	1dxy_A	2e54_A	2oat_A
2a5w_A	2v4j_C	1wr8_A	1s2o_A	2ggs_A	1n2s_A	2fnc_A	1anf_A
2b2h_A	3bhs_A	1wwk_A	1dxy_A	2bo1_A	1vqo_F	2h3h_B	2dri_A
2cyb_A	2yxn_A	1wy1_A	2zhz_A	1mgt_A	1sfe_A	2p3n_A	2qfl_A
3c7b_A	2v4j_A						

Table S3. *p*-values regarding equality in discrimination accuracies between mesophilic and (hyper)thermophilic protomers for clustering based on different residue-wise energy components *versus* a random discrimination.[a]

	Hydrogen bond	**van der Waals**	**Hydrophobic interaction**	**Random**
Hydrogen bond	- [b]	0.3245	<0.0001	0.2196
van der Waals	0.0005	- [b]	<0.0001	0.8056
Hydrophobic interaction	0.0042	0.5829	- [b]	<0.0001
Random	0.0610	<0.0001	<0.0001	- [b]

[a] The discrimination analysis is based on clustering by residue-wise energy components. The *p*-values were computed by a bootstrap hypothesis of equality generating 10000 bootstrap samples. Values in shaded shells correspond to mesophilic/hyperthermophilic protomers, other values correspond to mesophilic/thermophilic protomers.

[b] Not determined.

Table S4. *p*-values regarding equality in discrimination accuracies between mesophilic and (hyper)thermophilic protein bio-assemblies for clustering based on different residue-wise energy components *versus* a random discrimination. [a]

	Hydrogen bond	**van der Waals**	**Hydrophobic interaction**	**Random**
Hydrogen bond	- [b]	1.0000	<0.0001	0.6020
van der Waals	0.0572	- [b]	<0.0001	0.6068
Hydrophobic interaction	0.0534	1.0000	- [b]	<0.0001
Random	0.1141	0.0004	0.0007	- [b]

[a] The discrimination analysis is based on clustering by residue-wise energy components. The *p*-values were computed by a bootstrap hypothesis of equality generating 10000 bootstrap samples. Values in shaded shells correspond to mesophilic/hyperthermophilic protein bio-assemblies, other values correspond to mesophilic/thermophilic protein bio-assemblies.

[b] Not determined.

Table S5. Discrimination between mesophilic and (hyper)thermophilic protomers when clustering residues of type "hydrophobic" by inter-residue spatial distance.

Pairs	**Discrimination accuracy[b]**	***p*-value[a]**
Mesophilic/thermophilic	53.03	0.6175
Mesophilic/hyperthermophilic	61.74	0.0369

[a] The *p*-values were computed by a bootstrap hypothesis of equality between the given discrimination accuracy and a random discrimination (50% correct discrimination) generating 10000 bootstrap samples.
[b] In %.

Table S6. Thermostability of *E. coli* DHFR mutants.

Mutant	ΔΔ*G* (H_2O) [kcal mol^{-1}][a]	Mutant	ΔΔ*G* (H_2O) [kcal mol^{-1}] [a]
G15A	**0.70**	W74F	-1.20
P21L	-0.10	V88I	**0.75**
W22L	**0.10**	V88A	**0.39**
L24V	-1.90	G95A	**1.30**
D27N	**1.40**	T113V	-1.20
L28R	**1.72**	D122A	-1.60
W30M	-2.03	E139Q	-0.42
W30Y	-2.16	E139K	-1.00
W30A	-2.33	S148K	-0.26
W30R	-2.49	S148P	-0.26
W30N	-2.52	S148V	-0.33
W30S	-2.74	S148A	-0.47
W30H	-2.78	S148T	-0.51
W30E	-2.89	S148E	-0.52
F31V	-1.50	S148R	-0.75
F31A	-1.90	S148N	-0.89
T35A	-1.10	I155V	-0.58
P39C	-3.00[b]	I155L	-2.27
V40I	-0.85	I155L	-2.80
V40L	-1.35	I155E	-3.26
V40A	-1.55	I155R	-3.28
V40R	-1.72	I155T	-3.30
V40M	-2.00	I155K	-3.35
V40F	-2.15	I155Y	-3.64
V40N	-2.17	I155A	-3.82
V40S	-2.52	I155Q	-3.86
V40H	-3.27	I155S	-3.93
G43A	-0.40	I155A	-4.00
L54V	**0.40**	I155D	-4.10
P66A	**1.30**	I155W	-4.31

[a] ΔG (mutant) − ΔG (wildt-ype) where ΔG is the free energy of unfolding in water, determined by denaturant (urea; guanidine hydrochloride; glutathione disulfide/glutathione; guanidinium thiocyanate) denaturation of proteins and extrapolation of the data to zero concentration of the denaturant. A positive value (marked in bold) indicates that the mutant is more thermostable than the wild-type.

[b] T_m (mutant) − T_m (wild-type) in K where T_m is the melting temperature identified as a mid-point temperature at which half of the protein is unfolded in a thermal unfolding method.

Table S7. Additional validation of weak spot prediction.

Protein	Mutations[a]	% correct prediction	Comment	Ref.
B. subtilis adenylate kinase	**Stabilizing mutations** *L3I*, G17A, D23K, **K69R**, **G73S**, **D75S**, I99S, *Y103M*, K105R, E114Q, **D118E**, V119E, *M121I*, E122A, S169T, Q180A, **D184A**, S187D, **E188S**, **G190E**, *Y191V*, *A193V*, *Y205F*, **D210V**, *L211I*, **K217Q**	34.61% (47.36% excluding mutations involving the exchange of one hydrophobic residue with another)	Mutations in two multiple-mutants that led to an increase of 11.6ºC and 12.5ºC in the melting temperature T_m compared to the wild type	7
E. coli maltose binding protein	**Destabilizing mutations** **V8G**, **W10A**, G19C, **I59A**, **I108A**, **L115A**, **L147A**, **P159A**, **I161A**, **L192A**, **L195A**, **I226A**, **A276G**, **Y283D**, **V347A**, **L361A**	93.75%	Single point mutations that led to a decrease in the T_m in a range of 0.1 to 7.5ºC or in the free energy of unfolding ΔG in a range of 0.3 to 5.5 kcal mol^{-1} compared to the wild type	8-11

[a] A correctly predicted mutation site is marked in bold. A mutation in italic involves the exchange of one hydrophobic residue with another.

Supplemental figures

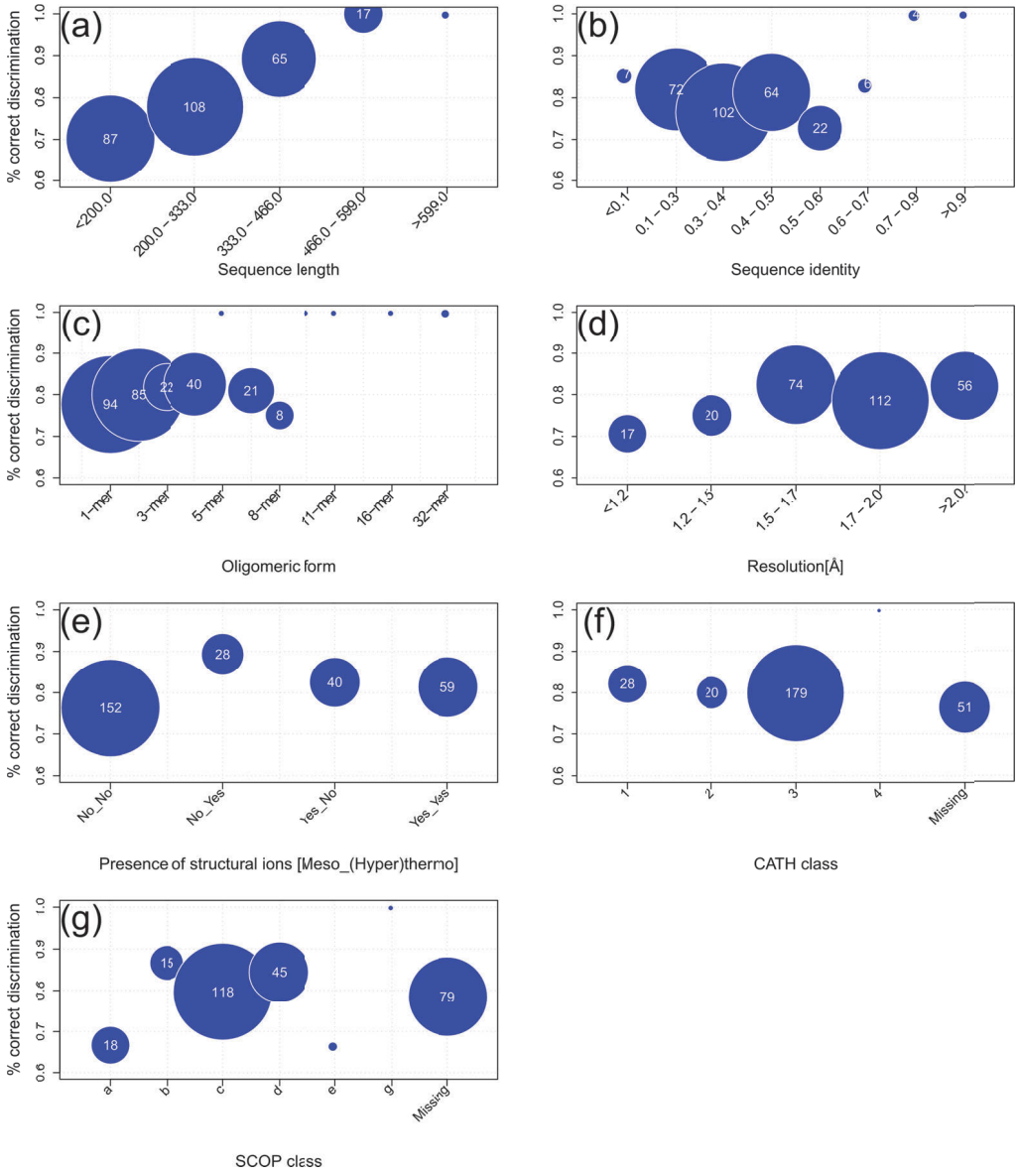


Figure S1. Discrimination accuracy between mesophilic and (hyper)thermophilic protomers, based on clusters of residues with good hydrophobic interaction energies, grouped according to the sequence length (a), sequence identity (b), oligomeric form (c), resolution of the X-ray structure (d), presence of structural ions in the structure (e), CATH class (f), and SCOP class (g). Mesophilic/(hyper)thermophilic pairs were grouped according to the property of the mesophilic protein chain in the structure unless indicated otherwise in the abscissa label of the plot. The size of a circle represents the number of pairs in each group (also indicated by the number in the circle), and the circle's position on the ordinate indicates the percentage of correct discrimination for these pairs.

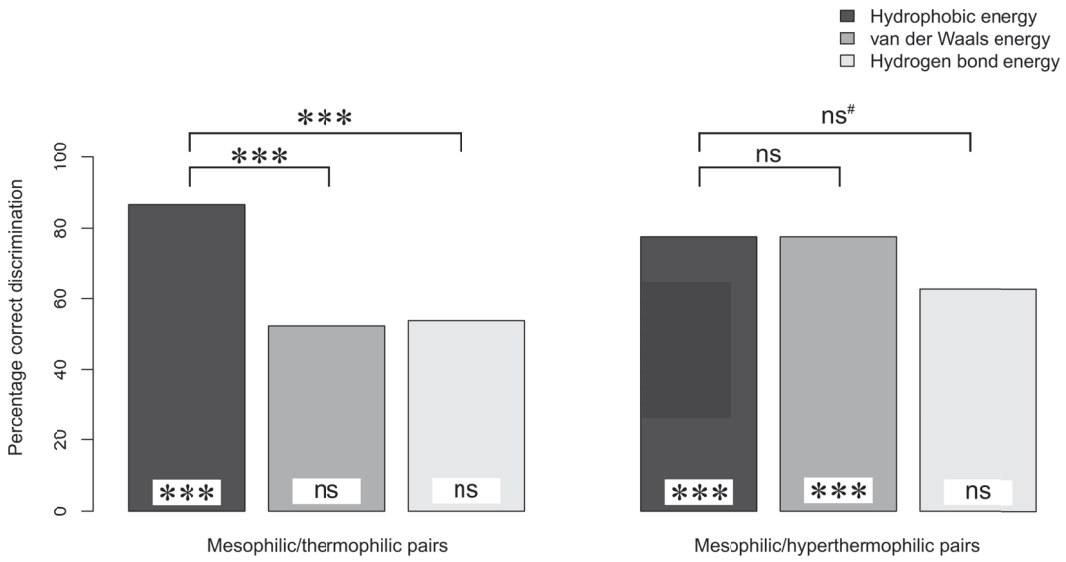


Figure S2. Discrimination accuracy between mesophilic and (hyper)thermophilic protein bio-assemblies based on clusters of residues with good residue-wise energy components. Lines connecting two bars indicate if the difference in discrimination accuracies for the two respective energy components is statistically significant. Marks at the bottom of a column indicate if the discrimination accuracy is significantly different from a random discrimination (50%). The statistical significance of the differences in discrimination accuracies is computed by a bootstrap hypothesis test of equality generating 10000 bootstrap samples; the significance levels are marked by ***: $p < 0.001$; **: $p < 0.01$; ns: $p > 0.05$. The p-value between hydrophobic and hydrogen bond energies in the case of mesophilic/hyperthermophilic pairs is 0.0534 (see ns$^{\#}$ in the figure).

Supplemental references

1. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000,** *28*, 235-242.
2. Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **2013,** *27*, 221-234.
3. Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **2002,** *320*, 597-608.
4. Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. A hierarchical approach to all-atom protein loop prediction. *Proteins: Struct., Funct., Bioinf.* **2004,** *55*, 351-367.
5. Dahiyat, B. I.; Gordon, D. B.; Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci.* **1997,** *6*, 1333-1337.
6. *FIRST, a program for analysing flexibility of networks*, http://flexweb.asu.edu/ (accessed January 17, 2014).
7. Bae, E.; Bannen, R. M.; Phillips, G. N., Jr. Bioinformatic method for protein thermal stabilization by structural entropy optimization. *Proc. Natl. Acad. Sci. USA* **2008,** *105*, 9594-9597.
8. Chun, S. Y.; Strobel, S.; Bassford, P., Jr.; Randall, L. L. Folding of maltose-binding protein. Evidence for the identity of the rate-determining step in vivo and in vitro. *J. Biol. Chem.* **1993,** *268*, 20855-20862.
9. Diamond, D. L.; Strobel, S.; Chun, S. Y.; Randall, L. L. Interaction of SecB with intermediates along the folding pathway of maltose-binding protein. *Protein Sci.* **1995,** *4*, 1118-1123.
10. Prajapati, R. S.; Lingaraju, G. M.; Bacchawat, K.; Surolia, A.; Varadarajan, R. Thermodynamic effects of replacements of Pro residues in helix interiors of maltose-binding protein. *Proteins* **2003,** *53*, 863-871.
11. Chang, Y.; Park, C. Mapping transient partial unfolding by protein engineering and native-state proteolysis. *J. Mol. Biol.* **2009,** *393*, 543-556.

Publication II

Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio

Prakash Chandra Rathi, Sebastian Radestock, and Holger Gohlke
Journal of Biotechnology, **2012**, 159:135-144.
2012 impact factor: 3.18; contribution: 65%

Journal of Biotechnology 159 (2012) 135–144

Contents lists available at SciVerse ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio

Prakash C. Rathi, Sebastian Radestock, Holger Gohlke*

Department of Mathematics and Natural Sciences, Heinrich Heine-University, Düsseldorf, Germany

ARTICLE INFO

Article history:
Received 3 October 2011
Received in revised form 16 January 2012
Accepted 24 January 2012
Available online 1 February 2012

Keywords:
Citrate synthase
Thermostability
Rigidity theory
Flexibility
Protein engineering
Constraint Network Analysis

ABSTRACT

We apply Constraint Network Analysis (CNA) to investigate the relationship between structural rigidity and thermostability of five citrate synthase (CS) structures over a temperature range from 37 °C to 100 °C. For the first time, we introduce an ensemble-based variant of CNA and model the temperature-dependence of hydrophobic interactions in the constraint network. A very good correlation between the predicted thermostabilities of CS and optimal growth temperatures of their source organisms ($R^2 = 0.88$, $p = 0.017$) is obtained, which validates that CNA is able to quantitatively discriminate between less and more thermostable proteins even within a series of orthologs. Structural weak spots on a less thermostable CS, predicted by CNA to be in the top 5% with respect to the frequency of occurrence over an ensemble, have a higher mutation ratio in a more thermostable CS than other sequence positions. Furthermore, highly ranked weak spots that are also highly conserved with respect to the amino acid type found at that sequence position are nevertheless found to be mutated in the more stable CS. As for mechanisms at an atomic level that lead to a reinforcement of weak spots in more stable CS, we observe that the thermophilic CS achieve a higher thermostability by better hydrogen bonding networks whereas hyperthermophilic CS incorporate more hydrophobic contacts to reach the same goal. Overall, these findings suggest that CNA can be applied as a pre-filter in data-driven protein engineering to focus on residues that are highly likely to improve thermostability upon mutation.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Thermostability is an important property of an enzyme as increasing it can widen the enzyme's scope in biotechnological processes (Demirjian et al., 2001; Van den Burg, 2003). In general, enzymes from (hyper)thermophiles, *i.e.*, organisms that grow optimally at a temperature of more than 50 °C (85 °C), show a higher temperature tolerance than their orthologs from mesophiles, *i.e.*, organisms with an optimal growth temperature, T_{og}, of 25–50 °C (Demirjian et al., 2001; Vieille and Zeikus, 2001). However, not all enzymes found in nature are optimized to withstand extreme industrial process conditions including high temperature (Polizzi et al., 2007). The identification and development of thermostable enzymes is therefore an important aspect of research in biotechnology (Haki and Rakshit, 2003). Identifying thermostable enzymes by screening metagenomes is often problematic because of an insufficient and biased expression of the cloned genes in the expression systems (Uchiyama and Miyazaki, 2009). Engineering existing enzymes for improved thermostability is therefore a valuable alternative (Leisola and Turunen, 2007). The latter approach includes random mutagenesis and recombination followed by screening for thermostable mutants (Eijsink et al., 2005), rational design (Eijsink et al., 2004), and data-driven approaches (Chaparro Riggers et al., 2007). While random mutagenesis has a limitation in that only a restricted sequence space can be tested for the desired activity (Lehmann and Wyss, 2001), a rational design requires a thorough understanding of the mechanisms underlying thermostabilization (Eijsink et al., 2004). As a compromise, data-driven approaches are being pursued for reducing the library size for mutagenesis based on suggestions of interesting residue positions that, when mutated, would lead to a more thermostable protein (Chaparro Riggers et al., 2007).

One such data-driven approach has been introduced by Reetz et al. (2006). In that study, saturation mutagenesis was performed in an iterative manner on those residues of lipase A from *Bacillus subtilis* with the highest B-factors. This followed the guiding principle that thermostable proteins usually show a higher degree of structural rigidity than their counterparts from mesophilic organisms; hence, preferentially stabilizing the most mobile regions

Abbreviations: CNA, Constraint Network Analysis; CS, citrate synthase; DCM, distance constraint model; TUS, thermal unfolding simulation; RCD, rigid cluster decomposition; MD, molecular dynamics; PDB, protein data bank; *WSMR*, weak spot mutation ratio.

* Corresponding author at: Universitätsstr. 1, 40225 Düsseldorf, Germany. Tel.: +49 211 8113662; fax: +49 211 8113847.
E-mail address: gohlke@uni-duesseldorf.de (H. Gohlke).

0168-1656/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2012.01.027

should increase thermostability. In fact, a significantly more stable variant of lipase A was identified by screening less than 8000 colonies.

Following the same guiding principle, we have developed a graph theory-based approach termed Constraint Network Analysis (CNA) that predicts protein thermostability by characterizing the mechanical rigidity of a protein structure (Klein et al., 2011; Radestock and Gohlke, 2008, 2011; Rathi et al., 2011). For this, hydrogen bonds are removed in the order of increasing strength from a constraint network representation of a protein, which simulates a thermal unfolding of the protein (Hespenheide et al., 2002; Rader et al., 2002). At a phase transition temperature, T_p, the mechanical rigidity of the network is lost. At this point, the network represents an unfolded protein and, thus, T_p can be related to the melting temperature, T_m, of the system. When applied to a data set of 20 pairs of orthologs from meso- and thermophilic organisms, for 2/3 of the pairs a higher T_p was observed, in agreement with experiment. Furthermore, the approach allows understanding and exploiting the relationship between microscopic structure and macroscopic stability. Thus, it can be used in data-driven protein engineering to increase protein thermostability by introducing mutations at regions that are crucial for macroscopic stability. These regions are referred to as unfolding regions or weak spots. For thermolysin-like protein and 3-isopropylmalate-dehydrogenase, predicted weak spots were indeed involved in improving the proteins' thermostability as demonstrated by comparison to mutagenesis experiments in a retrospective analysis. CNA was also used in a prospective manner on phytase for identifying weak spots that were subsequently mutated, leading to an increase in thermostability in some cases (Radestock, 2010). A related approach is provided by the distance constraint model (DCM) (Jacobs et al., 2003) where an ensemble of constraint topologies is generated by considering mean-field probabilities of hydrogen bonds and torsion constraints in a Monte Carlo sampling. Average stability characteristics are then computed by performing a FIRST analysis on each constraint topology in the ensemble. Note, however, that DCM requires knowledge of experimentally determined heat capacity curves for a protein-specific parameterization of the model (Jacobs and Dallakyan, 2005).

Here, we extend our previous studies in three major aspects. (I) So far, CNA has been applied to only pairs of orthologs. In the present study, we analyze and compare a series of five citrate synthase (CS) proteins from organisms with T_{og} in the range of 37–100 °C. First, from a methodological point of view, this set of structures provides a thorough test for whether the model underlying CNA applies throughout this temperature range. Second, from a structural biology point of view, this set allows deciphering to what extent nature applies different mechanisms for achieving protein thermostability at different temperature stages. (II) Rather than applying CNA to a single crystal structure as done previously, here we introduce an ensemble-based CNA using ensembles generated by molecular dynamics (MD) simulations. The incentive for this is to circumvent the sensitivity of CNA with respect to details in the input structure (Jacobs et al., 2001) and, thus, to provide a more robust way of analysis. Using ensembles of protein conformations has already been shown to yield more robust results in the case of analyzing changes in protein flexibility upon protein-protein complex formation (Gohlke et al., 2004) as well as in the case of predicting the intrinsic flexibility of a protein (Mamonova et al., 2005). Furthermore, the ensemble-based analysis allows ranking predicted weak spots with respect to their frequency of occurrence in the ensemble. When related to information on the frequency of mutations, this allows analyzing whether nature preferentially mutates structural weak spots identified by CNA to achieve higher thermostability. When going from a less thermostable CS to a more thermostable one, it also allows to understand the stepwise thermal stabilization of CS. (III) We refine the model underlying thermal unfolding simulations (TUS) in that now hydrophobic contacts in the constraint network are modeled in a temperature-dependent manner, as has been done so far already for hydrogen bonds and salt bridges. This accounts for the fact that hydrophobic interactions become stronger with increasing temperatures (Privalov and Gill, 1988).

2. Materials and methods

2.1. Data set

CS is one of the rare examples for which crystal structures are available in the Protein Data Bank (PDB) (Berman et al., 2000) from organisms whose living temperatures span an extreme temperature range from 0 °C to 100 °C (Bell et al., 2002). This makes CS a particularly valuable model for understanding the thermal adaptation of proteins. CS is a homodimeric protein with two active sites; each monomer consists of a small and a large sub-domain (Fig. 1). CSs have been crystallized in two different conformations, an "open" *holo* and a "closed" *apo* form (Remington et al., 1982). CS proteins for this study were selected from more than 40 CS entries in the PDB based on the following criteria: (I) Only structures in the open form were considered. (II) Structures whose quality were scored "bad" according to the PDBREPORT database (Hooft et al., 1996) were excluded. (III) More than one structure was available from organisms with T_{og} = 37 °C. Only one structure was chosen from this temperature range.

This resulted in four structures from organisms with T_{og} = 37 °C, 75 °C, 87 °C, and 100 °C, respectively. A CS structure in the open form was not available from any organism having a T_{og} between 37 °C and 75 °C. Hence, a homology model of CS from *Thermoplasma acidophilum* (T_{og} = 59 °C) in the open form was built using the MODELLER software (Sali and Blundell, 1993). For this, the closed form of this CS (PDB ID: 2r26) was used as a template to model the large sub-domain, and the CS in the open form from *Sulfolobus solfataricus* (PDB ID: 1o7x) was used as a template to model the small sub-domain. The final data set contained five CS structures, which are summarized in Table 1. All water molecules and ligands were removed from the structures. All structures were protonated, and side chains of Asn, Gln, and His were flipped if necessary to optimize the hydrogen bond network, using REDUCE (Word et al., 1999). The following abbreviations will be used to refer to CSs, including the T_{og} of their source organisms, throughout this manuscript: *Sus scrofa*: PigCS_37, *T. acidophilum*: TaCS_59, *Thermus thermophilus* HB8: TtCS_75, *S. solfataricus*: SsCS_87 and *Pyrobaculum aerophilum* IM2: PaCS_100.

2.2. Generation of the structure ensemble

MD simulations of all five CS structures were performed using the AMBER 10 suite of programs (Case et al., 2005) together with the parm99 force field (Cornell et al., 1995) with a modification suggested by Simmerling et al. (2002). The system was neutralized by adding sodium counter-ions and solvated with a truncated octahedral box of TIP3P water molecules (Jorgensen et al., 1983) such that the distance between the edges of the water box and the closest atom of the protein was at least 11 Å. The particle mesh Ewald method (Darden et al., 1993) was used with a direct-space non-bonded cutoff of 8 Å. Bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm (Ryckaert et al., 1977), and the time step for all simulations was 2 fs. After equilibration, a production run of unrestrained MD in the canonical ensemble (NVT) was performed to generate a trajectory of 10 ns length, with

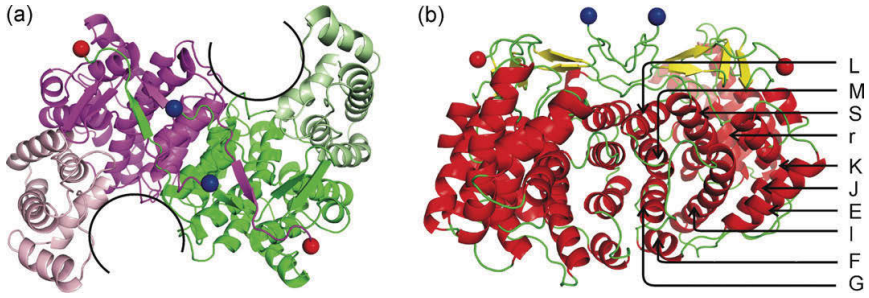


Fig. 1. Cartoon representation of SsCS_87 as an example for CS structures. Two different views (a and b) differ from each other by a rotation of ~90° about a horizontal axis. N- and C-termini are represented as blue and red spheres, respectively. In (a), two monomers are represented in different colors with the small sub-domains colored in a lighter shade. Two active sites are represented by arcs (a). In (b), CS is colored by secondary structure elements (α-helices in red, β-sheets in yellow, and loops in green). α-helices referred to in the text are labeled, with α-helices of the small sub-domain denoted in lower case letters. All figures of CS structures were generated with PyMOL (www.pymol.org).

conformations extracted every 40 ps from the last 8 ns. This resulted in 200 conformations that were subjected to CNA.

2.3. Construction of constraint networks and rigid cluster decomposition

The folded state of a protein is stabilized by non-covalent interactions (Dill, 1990), and proteins can be modeled as molecular networks to study these stabilizing features (Böde et al., 2007; Greene and Higman, 2003). In this study, we go beyond a topological network representation by modeling proteins as constraint networks (also referred to as molecular frameworks). As such, a protein structure is modeled as a *body-and-bar* network where each atom is considered as a rigid body having six degrees of freedom (Hespenheide et al., 2004). Any number of bars between one and six can be placed between two such atoms to represent an interaction, and every such bar removes one degree of freedom. A covalent bond is modeled as five bars, allowing for a dihedral rotation about it. Peptide bonds and double bonds are modeled with six bars, disallowing any bond rotation. Hydrogen bonds and salt bridges, together referred to as hydrogen bonds in this study, were modeled with five bars whereas hydrophobic interactions were modeled with two bars. Weaker interactions such as van der Waals interactions are not modeled as constraints. These constraint networks were constructed using the FIRST software (version6.2) (Jacobs et al., 2001). A hydrogen bond energy E_{hb} is calculated using a geometry-based empirical function (Dahiyat et al., 1997), and only hydrogen bonds with a lower energy (*i.e.*, a higher stability) than a certain cutoff $E_{cut,hb}$ are included in the network. Hydrophobic contacts are considered between all carbon and sulfur atoms separated by a distance less than the sum of their van der Waals radii (1.7 Å for C and 1.8 Å for S) plus a certain cutoff $D_{cut,hp}$. Cutoff values for inclusion of hydrogen bonds and hydrophobic contacts can vary with temperature, as described in more detail in Section 2.4.

The pebble game algorithm (Jacobs, 1998; Jacobs and Hendrickson, 1997; Jacobs and Thorpe, 1995) implemented in the FIRST program (Jacobs et al., 2001) is then applied to characterize the rigidity of such networks by constraint counting. FIRST determines whether a bond is either flexible or rigid and, subsequently, decomposes the constraint network into rigid clusters and flexible regions. A rigid cluster is a set of atoms that move together as a rigid body in any floppy motion. Atoms that are not part of a rigid cluster are in a flexible region. The size of a rigid cluster is defined by the number of atoms in it. This approach has been successfully applied for characterizing rigidity in proteins (Ahmed and Gohlke, 2006; Gohlke et al., 2004; Gohlke and Thorpe, 2006; Hespenheide et al., 2002; Jacobs et al., 2001; Rader and Bahar, 2004; Rader et al., 2002), RNAs (Fulle and Gohlke, 2008, 2009a), and the ribosome (Fulle and Gohlke, 2009b).

2.4. Thermal unfolding simulation

During the thermal unfolding of a protein, non-covalent bonds are broken sequentially until the protein becomes unfolded with very few non-covalent interactions remaining. This can be simulated by gradually removing non-covalent constraints from the constraint network and applying the pebble game algorithm to each of the resulting networks (Hespenheide et al., 2002; Rader et al., 2002; Radestock and Gohlke, 2008, 2011).

In previous studies, only hydrogen bonds were removed from the network in the order of increasing strength to simulate an increase in temperature. As such, at a temperature T, all hydrogen bonds with $E_{hb} \geq E_{cut,hb} = (300\,\mathrm{K} - T) \times (\mathrm{kcal\,mol^{-1}})/20\,\mathrm{K}$ were removed. The relation between T and $E_{cut,hb}$ had been determined previously (Radestock and Gohlke, 2008, 2011). For hydrophobic contacts, a temperature-independent $D_{cut,hp} = 0.25$ Å was used (Radestock and Gohlke, 2008, 2011). We will refer to this type of thermal unfolding simulation as TUS1.

Table 1
CS proteins used in this study.

PDB ID	Source organism	T_{og}[a]	Sequence length	Resolution[b]	Reference
3ENJ	*Sus scrofa*	37	437	1.78	Larson et al. (2009)
2R26[c]	*Thermoplasma acidophilum*	59	384	2.50[d]	Darland et al. (1970)
1IOM	*Thermus thermophilus* HB8	75	377	1.50	Oshima and Imahori (1974)
1O7X	*Sulfolobus solfataricus*	87	377	2.70	Bell et al. (2002) and Zillig et al. (1980)
2IBP	*Pyrobaculum aerophilum* IM2	100	409	1.60	Boutz et al. (2007) and Volkl et al. (1993)

[a] In °C.
[b] In Å.
[c] A homology model in the open form was used.
[d] Resolution of the structure in the closed form upon which the model in the open form was built.

In the present study, we modify the TUS procedure by modeling hydrophobic contacts as temperature-dependent, too. As the strength of hydrophobic interactions increases with increasing temperature (Privalov and Gill, 1988; Schellman, 1997), we include more and more hydrophobic contacts into the network by increasing $D_{cut,hp}$ with increasing temperature. Preliminary tests showed that the number of hydrophobic contacts per atom increases roughly linearly when modifying $D_{cut,hp}$ from 0.25 Å to 0.40 Å. $D_{cut,hp} > 0.40$ Å resulted in very rigid networks that were not amenable to TUS. Therefore, we decided to linearly increase $D_{cut,hp}$ from 0.25 Å at 300 K to 0.40 Å at 420 K. We will refer to this new type of thermal unfolding simulation as TUS2.

2.5. Identification of the folded-unfolded transition and of weak spots

When going from a rigid network at low temperature to a flexible network at high temperature, a rather pronounced phase transition is observed. At this point, the percolating ("giant") rigid cluster stops dominating the network, and many smaller rigid clusters appear. Such a percolation can be observed in both network glasses (Rader et al., 2002) and proteins (Rader et al., 2002; Radestock and Gohlke, 2008, 2011). However, the percolation behavior of proteins is more complex than that of glasses in that multiple phase transitions can be observed in the former (Rader et al., 2002; Radestock and Gohlke, 2008, 2011), in contrast to a single transition in the latter (Rader et al., 2002). This can be understood by the fact that protein structures are modular, *i.e.*, they consist of secondary structure elements, sub-domains, and domains, which often break away from the giant cluster as a whole. As demonstrated by us, the last one of these transitions is most relevant from a structural biology point of view in that this transition relates to a protein going from the folded to the unfolded state (Radestock and Gohlke, 2011). The temperature T_p related to this phase transition is identified as the inflection point in a curve of the cluster configuration entropy H *versus* T, after fitting a five-parameter double sigmoid function (Cairns et al., 2008) to value pairs (H, T) determined by TUS. H has been introduced by Andraud et al. (1994) as a morphological descriptor for heterogeneous materials and is adapted from Shannon's information theory. Here, the definition of H as given in Radestock and Gohlke (2008, 2011) is used.

Once the biologically relevant folded-unfolded transition is observed, locally weak regions (weak spots) in the constraint network are identified. For this, rigid cluster decompositions (RCD) directly before and after this folded-unfolded transition are compared. Residues for which C_α atoms are part of the giant cluster before the transition, and that become flexible afterwards, are considered weak spots. Here, a residue is considered flexible if its C_α atom is either in a flexible region or part of a small rigid cluster of less than four atoms. The identification of weak spots is carried out for each snapshot of the ensemble individually. This leads to a set of residues being predicted as weak spots for each snapshot. These sets do not necessarily contain the same residues across different snapshots as the RCD may vary with the conformation of the protein. The frequency of all residues being predicted as a weak spot throughout the ensemble is counted and, finally, all weak spots are assigned a rank according to the decreasing order of their frequency.

The thermal unfolding simulation, determination of T_p, and identification of weak spots is performed by the CNA software package developed in our group (Radestock and Gohlke, 2008, 2011; Rathi et al., 2011), which functions as a front- and back-end to the FIRST software (Jacobs et al., 2001).

2.6. Mutation ratio of weak spot residues

For validating the weak spots predicted by CNA and tracing the stepwise thermal stabilization of CS, a weak spot mutation ratio (*WSMR*) is computed for each weak spot up to rank r by comparing pairs of proteins with distinct thermostability.

$$WSMR_r = \frac{\sum_{i=1}^{r} \mathrm{Res}_{i,\mathrm{mutated}}}{\sum_{i=1}^{r} \mathrm{Res}_{i,\mathrm{conserved}}} \times \frac{\text{total conserved residues}}{\text{total mutated residues}} \quad (1)$$

Here, $Res_{i,\mathrm{mutated}}$ is the number of residues in a weak spot of rank i of the protein with lower thermostability that have been mutated in the more thermostable protein; conversely, $Res_{i,\mathrm{conserved}}$ is the number of residues in a weak spot of rank i of the protein with lower thermostability that are conserved in the more thermostable protein. The second term is a normalization factor regarding the total numbers of mutated and conserved residues between both proteins, irrespective of their assignment to weak spots. For determining whether a residue has been mutated or remained conserved between pairs of CS structures, a multiple sequence alignment of 549 CS sequences extracted from the UniProt database (http://www.uniprot.org) was used.

If $WSMR_r > 1$, residues in weak spots up to rank r have been found to be more frequently mutated on going from the less thermostable to the more thermostable protein compared to the average mutation ratio for this pair of proteins. Conversely, if $WSMR_r < 1$, weak spots up to rank r have a mutation ratio that is lower than the average mutation ratio for this pair of proteins.

2.7. Position-specific conservation of residues

We calculated the conservation of an amino acid a at position i, independent of all other positions, for all amino acids in the multiple sequence alignment. The conservation was calculated in terms of a relative entropy, $D_i^{(a)}$ (Cover and Thomas, 2006) as also used by Ranganathan and coworkers (Halabi et al., 2009; Lockless and Ranganathan, 1999). $D_i^{(a)}$ is the divergence of the observed frequency of a at i ($f_i^{(a)}$) from the background frequency of a in all proteins and rises more and more steeply as $f_i^{(a)}$ approaches one (see Halabi et al., 2009 for more details). This was done to validate the weak spot prediction on the account that a prediction of a highly conserved amino acid as a weak spot in a less stable CS, which is then mutated in a more thermostable CS, is more significant than if the weak spot were predicted at a position that has already been mutated in many CSs.

3. Results and discussion

3.1. Data set

The data set used for this study contains CSs from five different organisms, which have T_{og} from 37 °C to 100 °C (Table 1). Only structures in the open form are considered for the analysis because this is the prevalent form in solution as evident from the crystallization of CS in the open form in the absence of a ligand (Remington et al., 1982; Wiegand and Remington, 1986); in turn, a closed form is only adopted when the substrate oxaloacetate is bound. This has also been supported by a molecular dynamics study that suggested that the closed form is separated from the open form by a large energy barrier, and, hence, is inaccessible to the open form without substrate binding (Daidone et al., 2004). Thus, the open state predominantly determines the overall stability of CS. Unfortunately, melting temperatures (T_m) were not available for all five CSs, and, hence, the T_{og} of their source organisms were used as a descriptor of thermostability to which T_p predictions by CNA will be

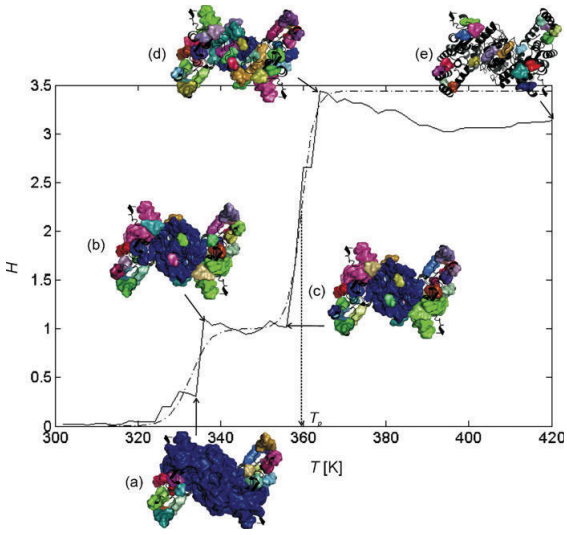


Fig. 2. Rigidity percolation of a TtCS_75 structure monitored by an *H versus T* plot (continuous line). Rigid clusters just before (a and c) and after (b and d) the two-phase transitions and at a temperature of 420 K (e) have been depicted as uniformly colored bodies. The blue body represents the largest ("giant") cluster. The folded–unfolded transition (T_p) is identified by the inflection point of the second sigmoid (broken line) on the temperature axis.

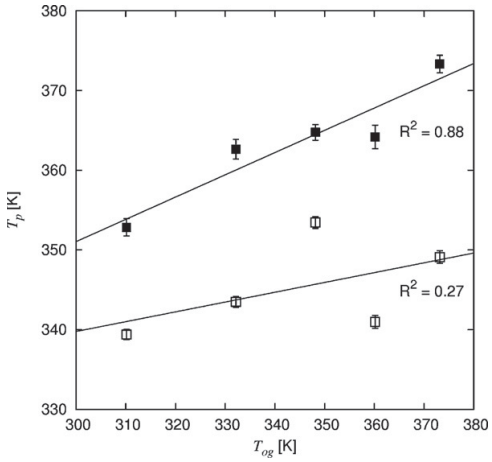


Fig. 3. Correlation between T_p and T_{og}. T_p values obtained with TUS1 are marked by empty squares whereas filled squares denote T_p values obtained with TUS2. Error bars represent the standard error in the mean. Least squares fit lines have been drawn for both the correlations.

compared. Optimal growth temperatures have previously been used in studies comparing meso- and thermophilic proteins (Gromiha et al., 1999; Radestock and Gohlke, 2008, 2011). CSs are almost entirely α-helical with the PigCS_37 containing 20 α-helices, four more than any other CS. The dimer interface of all CSs primarily consists of an eight-fold α-helical sandwich composed of four antiparallel pairs of helices (F, G, M and L). Additionally, N- and C-termini make interactions with the respective other monomer (Fig. 1). The CS structures (except TsCS_59, which is a homology model) have a resolution in the range of 1.60–2.70 Å. All CSs used in this study are structurally very similar as shown by C_α atom root mean-square deviations of 1.22–2.32 Å. The pairwise sequence identities lie between ∼20 and ∼60%.

3.2. Loss of rigidity percolation in CS structures upon thermal unfolding

In general, the loss of rigidity in proteins appears as multiple phase transitions on going from a folded to an unfolded state, owing to the modular architecture of protein structures. For most conformations of CSs, two prominent transitions are observed in the plot of *H versus T* (Fig. 2). The first transition corresponds to the appearance of a rigid core formed by helices G, I, L, M, S and loops connecting these (which are all part of the large sub-domain) from both monomers (RCD (b) in Fig. 2), originating from an almost rigid network (RCD (a) in Fig. 2). This rigid core across the dimer interface suggests strong interactions at the interface. Active site residues from the large sub-domain are also part of this rigid core. This particular transition is not relevant here because the protein network after the transition still reflects a structurally stable protein, not an unfolded one. The second transition involves a breakdown of this rigid core (RCD (c) in Fig. 2) into many smaller rigid clusters and flexible links in between (RCD (d) in Fig. 2) so that now the rigidity across the interface is lost and no folding core is left. After this transition, the network becomes largely flexible. Thus, this second transition is the relevant one with respect to going from a structurally stable, folded state of the protein to an unfolded one, so that the temperature T_p associated with it relates to the experimental T_m and indicates the thermostability of the protein (Radestock and Gohlke, 2008, 2011). Finally, at a very high temperature, the network becomes fully flexible (RCD e) in Fig. 2.

3.3. Macroscopic analysis of constraint networks: thermostability prediction

From the TUS, the thermostability for all CSs was predicted according to the computed T_p values, which are ensemble averages over 200 conformations per CS structure. When correlated to T_{og} values, no significant correlation ($R^2 = 0.27$, $p = 0.374$) was found if the thermal unfolding simulation type TUS1 was applied (Fig. 3). An even worse correlation ($R^2 = 3 \times 10^{-5}$, $p = 0.992$) was obtained if only the single X-ray structures (or the homology model in the case of TaCS_59) were used instead of the ensembles of structures. The absolute values of predicted T_p were lower in comparison to T_{og} with a slope of 0.12 for five CSs analyzed. In particular, the phase transition temperatures of the hyperthermophilic CSs, SsCS_87 and PaCS_100, were predicted too low when compared to the respective T_{og}, whereas TtCS_59 was predicted as the most stable CS. A reason for this may be that the T_p rely on the empirical relationship between *E* and *T* given in Section 2. This relationship has been determined for a dataset of orthologous meso- and thermophilic protein pairs with T_{og} in the range of 30–83 °C. The deviation indicates that a system-specific reparametrization might be necessary here. Note however, that our primary goal here is to predict a correct rank ordering of the CS structures according to their thermostabilities.

Initially, we examined the first transition points on the *H vs. T* curves (Fig. 2) for all CSs to see if these provide a better correlation with T_{og} than the T_p values (obtained for the last transition). No significant improvement of the correlation was found, however (data not shown). Then, we anticipated that the mispredition may arise from neglecting the effect of temperature on the strength of hydrophobic contacts in the constraint network model underlying TUS1. While this model has worked well for predicting relative thermostabilities for pairs of meso- and thermophilic proteins (Radestock and Gohlke, 2008, 2011), we note that in neither of these studies proteins with a T_m or T_{og} as high as 100 °C were included, nor were differences in the thermostability as large as

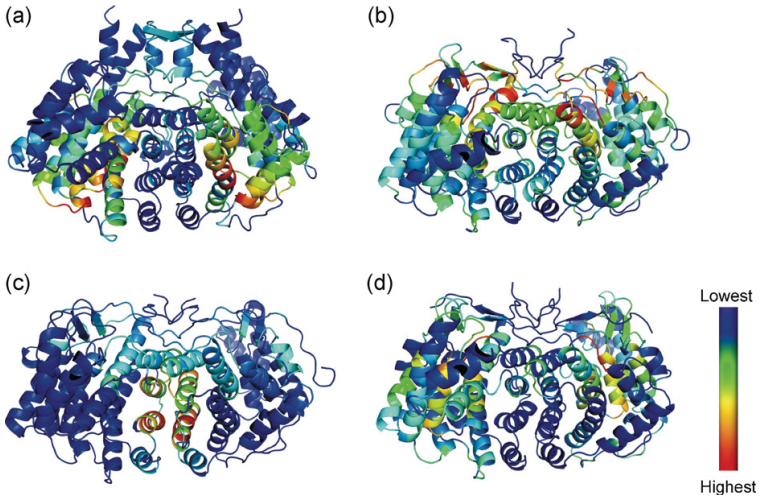


Fig. 4. Mapping of predicted weak spots on CS structures using a color range from red (highest ranking weak spot) to blue (lowest ranking weak spot). Weak spots for four CSs are presented: PigCS_37 (a), TaCS_59 (b), TtCS_75 (c), and SsCS_87 (d).

63 °C considered, in contrast to the present study. Notably, the contribution of hydrophobic interactions to the free energy of protein folding increases with increasing temperatures (Privalov and Gill, 1988). Accordingly, we have devised a new constraint network representation for thermal unfolding simulations, which now takes into account the strengthening of hydrophobic contacts in that the number of these contacts linearly increases with temperature (see Section 2).

Thermal unfolding simulations of type TUS2 on single X-ray structures (or the homology model in the case of TaCS_59) of CS still resulted in an insignificant correlation between T_p and T_{og} ($R^2 = 0.35$, $p = 0.295$). However, the thermostability prediction for the CS structures improves considerably when applying TUS2 to ensembles of CS structures: the T_p for SsCS_87 and PaCS_100 increase by ∼24 °C, whereas the T_p for the three less stable CS (PigCS_37, TaCS_49 and TtCS_75) increase by only ∼15 °C on average. This results in a significant and very good correlation with respect to experimental T_{og} ($R^2 = 0.88$, $p = 0.017$) (Fig. 3). The absolute values of predicted T_p were higher than T_{og} for PigCS_37, TaCS_59, and TsCS_75, indicating that an estimate of absolute T_p might need to consider both $E_{cut,hb}$ and $D_{cut,hp}$ (rather than rely on $E_{cut,hb}$ only as in the present empirical relationship). However, it is more important for this study that the order of thermostability was correctly predicted except for TtCS_75 and SsCS_87, the T_p of which were computed to be roughly equal. Notably, the analysis of the constraint networks at a macroscopic level already suggests that the hyperthermophilic CS adapt to high temperature by incorporating more hydrophobic contacts, that way rigidifying their structures. Indeed, the constraint network of PaCS_100 contains, as an average over 200 conformations, 437 hydrophobic contacts at $D_{cut,hp} = 0.25$ Å, which are at least 20 (∼5%) more than in any other CS constraint network.

3.4. Microscopic analysis of constraint networks: weak spot prediction

Once the biologically relevant folded-unfolded transition is identified, structurally weak regions can be located in the protein structures. The weak spots are identified as those residues for which C_α atoms are part of the giant cluster before the transition but are in a flexible region afterwards. The weak spots are ranked according to their frequency of occurrence throughout the structural ensemble of the CS. Thus, higher ranked weak spots occur more often and, hence, should be primarily considered as sites for performing (saturation) mutagenesis in order to generate a more thermostable variant of the protein.

Weak spots for four of the CS structures for which a corresponding more thermostable CS was analyzed are shown in Fig. 4. The weak spots of PigCS_37 and TaCS_59 are found in similar regions of the structures, *i.e.*, the structurally weakest regions are found in both cases on the helixes I and S, which are located directly below helices L, M, G and F at the dimer interface (Fig. 4a and b). Note that the weak spot locations do not necessarily mean that helices themselves weakened upon temperature increase; rather (tertiary) interactions with surrounding helices were broken at T_p, which causes the decay of the giant cluster. The finding of weak spots in similar regions supports the principle of "corresponding states" (Jaenicke and Böhm, 1998; Somero, 1978) according to which homologs from mesophilic and thermophilic organisms are in corresponding states of similar rigidity and flexibility at their respective optimal temperatures. Additionally, TaCS_59 has a few prominent weak spots on the interfacial helix L (Fig. 4b). However, in general, the other interfacial helices are more rigid than other structural parts in these two structures. Surprisingly, weak spots of TtCS_75 predominantly lie on the helices G, L, and M at the dimer interface (Fig. 4c) rather than within the monomers as found for PigCS_37 and TaCS_59; these interface regions have been reinforced in ScCS_87 by incorporating, on average, four additional hydrophobic contacts. Finally, a loop K–L is predicted to be the weakest part in SsCS_87 apart from residues on helix R (Fig. 4d). Thus, the spatial

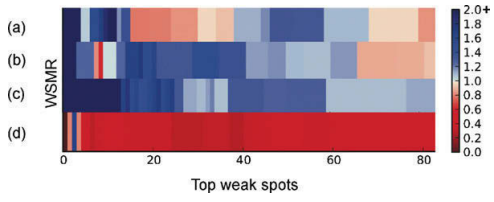


Fig. 5. Weak spot mutation ratios for PigCS_37 *vs.* TaCS_59 (a), TaCS_59 *vs.* TtCS_75 (b), TtCS_75 *vs.* SsCS_87 (c), and SsCS_87 *vs.* PaCS_100 (d).

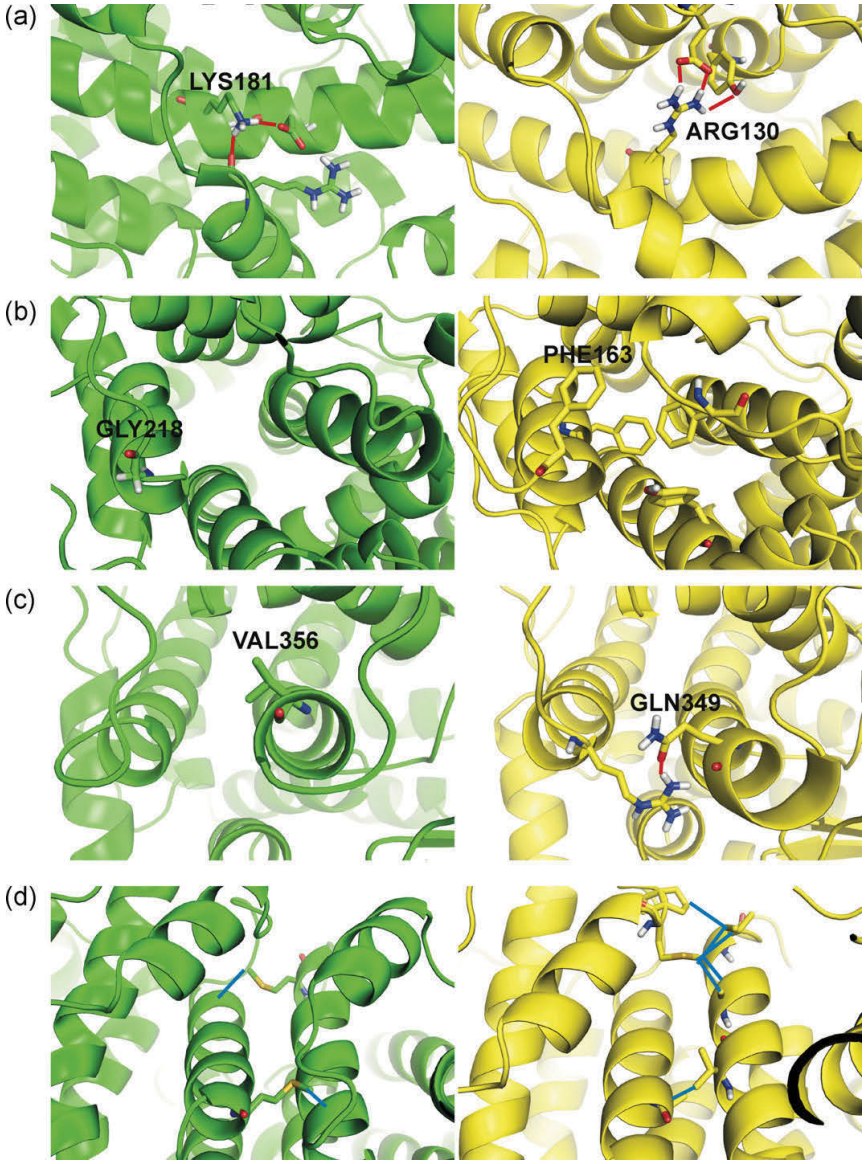


Fig. 6. Top ranking weak spots in a less stable (green) CS are compared with mutated residues in a more stable CS (yellow) for PigCS_37 *vs.* TaCS_59 (a and b), TaCS_59 *vs.* TtCS_75 (c), and TtCS_75 *vs.* SsCS_87 (d). Hydrogen bonds are represented as red lines whereas hydrophobic contacts are represented with blue lines.

distribution of weak spots in SsCS_87 resembles those in PigCS_37 and TaCs_59 in that the interface is stable whereas weak spots are predominantly found on helices within the monomers.

3.5. Validation of weak spot predictions by analyzing sequence information on stepwise CS thermal adaptation

To validate the weak spot predictions by CNA and to trace the stepwise thermal adaptation in the series of CS structures, the sequence of a less stable CS is compared to that of the next more stable CS (*i.e.*, PigCS_37 *vs.* TaCS_59, TaCS_59 *vs.* TtCS_75, TtCS_75 *vs.* SsCS_87, and SsCS_87 *vs.* PaCS_100) with respect to whether residues in weak spots are more frequently mutated than others. For this, the pairwise sequence alignments were extracted from a multiple sequence alignment of 549 CS sequences. For each weak spot rank, a cumulative weak spot mutation ratio $WSMR_r$ (see Section 2.6) was computed. $WSMR_r > 1$ shows that, for weak spots up to rank r, residues in these weak spots are more frequently mutated on going from the less thermostable to the more thermostable protein compared to the average mutation rate for this pair of proteins. Phrased differently, $WSMR_r > 1$ means that CNA successfully identified those (structurally weak) parts of CS where (thermostabilizing) mutations occur more preferentially. This ratio has been plotted against the number of top ranked weak spots in Fig. 5.

Remarkably, *WSMR* for the top 9 weak spot ranks (~5% of the total number of residues) are consistently larger than 1 for all consecutive CS pairs except SsCS_87 *vs.* PaCS_100 (Fig. 5), with average *WSMR* values over the top 9 weak spots ranks of 1.71 (PigCS_37 *vs.* TaCS_59), 1.30 (TaCS_59 *vs.* TtCS_75), 2.07 (TtCS_75 *vs.* SsCS_87), and 0.62 (ScCS_87 *vs.* PaCS_100). This observation has an important implication for applying CNA in data-driven protein engineering: It suggests for the first three pairs that if only the top 9 weak spots had been considered for (saturation) mutagenesis in a less stable CS, those residues would have been preferentially picked that also show an above-average mutation rate in the corresponding, more thermostable CS found in nature. As these weak spots comprise at most 5% of the CS residues, restricting mutations to weak spots apparently leads to focusing on those residues that have a high propensity to improve thermostability upon mutation. For PaCS_100, a rare mechanism of thermostabilization has been reported where a strategically placed disulfide bond within each CS monomer results in a topological cross-link of the two chains. This abrogates the separability of the chains and so leads to an increase in the thermostability (Boutz et al., 2007). It may thus come as no surprise that $WSMR_r > 1$ is only found for a few ($r = 3$) weak spots in the case of the SsCS_87 *vs.* PaCS_100 pair (Fig. 5).

3.6. Validation of weak spot prediction considering sequence conservation across multiple sequences

For further verification of the weak spot predictions, we analyzed the complete multiple sequence alignment based on the following account: The finding that a residue in a weak spot of a less stable CS is mutated in a more stable CS is all the more significant the higher the degree of sequence conservation for this residue is at this position across the multiple sequence alignment. The sequence conservation was calculated in terms of a relative entropy (see Section 2.7). We considered a residue highly conserved at its position if the value of D is greater than 1.0 in the multiple sequence alignment, which corresponds to a frequency of occurrence of >40% for the least frequent amino acid Trp and to a frequency of occurrence of >65% for the most frequent amino acid Ala. Indeed, residues in such weak spots that showed a high degree of sequence conservation are predicted at high ranks, *i.e.*, at ranks 1, 2, 4, and 7 for PigCS_37, at ranks 1, 3, 6, and 10 for TaCS_59 and at ranks 1, 12, and 18 for TtCS_75. For SsCS_87, such weak spots are still found at ranks 9, 13, and 19. Identifying such generally conserved amino acids as weak spots that are then mutated in a more thermostable CS is significant in that such positions constitute <4% of the total number of residues of CS. Thus, these identified weak spots at highly conserved sequence positions do strongly affirm the predictive power of CNA. Furthermore, as it would be difficult to identify such positions from sequence information alone, this demonstrates the added value of applying a structure-based approach as done with CNA.

3.7. Structural basis of weak spot reinforcement

Finally, we analyzed in atomic detail by what mechanisms weak spots are reinforced in more thermostable CS. This was done for those pairs of structures for which the prediction of weak spots was successful, *i.e.*, where $WSMR > 1.0$ was found for weak spots on high ranks. For the analysis, interactions of weak spot residues in the less stable CS were compared to interactions of the residue at the same positions in the corresponding, more stable CS.

In the case of PigCS_37 *vs.* TaCS_59, a better hydrogen bonding network and the formation of an aromatic cluster contribute to the higher thermostability. As such, residue Lys181 (helix I) in PigCS_37, which is highly conserved ($D = 1.87$) and predicted to be in a weak spot of rank two, has only one hydrogen bond with Arg117 (loop E–F) (besides one additional hydrogen bond within its own helix). In contrast, Arg130 at the same position in TaCS_59 forms a hydrogen bond with Tyr207 (helix M) and another one with Glu110 (loop G–I) (Fig. 6a), that way connecting two parts of the structure that are topologically distant. As another pronounced mutation, Gly218 (loop J–K) in PigCS_37, predicted to be in a weak spot of rank two, is replaced by Phe165 in TaCS_59. The side chain of the latter engages in the formation of an aromatic cluster (Fig. 6b). The formation of such aromatic clusters has been described to be among the major factors that lead to higher thermostability (Kannan and Vishveshwara, 2000; Puchkaev et al., 2003). In the case of TaCS_59 *vs.* TtCS_75, residue Gln349 (helix S) in TtCS_75 is involved in a hydrogen bond with Arg114 (helix I), whereas no such interaction is found for the corresponding Val356 in TaCS_59, predicted to be in a weak spot of rank three (Fig. 6c). Additionally, a highly conserved residue Thr21 ($D = 1.69$) in TaCS_59 predicted at weak spot rank three is mutated into Cys in TtCS_75, which is tightly packed to neighboring carbon atoms to form hydrophobic contacts. Finally, the top ranking weak spots in TtCS_75 are predicted to be at the dimer interface. Accordingly, six interfacial hydrophobic contacts have been identified in SsCS_87, in contrast to only two hydrophobic contacts in TtCS_75 (Fig. 6d). Notably, the highly conserved residues Met92 ($D = 3.37$) and Val99 ($D = 1.10$) in the interfacial helix cluster of TtCS_75 are mutated in SsCS_87. The higher degree of hydrophobicity at the SsCS_87 dimer interface has also been reported in a comparison of five CSs by Bell et al. (2002). The latter finding, based on analyzing microscopic properties in the constraint networks, points again to an increased number of hydrophobic contacts as a mechanism of hyperthermophilic CSs to maintain their structural rigidity at high temperature; this is similar to the above finding, based on analyzing macroscopic properties in terms of phase transitions, according to which appropriate T_p could only be predicted after strengthening hydrophobic contacts at high temperature.

4. Conclusions

In this study, we have analyzed and compared the stepwise thermal adaptation of CSs from five different organisms with T_{og} from 37 °C to 100 °C using the graph theory-based Constraint Network Analysis (CNA). In this way, the present study extends our previous studies (Radestock and Gohlke, 2008, 2011) wherein only pair-wise comparisons between orthologous proteins from a mesophilic and a thermophilic organism were performed. From a methodological point of view, CNA is advanced in that now multiple conformations of a protein (generated from MD simulations, but they could also be taken from an experimental source) are analyzed in order to circumvent the problem of sensitivity of CNA on the quality of the input structure. Indeed, this procedure allowed applying CNA to a homology model of TaCS_59 and correctly predicting its thermostability along with four other CSs. A further methodological advancement in CNA has been to model the effect of temperature on hydrophobic interactions during TUS, which helped correctly predicting the thermostability of (hyper)thermophilic CSs.

CNA correctly predicted the thermostability of five CSs with a correlation of $R^2 = 0.88$ between T_p and T_{og}. Most significantly, for the first time, we have analyzed weak spots predicted by CNA with respect to the mutation ratio at these locations on going from a less thermostable to a more thermostable CS. We have also analyzed the weak spots with respect to their degree of conservation in a multiple sequence alignment. Remarkably, weak spots predicted by CNA for a less stable CS were found to be indeed more often mutated in a more stable CS in three out of four cases. Even more convincingly, weak spots at very high ranks and that are highly conserved were nevertheless mutated in the more stable CS. These predictions render CNA a useful pre-filter in protein engineering projects that aim

at developing thermostable mutants because that way the number of residues where mutations should be preferentially introduced can be significantly (to just ∼5% of the total number of residues in the case of CS) reduced. Hence, even suggestions for multiple mutations are in place because the number of possible combinations still remains small that way. Finally, the present study elucidates mechanisms at an atomic level that lead to a reinforcement of weak spots in more stable CS and, hence, to improved thermostability. As such, we observed that the thermophililic CSs achieve a higher thermostability by better hydrogen bonding networks whereas hyperthermophilic CSs incorporate more hydrophobic contacts to reach the same goal.

As for shortcomings of our method, we note that a structure of high quality is a prerequisite for a meaningful analysis although the ensemble approach introduced here alleviates this to a large extent. Another limitation lies in the fact that all hydrogen bonds remove the same number of degrees of freedom from the network, irrespective of their strength, and so do the hydrophobic contacts. Refining the constraint network in terms of constraints with varying influence on the structural flexibility/rigidity will likely be a valuable goal to pursue. As yet another limitation, extrinsic factors like glycosylation, salt concentration, pressure effects, and solvent viscosity, which may influence the thermostability of a protein, are not considered in CNA. Finally, CNA does not yet predict actual amino acid substitutions at the weak spots that would improve the thermostability. Despite all this, we are convinced that being able to identify weak spots in proteins of low thermostability already makes CNA a valuable method for prospective studies that aim at improving the thermostability of a protein by means of data-driven protein engineering.

Acknowledgements

We are grateful to the Ministry of Innovation, Science, and Research of North Rhine-Westphalia and Heinrich-Heine-University Düsseldorf for a scholarship to PCR within the CLIB-Graduate Cluster Industrial Biotechnology. We are grateful to Doris L. Klein and Christopher Pfleger (Heinrich-Heine-University, Düsseldorf) for fruitful discussions. We acknowledge the "Zentrum fuer Informations- und Medientechnologie" (ZIM) at the Heinrich-Heine-University, Düsseldorf, for computational support.

References

Ahmed, A., Gohlke, H., 2006. Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. Proteins 63, 1038–1051.
Andraud, C., Beghdadi, A., Lafait, J., 1994. Entropic analysis of random morphologies. Physica A 207, 208–212.
Bell, G.S., Russell, R.J.M., Connaris, H., Hough, D.W., Danson, M.J., Taylor, G.L., 2002. Stepwise adaptations of citrate synthase to survival at life's extremes. Eur. J. Biochem. 269, 6250–6260.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.
Böde, C., Kovács, I.A., Szalay, M.S., Palotai, R., Korcsmáros, T., Csermely, P., 2007. Network analysis of protein dynamics. FEBS Lett. 581, 2776–2782.
Boutz, D.R., Cascio, D., Whitelegge, J., Perry, L.J., Yeates, T.O., 2007. Discovery of a thermophilic protein complex stabilized by topologically interlinked chains. J. Mol. Biol. 368, 1332–1344.
Cairns, S.P., Robinson, D.M., Loiselle, D.S., 2008. Double-sigmoid model for fitting fatigue profiles in mouse fast- and slow-twitch muscle. Exp. Physiol. 93, 851–862.
Case, D.A., Cheatham III, T.E., Darden, T., Gohlke, H., Luo, R., Merz Jr., K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.J., 2005. The Amber biomolecular simulation programs. J. Comput. Chem. 26, 1668–1688.
Chaparro Riggers, J.F., Polizzi, K.M., Bommarius, A.S., 2007. Better library design: data driven protein engineering. Biotechnol. J. 2, 180–191.
Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., Kollman, P.A., 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. 117, 5179–5197.
Cover, T.M., Thomas, J.A., 2006. Elements of Information Theory, second ed. Wiley-Interscience, New Jersey.
Dahiyat, B.I., Gordon, D.B., Mayo, S.L., 1997. Automated design of the surface positions of protein helices. Protein Sci. 6, 1333–1337.
Daidone, I., Roccatano, D., Hayward, S., 2004. Investigating the accessibility of the closed domain conformation of citrate synthase using essential dynamics sampling. J. Mol. Biol. 339, 515–525.
Darden, T., York, D., Pedersen, L., 1993. Particle mesh Ewald: an N.log (N) method for Ewald sums in large systems. J. Chem. Phys. 98, 10089–10092.
Darland, G., Brock, T.D., Samsonoff, W., Conti, S.F., 1970. A thermophilic, acidophilic mycoplasma isolated from a coal refuse pile. Science 170, 1416–1418.
Demirjian, D.C., Moris-Varas, F., Cassidy, C.S., 2001. Enzymes from extremophiles. Curr. Opin. Chem. Biol. 5, 144–151.
Dill, K.A., 1990. Dominant forces in protein folding. Biochemistry 29, 7133–7155.
Eijsink, V.G.H., Bjørk, A., Gåseidnes, S., Sirevåg, R., Synstad, B., Burg, B., Vriend, G., 2004. Rational engineering of enzyme stability. J. Biotechnol. 113, 105–120.
Eijsink, V.G.H., Gaseidnes, S., Borchert, T.V., van den Burg, B., 2005. Directed evolution of enzyme stability. Biomol. Eng. 22, 21–30.
Fulle, S., Gohlke, H., 2008. Analysing the flexibility of RNA structures by constraint counting. Biophys. J. 94, 4202–4219.
Fulle, S., Gohlke, H., 2009a. Constraint counting on RNA structures: linking flexibility and function. Methods 49, 181–188.
Fulle, S., Gohlke, H., 2009b. Statics of the ribosomal exit tunnel: implications for cotranslational peptide folding, elongation regulation, and antibiotics binding. J. Mol. Biol. 387, 502–517.
Gohlke, H., Kuhn, L.A., Case, D.A., 2004. Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. Proteins 56, 322–337.
Gohlke, H., Thorpe, M.F., 2006. A natural coarse graining for simulating large biomolecular motion. Biophys. J. 91, 2115–2120.
Greene, L.H., Higman, V.A., 2003. Uncovering network systems within protein structures. J. Mol. Biol. 334, 781–791.
Gromiha, M.M., Oobatake, M., Sarai, A., 1999. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys. Chem. 82, 51–67.
Haki, G.D., Rakshit, S.K., 2003. Developments in industrially important thermostable enzymes: a review. Bioresour. Technol. 89, 17–34.
Halabi, N., Rivoire, O., Leibler, S., Ranganathan, R., 2009. Protein sectors: evolutionary units of three-dimensional structure. Cell 138, 774–786.
Hespenheide, B.M., Jacobs, D.J., Thorpe, M.F., 2004. Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. J. Phys.: Condens. Matter 16, S5055–S5064.
Hespenheide, B.M., Rader, A.J., Thorpe, M.F., Kuhn, L.A., 2002. Identifying protein folding cores from the evolution of flexible regions during unfolding. J. Mol. Graph. Model. 21, 195–207.
Hooft, R.W.W., Vriend, G., Sander, C., Abola, E.E., 1996. Errors in protein structures. Nature 381, 272.
Jacobs, D.J., 1998. Generic rigidity in three-dimensional bond-bending networks. J. Phys. A: Math. Gen. 31, 6653–6668.
Jacobs, D.J., Dallakyan, S., 2005. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. Biophys. J. 88, 903–915.
Jacobs, D.J., Dallakyan, S., Wood, G.G., Heckathorne, A., 2003. Network rigidity at finite temperature: relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems. Phys. Rev. E: Stat. Nonlin. Soft Matter Phys. 68, 061109.
Jacobs, D.J., Hendrickson, B., 1997. An algorithm for two-dimensional rigidity percolation: the pebble game. J. Comput. Phys. 137, 346–365.
Jacobs, D.J., Rader, A.J., Kuhn, L.A., Thorpe, M.F., 2001. Protein flexibility predictions using graph theory. Proteins 44, 150–165.
Jacobs, D.J., Thorpe, M.F., 1995. Generic rigidity percolation: the pebble game. Phys. Rev. Lett. 75, 4051–4054.
Jaenicke, R., Böhm, G., 1998. The stability of proteins in extreme environments. Curr. Opin. Struct. Biol. 8, 738–748.
Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L., 1983. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 79, 926–935.
Kannan, N., Vishveshwara, S., 2000. Aromatic clusters: a determinant of thermal stability of thermophilic proteins. Protein Eng. 13, 753–761.
Klein, D.L., Radestock, S., Gohlke, H., 2011. Analyzing protein rigidity for understanding and improving thermal adaptation. In: Sen, S., Nilsson, L. (Eds.), Thermophilic Proteins: Structural Stability and Design Strategies. CRC Press Taylor & Francis Group, Boca Raton, pp. 47–67.
Larson, S.B., Day, J.S., Nguyen, C., Cudney, R., McPherson, A., 2009. Structure of pig heart citrate synthase at 1.78 Å resolution. Acta Crystallogr. Sect. F: Struct. Biol. Cryst. Commun. 65, 430–434.
Lehmann, M., Wyss, M., 2001. Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. Curr. Opin. Biotechnol. 12, 371–375.
Leisola, M., Turunen, O., 2007. Protein engineering: opportunities and challenges. Appl. Microbiol. Biotechnol. 75, 1225–1232.
Lockless, S.W., Ranganathan, R., 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286, 295–299.
Mamonova, T., Hespenheide, B., Straub, R., Thorpe, M.F., Kurnikova, M., 2005. Protein flexibility using constraints from molecular dynamics simulations. Phys. Biol. 2, S137–S147.

Oshima, T., Imahori, K., 1974. Description of *Thermus thermophilus* (Yoshida and Oshima) comb. nov., a nonsporulating thermophilic bacterium from a Japanese thermal spa. Int. J. Syst. Bacteriol. 24, 102–112.
Polizzi, K.M., Bommarius, A.S., Broering, J.M., Chaparro-Riggers, J.F., 2007. Stability of biocatalysts. Curr. Opin. Chem. Biol. 11, 220–225.
Privalov, P.L., Gill, S.J., 1988. Stability of protein structure and hydrophobic interaction. Adv. Protein Chem. 39, 191–234.
Puchkaev, A.V., Koo, L.S., Ortiz de Montellano, P.R., 2003. Aromatic stacking as a determinant of the thermal stability of CYP119 from *Sulfolobus solfataricus*. Arch. Biochem. Biophys. 409, 52–58.
Rader, A.J., Bahar, I., 2004. Folding core predictions from network models of proteins. Polymer 45, 659–668.
Rader, A.J., Hespenheide, B.M., Kuhn, L.A., Thorpe, M.F., 2002. Protein unfolding: rigidity lost. Proc. Natl. Acad. Sci. U. S. A. 99, 3540–3545.
Radestock, S., 2010. Entwicklung eines rechnerischen Verfahrens zur Simulation der thermischen Entfaltung von Proteinen und zur Untersuchung ihrer Thermostabilitaet. Goethe University, Frankfurt am Main.
Radestock, S., Gohlke, H., 2008. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. Eng. Life Sci. 8, 507–522.
Radestock, S., Gohlke, H., 2011. Protein rigidity and thermophilic adaptation. Proteins 79, 1089–1108.
Rathi, P.C., Pfleger, C., Fulle, S., Klein, D.L., Gohlke, H., 2011. Statics of biomacromolecules. In: Comba, P. (Ed.), Molecular Modeling. Wiley-VCH, Weinheim, pp. 281–299.
Reetz, M.T., Carballeira, J.D., Vogel, A., 2006. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. Angew. Chem. Int. Ed. Engl. 45, 7745–7751.
Remington, S., Wiegand, G., Huber, R., 1982. Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 Å resolution. J. Mol. Biol. 158, 111–152.
Ryckaert, J.P., Ciccotti, G., Berendsen, H.J.C., 1977. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J. Comput. Phys. 23, 327–341.
Sali, A., Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779–815.
Schellman, J.A., 1997. Temperature, stability, and the hydrophobic interaction. Biophys. J. 73, 2960–2964.
Simmerling, C., Strockbine, B., Roitberg, A.E., 2002. All-atom structure prediction and folding simulations of a stable protein. J. Am. Chem. Soc. 124, 11258–11259.
Somero, G.N., 1978. Temperature adaptation of enzymes: biological optimization through structure–function compromises. Annu. Rev. Ecol. Syst. 9, 1–29.
Uchiyama, T., Miyazaki, K., 2009. Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr. Opin. Biotechnol. 20, 616–622.
Van den Burg, B., 2003. Extremophiles as a source for novel enzymes. Curr. Opin. Microbiol. 6, 213–218.
Vieille, C., Zeikus, G.J., 2001. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol. Mol. Biol. Rev. 65, 1–43.
Volkl, P., Huber, R., Drobner, E., Rachel, R., Burggraf, S., Trincone, A., Stetter, K.O., 1993. *Pyrobaculum aerophilum* sp. nov., a novel nitrate-reducing hyperthermophilic archaeum. Appl. Environ. Microbiol. 59, 2918–2926.
Wiegand, G., Remington, S.J., 1986. Citrate synthase: structure, control, and mechanism. Annu. Rev. Biophys. Biophys. Chem. 15, 97–117.
Word, J.M., Lovell, S.C., Richardson, J.S., Richardson, D.C., 1999. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J. Mol. Biol. 285, 1735–1747.
Zillig, W., Stetter, K.O., Wunderl, S., Schulz, W., Priess, H., Scholz, I., 1980. The Sulfolobus-Caldariella group: taxonomy on the basis of the structure of DNA-dependent RNA polymerases. Arch. Microbiol. 125, 259–269.

Publication III

Constraint Network Analysis (CNA): A Python software package for efficiently linking biomacromolecular structure, flexibility, (thermo)stability, and function

Christopher Pfleger[§], Prakash Chandra Rathi[§], Doris L. Klein, Sebastian Radestock, and Holger Gohlke

Journal of Chemical Information and Modeling, **2013**, 53:1007-1015.
2012 impact factor: 4.20; contribution: 35%

[§] Both authors share first authorship

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

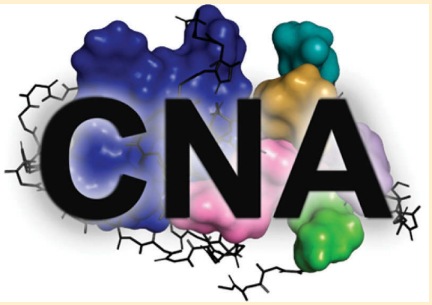
pubs.acs.org/jcim

Constraint Network Analysis (CNA): A Python Software Package for Efficiently Linking Biomacromolecular Structure, Flexibility, (Thermo-)Stability, and Function

Christopher Pfleger,[‡] Prakash Chandra Rathi,[‡] Doris L. Klein, Sebastian Radestock,[†] and Holger Gohlke*

Institute for Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich-Heine-University, Universitätsstr. 1, 40225, Düsseldorf, Germany

ABSTRACT: For deriving maximal advantage from information on biomacromolecular flexibility and rigidity, results from rigidity analyses must be linked to biologically relevant characteristics of a structure. Here, we describe the Python-based software package Constraint Network Analysis (CNA) developed for this task. CNA functions as a front- and backend to the graph-based rigidity analysis software FIRST. CNA goes beyond the mere identification of flexible and rigid regions in a biomacromolecule in that it (I) provides a refined modeling of thermal unfolding simulations that also considers the temperature-dependence of hydrophobic tethers, (II) allows performing rigidity analyses on ensembles of network topologies, either generated from structural ensembles or by using the concept of fuzzy noncovalent constraints, and (III) computes a set of global and local indices for quantifying biomacromolecular stability. This leads to more robust results from rigidity analyses and extends the application domain of rigidity analyses in that phase transition points ("melting points") and unfolding nuclei ("structural weak spots") are determined automatically. Furthermore, CNA robustly handles small-molecule ligands in general. Such advancements are important for applying rigidity analysis to data-driven protein engineering and for estimating the influence of ligand molecules on biomacromolecular stability. CNA maintains the efficiency of FIRST such that the analysis of a single protein structure takes a few seconds for systems of several hundred residues on a single core. These features make CNA an interesting tool for linking biomacromolecular structure, flexibility, (thermo-)stability, and function. CNA is available from http://cpclab.uni-duesseldorf.de/software for nonprofit organizations.



INTRODUCTION

The concepts of biomacromolecular flexibility and its opposite, rigidity, are crucial for understanding the relationship between biomacromolecular structure, (thermo-)stability, and function. In the field of statics, flexibility and rigidity denote the possibility (or impossibility) of internal motion but are not associated with information about directions and magnitudes of movements. Identifying and modulating the heterogeneous composition of biomacromolecules in terms of flexible and rigid regions is becoming increasingly important for successful protein engineering and rational drug-design.[1−5] Several computational approaches have been developed that identify flexible and rigid regions by either determining spatial variations in the local packing density[6] or representing and analyzing a structure as a connectivity network of interacting atoms or residues.[7−12] The approaches benefit from being computationally highly efficient. A related concept has been introduced by Jacobs et al.[13] Here, biomacromolecules were initially represented as bond-bending networks in which each atom has three degrees of freedom representing the dimensions of motion in 3-space. In later versions, the equivalent body-bar representation is used where atoms are modeled as bodies with six degrees of freedom.[13−15] By adding constraints (representing covalent and noncovalent bonds in a biomacromolecular context) between the bodies, internal motions become restricted. Each constraint is modeled as a set of bars, and each bar removes one degree of freedom. According to the type of interaction, the number of bars varies in that stronger interactions are modeled with a higher number of bars than weaker ones. Noncovalent interactions such as hydrogen bonds, salt bridges, hydrophobic tethers, and stacking interactions contribute most to the biomacromolecular stability; hence, these interactions are modeled as constraints in addition to covalent bonds. Once the network is constructed, the Pebble Game algorithm, available within the FIRST (Floppy Inclusions and Rigid Substructure Topography) software, efficiently decomposes the network into rigid clusters and flexible hinge regions from the number and spatial distribution of bond-rotational degrees of freedom.[16,17] A rigid region is a collection of interlocked bonds allowing no relative motion of the bodies. Such a region can either be overconstrained, if it has redundant constraints, or is isostatically rigid. In a flexible region, dihedral rotation is not locked in by other bonds. The theory underlying this approach is rigorous[18] and has been applied in different areas of biomacromolecular research.[5,19−35]

Received: January 20, 2013
Published: March 21, 2013

ACS Publications © 2013 American Chemical Society

We developed the command-line Python-based software package Constraint Network Analysis (CNA) for analyzing structural features of biomacromolecules that are important for the molecule's stability. CNA functions as a front- and backend to the FIRST software and allows (I) setting up a variety of constraint network representations for analysis by FIRST, (II) processing the results obtained from FIRST, and (III) calculating seven indices for quantifying biomacromolecular stability, both globally and locally.[36] As to the latter, the indices are calculated by monitoring changes of the network stability along a thermal unfolding simulation. The thermal unfolding is simulated by consecutively removing hydrogen bond (including salt bridge) constraints from the network with increasing temperature. Thermal unfolding simulations have been successfully applied in several studies on proteins, RNAs, and the ribosome in order to understand how flexibility and rigidity is linked to biomacromolecular stability and function.[4,5,14,19,28,31,34,35,37]

CNA goes beyond the mere identification of flexible and rigid regions in a biomacromolecular structure in that it allows linking results from constraint network analysis to biologically relevant characteristics of a structure. This is key for deriving maximal advantage from information on biomacromolecular flexibility and rigidity. Here, we describe the design and implementation of the CNA software package. We then demonstrate its application scope in a showcase example on Hen Egg White Lysozyme (HEWL) structures. The CNA software package is available under an academic license from http://cpclab.uni-duesseldorf.de/software.

METHODS AND IMPLEMENTATION

General Overview. The CNA software package allows three different types of rigidity analysis: (I) based on a *single network topology* generated from a single input structure, (II) based on an *ensemble of network topologies* generated from a conformational ensemble provided as input,[21,35] and (III) based on an *ensemble of network topologies* generated from a single input structure by considering fuzzy noncovalent constraints (FNC) (C. Pfleger, H. Gohlke, to be published elsewhere). The last variant mimics that noncovalent constraints thermally break and reform even in the native state of a biomacromolecule.[38] In short, we developed a system-independent parametrization of fuzzy noncovalent constraints by analyzing the atom type and location-dependent persistence characteristics of noncovalent constraints (hydrogen bonds, salt-bridges, and hydrophobic tethers) during MD simulations. With this, the number and distribution of noncovalent constraints are modulated by random components within certain ranges, simulating thermal fluctuations of a biomacromolecule without actually moving atoms. In the related distance constraint model (DCM), ensembles of network topologies are generated considering mean-field probabilities of hydrogen bond and torsion constraints in a Monte Carlo sampling.[20,39] Average stability characteristics are then calculated by constraint counting on each topology in the ensemble.[40] As a downside, the DCM approach requires experimental data for a system-specific parametrization of the model.

The analysis of a *single network topology* by CNA consists of the following steps. Initially, a constraint network is generated from the input structure by placing covalent and noncovalent constraints according to rules described in refs 13−15. Next, a thermal unfolding simulation is carried out by sequentially removing noncovalent constraints from the network (see section Thermal Unfolding Simulation for details). For each network during the simulation, a rigidity analysis by FIRST is performed and then post-processed to calculate global and local indices to characterize biomacromolecular flexibility and rigidity. The workflow of the software is illustrated in Figure 1. In the case of analyzing an *ensemble of network topologies*, these steps are repeated for each network, and the results are averaged over the ensemble.

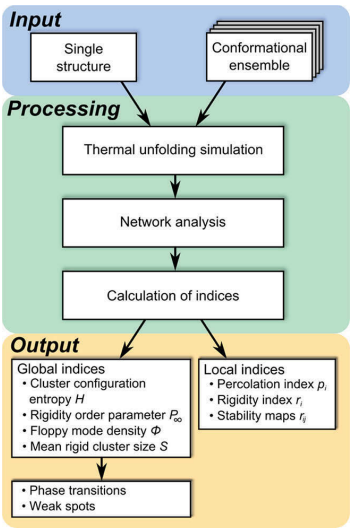


Figure 1. Schematic workflow of the CNA software.

Upon running a thermal unfolding simulation (a) phase transition(s) can be identified at which the network changes from mainly rigid to flexible. For this, the change in the global indices is monitored during the simulation. Four different global indices are implemented in CNA. They monitor (I) the normalized number of independent internal degrees of freedom (floppy mode density, Φ), (II) the fraction of the network belonging to a rigid component (rigidity order parameter, P_∞), (III) the degree of disorder in the network (cluster configuration entropy, H), and the rigid cluster size distribution (mean rigid cluster size, S). In addition, CNA calculates three local indices that characterize the flexibility and rigidity at the bond level: (I) the percolation index p_i monitors the percolation behavior of a biomacromolecule on a microscopic level and thus allows the identification of the hierarchical organization of the giant percolating cluster during a thermal unfolding simulation, (II) the rigidity index r_i monitors when a bond segregates from a rigid cluster, (III) a stability map is a two-dimensional itemization of the rigidity index r_i and is derived by identifying "rigid contacts" between two residues. Exact definitions of these indices and guidelines for when to use them are given in ref 36. Furthermore, the CNA software identifies unfolding nuclei, i.e., those residues that break apart from the giant cluster at the phase transition point.[4,28,35] The unfolding nuclei can be considered weak spots in the structure; accordingly, this knowledge can be exploited in data-driven protein engineering to focus on residues that are highly likely to improve thermostability upon mutation.

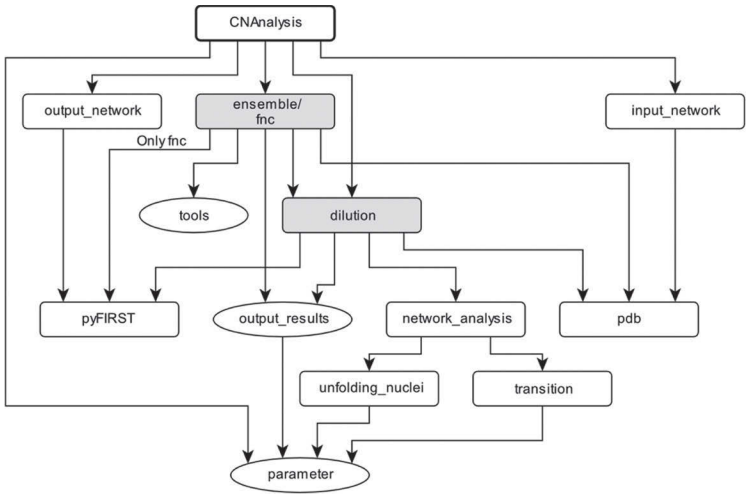


Figure 2. Hierarchical structure of the CNA software. All modules that contain (a) class definition(s) are shown in rectangles. The core module *CNAnalysis* is highlighted by a bold frame. Modules colored in gray contain the simulation methods for analyzing a single network topology and an ensemble of network topologies. Modules that solely contain methods are shown as ellipses. An arrow indicates the call of a module by another module.

CNA is implemented as a Python-based software package making use of an object-oriented design (Figure 2). Third party software is required for full functionality (Table 1): (I) The Biopython package[41] is needed to parse input PDB files and provides information on secondary structure from a DSSP analysis.[42,43] (II) For statistical analysis and detecting the phase transitions, the Numpy[44] and SciPy[45] extensions for Python are required. (III) The Open Babel[46,47] Python-bindings are required to determine the bond order of small-molecule ligands. To facilitate the installation of the CNA software package, the third party software is provided with the CNA source tree except for the DSSP program, which is available at http://swift.cmbi.ru.nl/gv/dssp/. The CNA source tree also contains a comprehensive documentation detailing the installation and usage of the software and a suite of test cases to check the validity of the installation. CNA is a command-line based software that is called by the shell script *CNA.sh*. A "--help" argument lists all available options and required arguments, their descriptions, default values, and the range of allowed values. An erroneous argument set for an option produces an informative error message. The CNA software has been successfully tested on Debian, OpenSuse, and CentOS Linux platforms.

Table 1. External Software Needed by the CNA Software

name	version	description and use
Python	2.73	Python interpreter used by the CNA software package
Biopython	1.58	for reading PDB files using the Bio.PDB package and parsing results from the DSSP program using the Bio.DSSP package
NumPy	1.6.1	for statistical analyses
SciPy	0.11.0	for statistical analyses
Open Babel	2.3.1	for identifying the connectivity and bond orders of ligand molecules
DSSP		for computing secondary structure information that is required by the FNC approach
SWIG	2.0.8	for compiling the *pyFIRST* interface module

Constraint Network Analysis Is the Core Module. The *CNAnalysis* module is the core of the CNA software. The *CNAnalysis* module consists of a single class ConstraintNetworkAnalysis. Upon creating an instance of the type ConstraintNetworkAnalysis, it (I) parses the command line options that specify the analysis type, (II) checks whether the values of the command line arguments conform to the desired data-type, and (III) performs the requested analysis. Depending on the type of analysis, the ConstraintNetworkAnalysis instance creates an instance of the class Dilution if the analysis of a *single network topology* is requested. Otherwise, it creates an instance of the class Fnc or Ensemble, which then creates instances of the class Dilution for each network of the ensemble. The command line options provided by the user are checked for validity by the module *Parameter*; this module also contains default values for the options and internal constants.

***PyFIRST* as an Interface.** We developed the *pyFIRST* interface module to directly access the functionality of the FIRST software (available at http://flexweb.asu.edu) within the Python environment of CNA. The interface module was implemented using the SWIG (Simplified Wrapper and Interface Generator) software tool (http://www.swig.org/).[48] SWIG automatically generates a wrapper code for C/C++ programs that then acts as an interface for other high level programming languages such as Python. The SWIG interface file is written in C++ and contains a single class pyFIRST. The class contains methods that are later on accessible within the Python environment of CNA. Upon instantiating a pyFIRST object, a data structure is generated that represents the constraint network topology of the input structure. Additionally, the pyFIRST object provides methods that are used to (I) read constraint information (covalent bonds, hydrogen bonds, salt bridges, hydrophobic tethers, and stacking interactions) from the network topology, (II) remove constraints from the network with respect to all or a certain type of constraints, and

(III) perform a rigid cluster decomposition. Finally, methods are available that return warnings issued by FIRST when initializing the data structure for the constraint network topology. Note that the *pyFIRST* interface module has been written such that it can be used in any other Python-based application requiring a rigidity analysis by FIRST, thus providing a general Python interface to FIRST.

Structural Information as Input. Single or multiple (in case of a conformational ensemble) input structures for CNA must be in PDB format.[49] Although the validity of the input structure(s) is checked upon creating an instance of the class PDB, we recommend subjecting only complete structures without missing residues or atoms. Hydrogen atoms must be present, too, because otherwise the identification of hydrogen bond and salt bridge constraints cannot be performed. Ligand molecules, if present, are extracted from the input structure and, subsequently, analyzed to determine the bond order by means of Open Babel.[46,47] The last step requires the presence of hydrogen atoms at the ligand. All identified rotatable bonds (single bonds) are then modeled by five bars, whereas nonrotatable bonds (double, triple, amide, and aromatic bonds) are modeled by six bars.[15] Finally, the covalent constraint information for the ligand is merged with the covalent and noncovalent constraint network of the biomacromolecule also generating noncovalent constraints between them. Ions, water, and buffer molecules are handled by FIRST. If an NMR structure is used as input, only the first model is considered. Furthermore, Amber-conform residue names (HIE, HID, HIP, and CYX) are replaced by standard residue names (HIS and CYS) in order to allow the use of PDB structures extracted from molecular dynamics (MD) trajectories created by the Amber software.[50] In the case of a conformational ensemble, a PDB object is instantiated for each conformation. Apart from checking the validity of and preparing the input structure, the PDB class provides several functions that can be used to work with the structure in terms of getting single atom and residue objects, finding neighbor residues within a certain distance cutoff, and writing out structures (including biomacromolecules and ligand molecules) in the PDB format.

Accessing the Network Topology. The *output_network* and *input_network* modules of CNA contain the OutputNetwork and InputNetwork class definitions. Upon instantiating an object, these classes are used to write and read the constraint network topology of a single structure or of each conformation of an ensemble. This is particularly useful for adding user-defined constraints that are not identified automatically, for example, constraints between ions and protein atoms. In the file containing the constraint network topology, each entry of a covalent constraint contains the identifiers of the involved atoms and number of bars of the constraint. For constraints representing hydrophobic or stacking interactions, in addition to the atom identifiers, the distance between the atoms is given plus an indicator whether the constraint occurs within a protein or between protein and ligand. For hydrogen bond and salt bridge constraints, the energy and type of interaction is written instead of the distance and indicator. This file can be modified and used as input for CNA again. In this case, user-defined constraints will overwrite constraint information identified from the input structure(s).

Thermal Unfolding Simulation. The thermal unfolding simulation allows analyzing changes in the network stability upon removing hydrogen bond (including salt bridge) constraints from the network.[4,14,28] To do so, the energy of a hydrogen bond E_{HB} is determined by an empirical energy function.[51] Then, during the thermal unfolding simulation,[4,28] intermediate networks σ are created such that hydrogen bonds with an energy $E_{HB} > E_{cut}(\sigma)$ are removed from the network.[51] This follows the idea that stronger hydrogen bonds will break at higher temperatures than weaker ones. By means of an empirically determined linear function, E_{cut} can be related to a temperature T.[28]

Consequently, the simulation mimics a rise in the temperature by analyzing a range of networks having many hydrogen bonds (equivalent to low temperatures) to having few hydrogen bonds (equivalent to high temperatures). Note that the temperatures should be considered relative values only because the absolute values may depend on the size and architecture of the analyzed protein.[4] Still, the temperatures are very helpful, for example, when it comes to comparing the thermostability of two or more homologous proteins or the stability of a wild-type with its mutant.[4,28,35] An alternative concept grounded in mean-field theory directly connects network rigidity and absolute temperature; while appealing, it requires experimental data for a system-specific parametrization.[20,40] Each of the intermediate networks σ is then subjected to rigidity analysis by FIRST. While the principal idea of the thermal unfolding simulation has been adapted from the FIRST software,[13] the method implemented here allows for additional settings that are not available in the FIRST implementation. These include specifying the energy range and step-size for removing hydrogen bonds. Furthermore, a modified method has been implemented that also considers the temperature dependence of hydrophobic tethers along the thermal unfolding simulation.[35] This approach follows the idea that hydrophobic interactions become stronger with increasing T.[52,53] Accordingly, more hydrophobic tethers are added to the network by linearly increasing the distance cutoff for including hydrophobic tethers $D_{cut}(\sigma)$ from a starting value of 0.25 Å at 300 K to an ending value of 0.40 Å at 420 K. Doing this has been shown to improve thermostability predictions of citrate synthases.[35]

The thermal unfolding simulation is done by the *dilution* module containing the Dilution class. Upon instantiating an object of this class, the object creates new intermediate networks σ and passes the networks through FIRST by instantiating a pyFirst object. Subsequently the module *networkAnalysis* is used to calculate the global and local indices (see section Analyzing the Results from the Rigidity Analysis). Via the global indices, phase transition(s) are identified by an object of the class Transitions. Finally, unfolding nuclei are identified by an object of the class UnfoldingNuclei.

Analyzing the Results from the Rigidity Analysis. The *network_analysis* module comprises in total four classes that process the results from the FIRST rigidity analysis. The main class NetworkAnalysis contains methods to calculate the size and size distribution of rigid clusters and to identify the actual largest rigid cluster as well as the giant percolating cluster of the network. The giant percolating cluster is the largest rigid cluster present at the highest E_{cut} value (i.e., at the lowest temperature) with all constraints in place. During the thermal unfolding simulation, the melting of the giant percolating cluster is monitored, and the largest rigid subcluster of the previous giant percolating cluster becomes the new giant percolating cluster of the present network state σ. Subsequently, the NetworkAnalysis object is passed to three classes for calculating the global and

local indices called GlobalIndices, LocalIndices, and LocalStabilityMaps.

The class GlobalIndices contains all methods that are required to calculate the floppy mode density Φ, the rigidity order parameter P_∞, the cluster configuration entropy H, and the mean rigid cluster size S.[36] Apart from this, the class GlobalIndices also instantiates objects of the classes Transitions and UnfoldingNuclei that are required for the identification of phase transition points and unfolding nuclei of the structure. For identifying phase transition points, two methods have been implemented that make use of the data of the global indices: fitting of a mono/double sigmoid curve and interpolating with a smoothed spline. By default, phase transition points are identified by the double sigmoid curve.[35] However, the user can choose as an option that Akaike's information criterion[54] be used to identify whether a mono or double sigmoid curve gives better fitting results. Finally, if more than two phase transitions are expected or shall be identified, interpolation with the smoothed spline is recommended. Multiple transitions can occur in multimeric proteins. The transition point is then identified for each global index as the point at which the maximal rigidity loss occurs in the structure. Occasionally, a Transitions object does not return a transition point; this occurs if no "sharp" transition can be detected or if multiple transitions with comparable rigidity losses are present.

The class LocalIndices is used to calculate the percolation index p_i and the rigidity index r_i. Both reflect structural stability on a per-residue basis[36] and, thus, can be used to identify the location and distribution of structurally weak or strong parts in biomacromolecules. Finally, the class LocalStabilityMaps is used to calculate the two-dimensional itemization of the rigidity index r_i, the stability map, and a so-called "neighbor stability map", where values of the stability map of residue pairs separated by more than 5 Å are masked. That way, the latter map provides useful information about the stability of neighboring residues only, which can be used for focusing on short-range weak and strong connections within a biomacromolecule.

Writing the Analysis Results. The module *output_results* is used to write results files containing information about global and local indices, phase transition points, and unfolding nuclei. For a phase transition point, the hydrogen bond energy cutoff E_{cut} and the respective temperature are listed. Unfolding nuclei are written out as a text file and PDB file; in the latter, the B-factor column is used to record whether or not a residue is an unfolding nucleus by setting the values to one or zero. If the analysis is performed on an *ensemble of network topologies*, an additional file summarizing the average local indices and standard deviations is written. Similarly, for the phase transition points, mean, median, and standard error are provided in addition. Furthermore, the percentage of network topologies in which a residue is predicted to be an unfolding nucleus is recorded.

Showcase Example: Flexibility Characteristics of HEWL. In a showcase example, we applied the CNA software to a HEWL structure. We show the results for two analysis types, analyzing a single network topology derived from a single input structure (PDB ID: 3LZT) and analyzing an ensemble of network topologies derived from a conformational ensemble. The conformational ensemble was generated by extracting 1500 conformations from a trajectory of 300 ns length obtained by MD simulations starting from an X-ray structure of HEWL (PDB ID: 3LZT). The MD simulation was carried out in explicit solvent at 300 K with the AMBER 11 package of molecular simulation programs.[50] The detailed simulation protocol is described elsewhere (C. Pfleger, H. Gohlke, to be published elsewhere). Water molecules were removed from each conformation before the ensemble was subjected to CNA. Analyzing a single network topology took about 40 s, and the ensemble of 1500 conformations required ~11 h on a single-core workstation computer, which demonstrates the computational efficiency of CNA and FIRST.

Snapshots from the thermal unfolding simulation of the single input structure are depicted in Figure 3. They show the

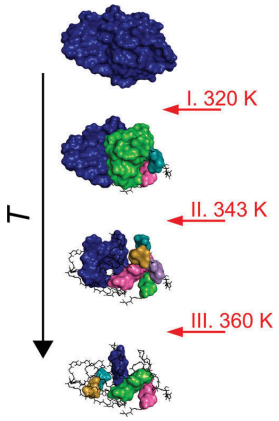


Figure 3. Rigid cluster decompositions along the thermal unfolding simulation of the showcase example HEWL. Rigid clusters are shown as uniformly colored bodies connected by flexible hinge regions (black). The roman numbers relate to three major steps of rigidity loss.

loss of rigidity in terms of the decay of rigid clusters with increasing temperature. The first transition relates to the beginning of the collapse of the giant rigid cluster, which occurs in the interface region of the α- and β-domains. At this state, the network is dominated by two large rigid components. During the next transition, the rigid cluster covering the α-domain collapses, and the helical elements remain as single rigid clusters. Finally, during the last transition, the rigid cluster covering the β-domain collapses, and nearly the whole system becomes flexible. The results from the thermal unfolding simulation agree, in reverse order, with the "fast track" folding pathway described in refs 55 and 56. Here, both domains of HEWL fold concurrently but with a slight preference to initially form native contacts in the β-domain.[57] Alternatively, a "slow track" folding reaction of HEWL has been described,[56,58,59] in which the majority of the protein molecules populate an intermediate state with persistent structures in only the α-domain.[57] Still, parts of the α-domain need to unfold again to enable the subsequent folding of the β-domain.

As an example for a global index, the cluster configuration entropy H is shown, which monitors the loss of network stability during the thermal unfolding simulation. In the analysis of the single network topology (Figure 4a), an early phase transition at 319 K indicates the beginning decay of structural stability, with most of the network still being captured in rigid clusters. The dominant phase transition at 343 K then refers to the point at which the network loses its ability to carry stress

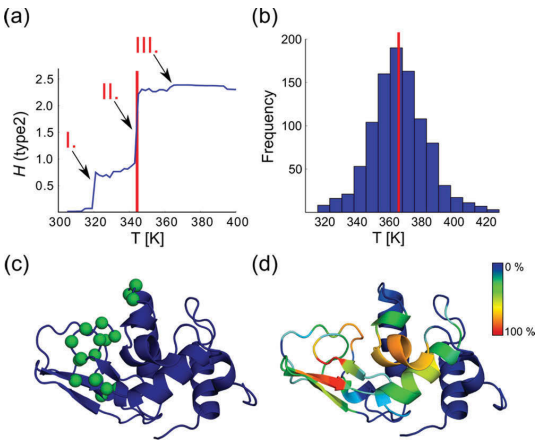


Figure 4. (a) Cluster configuration entropy H (type 2) derived from the single network topology. The entropy is plotted as a function of the temperature, and the roman numbers correspond to the three major steps depicted in Figure 3. The phase transition automatically identified by CNA is marked by the red vertical line. (b) Frequency distribution of phase transitions identified from analyzing the ensemble of network topologies. The median is marked with a red vertical line. (c) Weak spot detection for the single network topology. Green spheres highlight the identified weak spot residues in the HEWL structure. (d) Weak spot detection over the ensemble of network topologies. For depicting the probability of being a weak spot, each residue is colored according to a color scale ranging from blue (low probability) to red (high probability).

and, hence, corresponds to the folded−unfolded transition. The last transition indicates the loss of the remaining rigid components. In the case of analyzing the ensemble of network topologies, the frequency distribution of the identified phase transition points is shown (Figure 4b). From this, a median transition temperature of 358 K is revealed, which is 15 K higher than the dominant phase transition point identified from analyzing a single network topology. Note that, in general, phase transitions identified using a single input structure can be different from ensemble results, as shown in a previous study on citrate synthase.[35] We thus recommend performing CNA analyses on ensembles of network topologies, in particular, when quantitative results are desired. At the transition point, unfolding nuclei are identified (Figure 4c). Almost all unfolding nuclei are located in the β-domain of HEWL, which disintegrates at the dominant phase transition (Figures 3 and 4c). Furthermore, for the ensemble of network topologies, the probability of a residue being found as an unfolding nucleus over the entire ensemble is provided (Figure 4d). The higher this probability the more likely will it be that rigidifying this residue will improve protein stability. The ensemble results are more detailed than the ones from the single structure in that now unfolding nuclei are not only located in the β-domain but also in helix B, which agrees with the view that this helix plays a crucial role in stabilizing the tertiary structure of HEWL.[60]

As for local indices, we exemplary show the rigidity index r_i, which characterizes the stability of the HEWL structure down to the bond level (Figure 5a, b). As such, r_i monitors the point when a residue segregates from a rigid cluster along the thermal unfolding simulation: the lower r_i the longer is a residue part of a rigid cluster. Secondary structure elements are generally found to be more stable than loop regions. Furthermore, averaging r_i values over the ensemble of network topologies leads to a smoother r_i curve and to the spike located at residue 78 becoming less pronounced than in the case of analyzing the single network topology. The spike reveals a region that is highly stabilized by hydrophobic interactions; these regions only melt at a late stage of the thermal unfolding simulations. Notably, the stable regions identified for residues 53 and 62−65 are in very good agreement with those identified by high protection factors in H/D experiments for the native and denatured states of HEWL.[61] During the catalytic cycle, HEWL undergoes a reorientation of the α- and β-domains due to a bending movement around a central hinge region.[62] Along these lines, the identified flexible hinge regions (Figure 5a, b) are in agreement with those suggested by McCammon et al.[62] and coincide with results obtained from Gaussian network models and MD simulations.[60,63] Such a decomposition into rigid clusters and flexible regions is used as a first step in a normal mode-based geometric simulation approach (NMSim) working on a coarse-grained protein representation.[64] With this, stereochemically and energetically favorable conformations of HEWL were generated previously.[64]

As yet another local index, stability maps rc_{ij} are two-dimensional itemizations of the r_i and report when a "rigid contact" between two residues of the network vanishes during the thermal unfolding simulation. The upper triangles of Figure 5c and d show the stability maps for the single network topology and the ensemble of network topologies, respectively. Again, blocks of stable contacts are pronounced for secondary structures elements. In contrast, very weak contacts are identified for residues 81−87 that partially form a 3_{10} helix. This is in agreement with results from NMR experiments that reveal a disordered structure of this region.[65] The lower triangles of Figure 5c and d show a modification of the stability map that highlights solely those residue pairs with a "rigid contact" where the residues are within a distance of 5 Å. This map is referred to as "neighbor stability map". Accordingly, a rigid contact in such a map that melts early in the thermal unfolding simulation is a prominent target for rigidification and, hence, for improving protein stability.

■ CONCLUSIONS

In recent years, there has been encouraging progress in characterizing the flexibility and rigidity of biomacromolecules down to the residue level by graph theoretical approaches. However, for deriving maximal advantage from information on biomacromolecular flexibility and rigidity, results from rigidity analyses must be linked to biologically relevant characteristics of a structure, such as (thermo-)stability and function. This provided the incentive for us to develop the CNA software package presented here. CNA functions as a front- and backend to the FIRST software and allows setting up a variety of constraint network representations, processing the results obtained from FIRST, and calculating global and local indices for quantifying biomacromolecular stability.

Thus, while CNA relies on FIRST as a core engine, it goes beyond the mere identification of flexible and rigid regions in a biomacromolecular structure. Major advancements in that respect include (I) a refined modeling of thermal unfolding simulations that considers the temperature-dependence of hydrophobic tethers, (II) the ability to perform rigidity analyses on ensembles of network topologies, either generated from structural ensembles provided as input or by using the concept

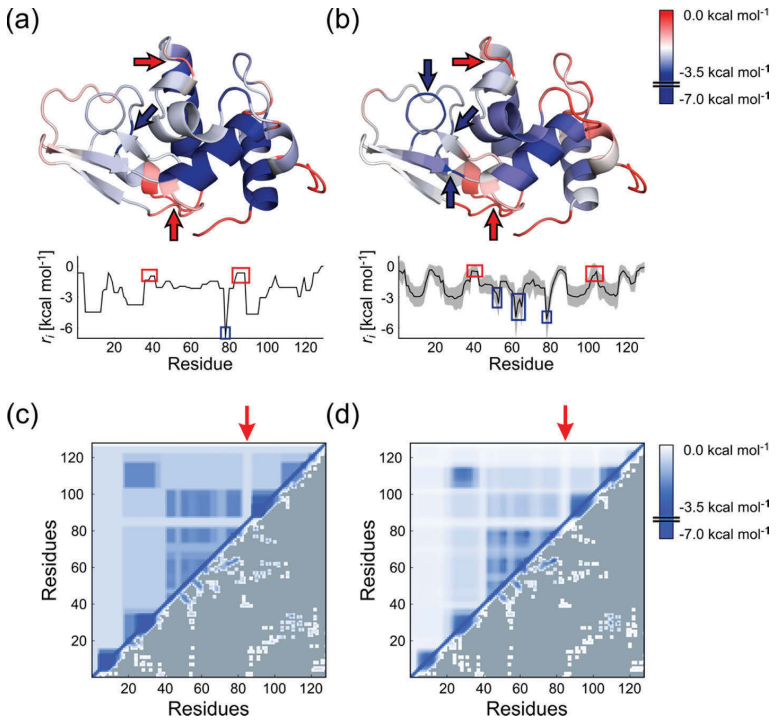


Figure 5. (a) Rigidity index r_i determined by analyzing the single network topology and (b) ensemble of network topologies plotted against a residue identifier and color coded onto the structure (range of color code: red (flexible) to blue (rigid)). In addition, the plot in (b) shows the standard deviation as a gray area. Blue rectangles and blue arrows in panels (a) and (b) highlight structurally stable regions for which high protection factors have been determined by H/D experiments. Red rectangles and red arrows in panels (a) and (b) highlight structurally flexible regions that are associated with hinge regions of HEWL. Stability maps (upper triangle) and neighbor stability maps (lower triangle) determined by analyzing the single network topology (c) and the ensemble of network topologies (d). The color depicts how stably two residues are connected and ranges from white (low stability) to blue (high stability). Red arrows highlight regions that reveal a disordered structure in NMR experiments. Gray areas in the neighbor stability map are displayed when residues are more than 5 Å away from each other.

of fuzzy noncovalent constraints, and (III) computing a set of global and local indices for characterizing biomacromolecular flexibility and rigidity, three of which have been introduced only recently by us.[36] The advancements allow (I) modeling in a more detailed manner the thermal unfolding of biomacromolecules, (II) obtaining more robust results from rigidity analyses due to a reduced sensitivity to the structural input, and (III) extending the application domain of rigidity analyses in that phase transition points ("melting points") and unfolding nuclei ("structural weak spots") are determined automatically. Such advancements are important for data-driven protein engineering, for example, for identifying structural parts that influence protein thermostability.[28] Furthermore, CNA robustly handles small-molecule ligands in general. This is important when it comes to estimating the influence of ligands on biomacromolecular stability, for example, for probing signal transmission across a protein structure for understanding and predicting "dynamic allostery"[66] and in assessing (changes in) flexibility characteristics of binding sites and interface regions.[67] How CNA can be applied in that respect has been demonstrated in a showcase example on HEWL.

CNA maintains the efficiency of FIRST. This has been achieved by linking CNA and FIRST via the *pyFIRST* interface module, minimizing the I/O overhead. The analysis of a single protein structure by CNA usually takes only a few seconds for systems of several hundred residues on a single core. The runtime for analyses of ensembles of network topologies, which is in the order of hours currently, could be further reduced given that processing individual members of such an ensemble is trivially parallelizable. Finally, the hierarchical design of the software makes CNA highly adaptable and extensible, for example, by adding new index definitions.

Overall, we believe that these unique features make CNA an interesting tool for linking biomacromolecular structure, flexibility, (thermo-)stability, and function.

AUTHOR INFORMATION

Corresponding Author

* Phone: (+49) 211-81-13662. Fax: (+49) 211-81-13847. E-mail: gohlke@uni-duesseldorf.de.

Present Address

[†]Elsevier Information Systems GmbH, Frankfurt am Main, Germany.

Author Contributions

[‡]These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to the Ministry of Innovation, Science, and Research of North Rhine-Westphalia and Heinrich-Heine-University Düsseldorf for a scholarship to PCR within the CLIB-Graduate Cluster Industrial Biotechnology. We are grateful to Daniel Mulnaes (Heinrich-Heine-University, Düsseldorf) for proofreading the manuscript. CNA is available from http://cpclab.uni-duesseldorf.de/software for nonprofit organizations.

■ ABBREVIATIONS

CNA, Constraint Network Analysis; FIRST, Floppy Inclusions and Rigid Substructure Topography; HEWL, Hen Egg White Lysozyme; FNC, Fuzzy Noncovalent Constraints; PDB, Protein Data Bank; DSSP, Define Secondary Structure of Proteins

■ REFERENCES

(1) Ahmed, A.; Kazemi, S.; Gohlke, H. Protein flexibility and mobility in structure-based drug design. *Front. Drug Des. Discovery* **2007**, *3*, 455−476.
(2) Heal, J. W.; Jimenez-Roldan, J. E.; Wells, S. A.; Freedman, R. B.; Romer, R. A. Inhibition of HIV-1 protease: The rigidity perspective. *Bioinformatics* **2012**, *28*, 350−357.
(3) Jagodzinski, F.; Hardy, J.; Streinu, I. Using rigidity analysis to probe mutation-induced structural changes in proteins. *J. Bioinf. Comput. Biol.* **2012**, 10.
(4) Radestock, S.; Gohlke, H. Protein rigidity and thermophilic adaptation. *Proteins* **2011**, *79*, 1089−1108.
(5) Tan, H. P.; Rader, A. J. Identification of putative, stable binding regions through flexibility analysis of HIV-1 gp120. *Proteins* **2009**, *74*, 881−894.
(6) Halle, B. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1274−1279.
(7) Dokholyan, N. V.; Li, L.; Ding, F.; Shakhnovich, E. I. Topological determinants of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 8637−8641.
(8) Vendruscolo, M.; Dokholyan, N. V.; Paci, E.; Karplus, M. Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E* **2002**, *65*, 1−4.
(9) Böde, C.; Kovács, I. A.; Szalay, M. S.; Palotai, R.; Korcsmáros, T.; Csermely, P. Network analysis of protein dynamics. *FEBS Lett.* **2007**, *581*, 2776−2782.
(10) Greene, L. H.; Higman, V. A. Uncovering network systems within protein structures. *J. Mol. Biol.* **2003**, *334*, 781−791.
(11) Heringa, J.; Argos, P. Side-chain clusters in protein structures and their role in protein folding. *J. Mol. Biol.* **1991**, *220*, 151−171.
(12) Heringa, J.; Argos, P.; Egmond, M. R.; Devlieg, J. Increasing thermal stability of subtilisin from mutations suggested by strongly interacting side-chain clusters. *Protein Eng.* **1995**, *8*, 21−30.
(13) Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins* **2001**, *44*, 150−165.
(14) Rader, A. J.; Hespenheide, B. M.; Kuhn, L. A.; Thorpe, M. F. Protein unfolding: Rigidity lost. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 3540−3545.
(15) Whiteley, W. Counting out to the flexibility of molecules. *Phys. Biol.* **2005**, *2*, S116−S126.
(16) Jacobs, D. J.; Thorpe, M. F. Generic rigidity percolation: The pebble game. *Phys. Rev. Lett.* **1995**, *75*, 4051−4054.
(17) Jacobs, D. J.; Hendrickson, B. An algorithm for two-dimensional rigidity percolation: The pebble game. *J. Comput. Phys.* **1997**, *137*, 346−365.
(18) Katoh, N.; Tanigawa, S. A proof of the molecular conjecture. *Discrete Comput. Geom.* **2011**, *45*, 647−700.
(19) Hespenheide, B. M.; Rader, A. J.; Thorpe, M. F.; Kuhn, L. A. Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graphics Modell.* **2002**, *21*, 195−207.
(20) Jacobs, D. J.; Dallakyan, S.; Wood, G. G.; Heckathorne, A. Network rigidity at finite temperature: Relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems. *Phys. Rev. E* **2003**, 68.
(21) Gohlke, H.; Kuhn, L. A.; Case, D. A. Change in protein flexibility upon complex formation: Analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins* **2004**, *56*, 322−337.
(22) Rader, A. J.; Anderson, G.; Isin, B.; Khorana, H. G.; Bahar, I.; Klein-Seetharaman, J. Identification of core amino acids stabilizing rhodopsin. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7246−7251.
(23) Rader, A. J.; Bahar, I. Folding core predictions from network models of proteins. *Polymer* **2004**, *45*, 659−668.
(24) Mamonova, T.; Hespenheide, B.; Straub, R.; Thorpe, M. F.; Kurnikova, M. Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.* **2005**, *2*, S137−S147.
(25) Wells, S.; Menor, S.; Hespenheide, B. M.; Thorpe, M. F. Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.* **2005**, *2*, S127−S136.
(26) Livesay, D. R.; Jacobs, D. J. Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins* **2006**, *62*, 130−143.
(27) Ahmed, A.; Gohlke, H. Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. *Proteins* **2006**, *63*, 1038−1051.
(28) Radestock, S.; Gohlke, H. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.* **2008**, *8*, 507−522.
(29) Fulle, S.; Gohlke, H. Analyzing the flexibility of RNA structures by constraint counting. *Biophys. J.* **2008**, *94*, 4202−4219.
(30) Fulle, S.; Gohlke, H. Constraint counting on RNA structures: Linking flexibility and function. *Methods* **2009**, *49*, 181−188.
(31) Fulle, S.; Gohlke, H. Statics of the ribosomal exit tunnel: Implications for cotranslational peptide folding, elongation regulation, and antibiotics binding. *J. Mol. Biol.* **2009**, *387*, 502−517.
(32) Fulle, S.; Christ, N. A.; Kestner, E.; Gohlke, H. HIV-1 TAR RNA spontaneously undergoes relevant apo-to-holo conformational transitions in molecular dynamics and constrained geometrical simulations. *J. Chem. Inf. Model.* **2010**, *50*, 1489−1501.
(33) Mottonen, J. M.; Jacobs, D. J.; Livesay, D. R. Allosteric response is both conserved and variable across three CheY orthologs. *Biophys. J.* **2010**, *99*, 2245−2254.
(34) Rader, A. J. Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys. Biol.* **2010**, *7*, 016002.
(35) Rathi, P. C.; Radestock, S.; Gohlke, H. Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.* **2012**, *159*, 135−144.
(36) Pfleger, C.; Radestock, S.; Schmidt, E.; Gohlke, H. Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comput. Chem.* **2013**, *34*, 220−233.
(37) Wells, S. A.; Jimenez-Roldan, J. E.; Romer, R. A. Comparative analysis of rigidity across protein families. *Phys. Biol.* **2009**, 6.
(38) Zaccai, G. Biochemistry - How soft is a protein? A protein dynamics force constant measured by neutron scattering. *Science* **2000**, *288*, 1604−1607.
(39) Jacobs, D. J.; Dallakyan, S. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys. J* . **2005**, *88*, 903−915.
(40) Gonzalez, L. C.; Wang, H.; Livesay, D. R.; Jacobs, D. J. Calculating ensemble averaged descriptions of protein rigidity without sampling. *PLoS One* **2012**, 7.
(41) Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422−1423.
(42) Joosten, R. P.; Beek, T. A. H. T.; Krieger, E.; Hekkelman, M. L.; Hooft, R. W. W.; Schneider, R.; Sander, C.; Vriend, G. A series of PDB

related databases for everyday needs. *Nucleic Acids Res.* **2011**, *39*, D411−D419.
(43) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(44) Ascher, D.; Dubois, P. F.; Hinsen, K.; Hugunin, J.; Oliphant, T. *Numerical Python*, 2001.
(45) Jones, E.; Oliphant, T.; Peterson, P. *SciPy: Open Source Scientific tools for Python*, 2001
(46) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, 3.
(47) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: A Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2008**, 2.
(48) Beazley, D. M. Automated scientific software scripting with SWIG. *Future Gener. Comput. Syst.* **2003**, *19*, 599−609.
(49) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(50) Case, D.A.; T. A. D., Cheatham, T.E.; , III, Simmerling, C.L.; Wang, J.; Duke, R.E.; Luo, R.; Walker, R.C.; Zhang, W.; Merz, K.M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; I. Kolossváry, Wong, K.F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S.R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D.R.; Mathews, D.H.; Seetin, M.G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P.A. *AMBER 11*; University of California: San Francisco, 2010.
(51) Dahiyat, B. I.; Gordon, D. B.; Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci.* **1997**, *6*, 1333−1337.
(52) Privalov, P. L.; Gill, S. J. Stability of protein−structure and hydrophobic interaction. *Adv. Protein Chem.* **1988**, *39*, 191−234.
(53) Schellman, J. A. Temperature, stability, and the hydrophobic interaction. *Biophys. J.* **1997**, *73*, 2960−2964.
(54) Burnham, K. P.; Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information−Theoretic Approach*, 2. ed.; Springer: New York, 2002; pp XXVI, 488 S.
(55) Radford, S. E.; Dobson, C. M.; Evans, P. A. The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature* **1992**, *358*, 302−307.
(56) Matagne, A.; Radford, S. E.; Dobson, C. M. Fast and slow tracks in lysozyme folding: Insight into the role of domains in the folding process. *J. Mol. Biol.* **1997**, *267*, 1068−1074.
(57) Dinner, A. R.; Sali, A.; Smith, L. J.; Dobson, C. M.; Karplus, M. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* **2000**, *25*, 331−339.
(58) Kiefhaber, T. Kinetic traps in lysozyme folding. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9029−9033.
(59) Wildegger, G.; Kiefhaber, T. Three-state model for lysozyme folding: Triangular folding mechanism with an energetically trapped intermediate. *J. Mol. Biol.* **1997**, *270*, 294−304.
(60) Haliloglu, T.; Bahar, I. Structure-based analysis of protein dynamics: Comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data. *Proteins* **1999**, *37*, 654−667.
(61) Radford, S. E.; Buck, M.; Topping, K. D.; Dobson, C. M.; Evans, P. A. Hydrogen-exchange in native and denatured states of hen egg-white lysozyme. *Proteins* **1992**, *14*, 237−248.
(62) McCammon, J. A.; Gelin, B. R.; Karplus, M.; Wolynes, P. G. The hinge-bending mode in lysozyme. *Nature* **1976**, *262*, 325−6.
(63) Kohn, J. E.; Afonine, P. V.; Ruscio, J. Z.; Adams, P. D.; Head-Gordon, T. Evidence of functional protein dynamics from X-ray crystallographic ensembles. *PLoS Comput. Biol.* **2010**, 6.
(64) Ahmed, A.; Rippmann, F.; Barnickel, G.; Gohke, H. A normal mode-based geometric simulation approach for exploring biologically relevant conformational transitions in proteins. *J. Chem. Inf. Model.* **2011**, *51*, 1604−1622.
(65) Smith, L. J.; Sutcliffe, M. J.; Redfield, C.; Dobson, C. M. Structure of hen lysozyme in solution. *J. Mol. Biol.* **1993**, *229*, 930−944.
(66) Tzeng, S. R.; Kalodimos, C. G. Dynamic activation of an allosteric regulatory protein. *Nature* **2009**, *462*, 368−U139.
(67) Metz, A.; Pfleger, C.; Kopitz, H.; Pfeiffer-Marek, S.; Baringhaus, K. H.; Gohlke, H. Hot spots and transient pockets: Predicting the determinants of small-molecule binding to a protein−protein interface. *J. Chem. Inf. Model.* **2011**, *52*, 120−133.
(68) Krüger, D. M.; Rathi, P. C.; Pfleger, C.; Gohlke, H. CNA Web server: Rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo)stability, and function. *Nucleic Acids Res.* **2013**, DOI: 10.1093/nar/gkt292.

NOTE ADDED IN PROOF

A CNA Web server is available at http://cpclab.uni-duesseldorf.de/cna/.[68]

NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on April 8, 2013, with an error in reference 68. The corrected version was published to the Web on April 9, 2013.

Publication IV

CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo)stability, and function

Dennis M. Krüger[§], Prakash Chandra Rathi[§], Christopher Pfleger, and Holger Gohlke
Nucleic Acids Research, **2013**, 41:W340-348.
2012 impact factor: 8.27; contribution: 30%

[§] Both authors share first authorship

W340–W348 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* *Published online 22 April 2013*
doi:10.1093/nar/gkt292

CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo-)stability, and function

Dennis M. Krüger, Prakash Chandra Rathi, Christopher Pfleger and Holger Gohlke*

Computational Pharmaceutical Chemistry Group, Department of Mathematics and Natural Sciences, Heinrich-Heine-University, 40225 Düsseldorf, Germany

Received January 31, 2013; Revised March 21, 2013; Accepted March 31, 2013

Downloaded from http://nar.oxfordjournals.org/ at UniversitaetsbibliothekDuesseldorf on August 20, 2013

ABSTRACT

The Constraint Network Analysis (CNA) web server provides a user-friendly interface to the CNA approach developed in our laboratory for linking results from rigidity analyses to biologically relevant characteristics of a biomolecular structure. The CNA web server provides a refined modeling of thermal unfolding simulations that considers the temperature dependence of hydrophobic tethers and computes a set of global and local indices for quantifying biomacromolecular stability. From the global indices, phase transition points are identified where the structure switches from a rigid to a floppy state; these phase transition points can be related to a protein's (thermo-)stability. Structural weak spots (unfolding nuclei) are automatically identified, too; this knowledge can be exploited in data-driven protein engineering. The local indices are useful in linking flexibility and function and to understand the impact of ligand binding on protein flexibility. The CNA web server robustly handles small-molecule ligands in general. To overcome issues of sensitivity with respect to the input structure, the CNA web server allows performing two ensemble-based variants of thermal unfolding simulations. The web server output is provided as raw data, plots and/or Jmol representations. The CNA web server, accessible at http://cpclab.uni-duesseldorf.de/cna or http://www.cnanalysis.de, is free and open to all users with no login requirement.

INTRODUCTION

Proteins carry out their biological functions by interacting with other biomacromolecules or small molecules (1). These interactions require a certain degree of conformational adaptation to better complement the binding partners. That way, structural flexibility of proteins is linked to molecular recognition and catalysis (2). Additionally, flexibility (and its opposite, rigidity) also plays a central role for a protein's structural stability (3). Particularly, thermophilic proteins are in general more rigid than their mesophilic homologues to retain their fold at higher temperatures (4). Hence, knowing what can move in a protein is important for linking a protein's structure to its function and (thermo-)stability. Finally, information on protein flexibility is increasingly incorporated in computer-aided drug discovery and design projects (5).

X-ray crystallography, cryo-electron microscopy, single molecule fluorescence and nuclear magnetic resonance spectroscopy are experimental means from which the flexibility of a protein can be inferred (6–9). As an alternative, computational methods such as molecular dynamics (MD) simulations (10) and normal mode analysis are used to probe protein flexibility and dynamics (11). As yet another computational approach, a computationally highly efficient graph theory-based rigidity analysis for probing protein flexibility has been implemented in the Floppy Inclusions and Rigid Substructure Topography (FIRST) software (12). FIRST builds a constraint network from a biomolecular structure and then decomposes this network into rigid clusters and flexible regions by using the pebble game algorithm (13,14). In the constraint network, atoms are represented as bodies, and covalent bonds and non-covalent interactions (including hydrogen bonds, salt bridges, stacked rings, and hydrophobic tethers) are represented as bars (constraints) between them. Building on the ideas of Rader *et al*. (15) on diluting non-covalent constraints in a constraint network, our group has developed the Constraint Network Analysis (CNA) approach that performs thermal unfolding simulations of proteins (16). CNA goes beyond the mere identification of rigid clusters and flexible regions in a biomolecule

*To whom correspondence should be addressed. Tel: +49 211 81 13662; Fax: +49 211 13847; Email: gohlke@uni-duesseldorf.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2013. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

in that it (i) provides a refined modeling of thermal unfolding simulations that also considers the temperature dependence of hydrophobic tethers; (ii) allows performing rigidity analyses on ensembles of network topologies, either generated from structural ensembles or by using the concept of fuzzy non-covalent constraints; and (iii) computes a set of global and local indices for quantifying biomacromolecular stability (17). Furthermore, CNA robustly handles small-molecule ligands in general. CNA has been successfully used for investigating protein thermostability, identifying unfolding nuclei ('structural weak spots'), and linking protein flexibility and function (18–20).

In this study, we present the CNA web server that allows (i) setting up a variety of constraint network representations of proteins from either single structures or ensembles of structures; (ii) performing rigidity analyses and thermal unfolding simulations on these networks; and (iii) processing the results. For that, the CNA web server provides a layer of user-friendly input and output interfaces around the CNA software. As input, the web server only requires a Protein Data Bank (PDB) code or user-provided PDB file(s) of the input structure(s), and choosing the simulation type. Results are presented in the browser in an interactive manner. For global indices, plots are provided; for local indices, plots and mappings onto the 3D structure [via a JmolApplet (http://jmol.sourceforge.net)] are provided. Weak spots predicted for a protein structure are also mapped onto the 3D structure. To the best of our knowledge, there are no other web servers that allow performing and analyzing thermal unfolding simulations of proteins in as much detail as the CNA web server does. Of the two most closely related web servers, KINARI (http://kinari.cs.umass.edu/Site/kinariWeb.html) (21) only performs rigid cluster decompositions, but no thermal unfolding simulations. Flexweb (http://flexweb.asu.edu/software/first/) does allow performing thermal unfolding simulations; however, it neither supports the use of ensembles of structures (which leads to more robust results from rigidity analyses) nor does it automatically determine phase transition points ('melting points') and unfolding nuclei (which extends the application domain of rigidity analysis to data-driven protein engineering). Thus, in addition to characterizing the distribution of flexible and rigid regions in a protein, the CNA web server can be used for probing changes in the flexibility/(thermo-)stability of a protein due to mutations or on ligand binding, and it aids in identifying structural weak spots in a protein that, when mutated, may improve the protein's (thermo-)stability. Note that the CNA web server solely characterizes what can move in a protein but does not simulate actual protein movements. To do the latter, the user is referred to the NMSim web server (http://cpclab.uni-duesseldorf.de/nmsim/ or http://www.nmsim.de) developed by us (22).

MATERIALS AND METHODS

Constructing a constraint network

Proteins are modeled as body-and-bar networks where each atom is represented as a body with six degrees of freedom (23). Interactions between the atoms (covalent and non-covalent bonds) are modeled as a set of bars that restrict internal motion between the atoms. A covalent single bond is modeled with five bars allowing for the dihedral rotation about it; peptide and double bonds are modeled with six bars, disallowing any bond rotation. For example, a diatomic molecule with a single bond, owing to the five constraints, has seven degrees of freedom $(6 \times 2 - 5)$, six of which represent the trivial overall rotations and translations and one of which represents the internal rotation around the single bond. Non-covalent interactions, which contribute significantly to protein stability, are also modeled as bars. As such, hydrogen bonds (and salt bridges) are modeled with five bars, whereas hydrophobic and ring stacking interactions are modeled with two and three bars, respectively (24). Figure 1A shows the structure of thermolysin-like protease (TLP; PDB code: 1NPC) from which a body-and-bar network is then generated (Figure 1B).

Performing a rigid cluster decomposition

Once the constraint network of a protein is built, the pebble game algorithm (13,14) as implemented in the FIRST software (12) decomposes the network into rigid clusters and flexible regions (Figure 1C). The pebble game algorithm then computes the rigidity of the protein network at a bond level by determining whether a bond is part of a rigid cluster or a flexible region. As such, a rigid cluster is a set of atoms for which no internal motions are allowed and that move together in a collective manner.

Simulating thermal unfolding

By successively removing non-covalent constraints from a biomolecular constraint network, new network representations at elevated temperatures are constructed. To do so, for a given network state $s = f(T)$, hydrogen bonds (including salt bridges) with an energy $E_{hb} > E_{cut,hb}$ are removed from the network (25). This follows the idea that stronger hydrogen bonds will break at higher temperatures than weaker ones. The hydrogen bond energy scale E_{hb} (25) is converted into a temperature scale T, using a linear equation proposed by Radestock and Gohlke (19). By default, the number of hydrophobic tethers is kept constant during the thermal unfolding simulation. However, as the strength of hydrophobic interactions increases with increasing temperature (26,27), hydrophobic tethers can also be treated in a temperature-dependent manner on request (20). Finally, a rigid cluster decomposition is performed on each constraint network state s to compute global and local flexibility indices (Figure 1C).

Computing global and local flexibility indices

The CNA web server computes in total six global and three local flexibility indices. Detailed definitions of the global and local flexibility indices are given elsewhere (17). In short, the global indices represent the macroscopic network flexibility (and rigidity) of each network state s: (i) The floppy mode density Φ refers to the number of internal independent degrees of freedom that are associated

Downloaded from http://nar.oxfordjournals.org/ at UniversitaetsbibliothekDuesseldorf on August 20, 2013

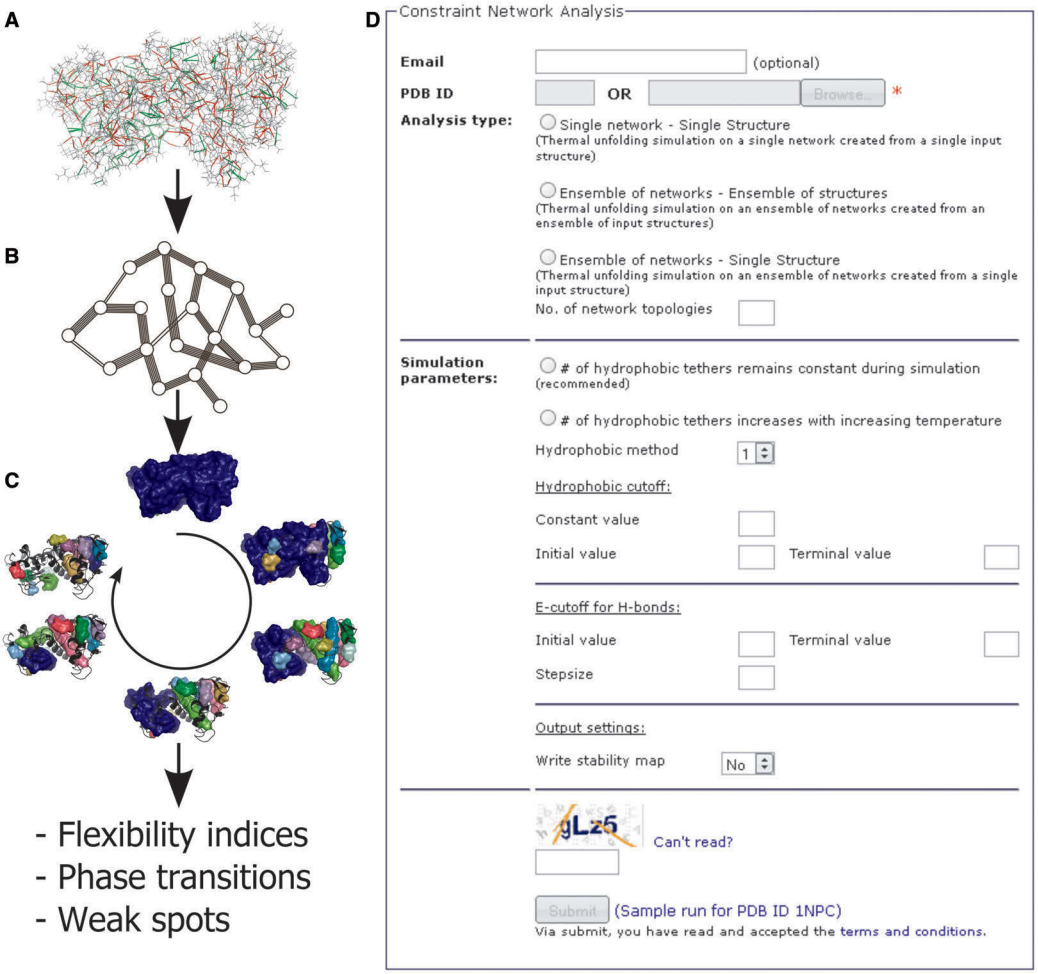


Figure 1. Covalent and non-covalent interactions in a protein structure (**A**) are modeled as bars in a body-and-bar network (**B**). A rigid cluster decomposition is carried out for all network states during a thermal unfolding simulation (**C**) and then post-processed to calculate flexibility indices, phase transitions, and weak spots. (**D**) Submission page to the CNA web server.

with dihedral rotations, normalized by the number of overall internal degrees of freedom associated with the total number of bodies in the network; (ii) the mean rigid cluster size S is computed with the size of the largest rigid cluster always being excluded; (iii) & (iv) the rigidity order parameter P_∞ denotes the fraction of the network belonging to the giant percolating cluster (type 1) or the actual largest rigid cluster (type 2). The giant percolating cluster is the largest rigid cluster in the network state at the lowest temperature, i.e. with all constraints in place. During the thermal unfolding simulation, the melting of the giant percolating cluster is monitored, and the largest rigid subcluster of the previous giant percolating cluster becomes the new giant percolating cluster of the present network state s. In contrast, the actual largest rigid cluster is the largest cluster present at a network state s, irrespective of its evolutionary history during the thermal unfolding simulation; (v) & (vi) the cluster configuration entropy H (type 1 and type 2) is a measure of the degree of disorder in the realization of a given network state.

Local indices characterize the flexibility of network at the bond level by monitoring the change in the flexibility for each bond during the thermal unfolding simulation: (i) the percolation index p_i is determined for each covalent bond by the $E_{cut,hb}$ value during the thermal unfolding simulation at which the bond segregates from the giant percolating cluster; (ii) the rigidity index r_i is determined for each covalent bond by the $E_{cut,hb}$ value during the thermal unfolding simulation at which the bond changes from rigid to flexible. For a C_α atom-based representation

Downloaded from http://nar.oxfordjournals.org/ at UniversitaetsbibliothekDuesseldorf on August 20, 2013

of a protein structure, the lower of the p_i values (the average of the r_i values) of the two backbone bonds is taken; (iii) a stability map rc_{ij} is a 2D itemization of the rigidity index r_i and is derived by identifying the $E_{cut,hb}$ value during the thermal unfolding simulation at which a rigid contact between a pair of residues represented by their C_α atoms is lost. Two residues are in rigid contact if they are part of the same rigid cluster.

Identifying phase transitions and unfolding nuclei

The global flexibility indices are used for identifying phase transitions during the thermal unfolding simulation when the network switches from being largely rigid to flexible. Such transitions ('melting points') can be related to the (thermo-)stability of proteins (18–20). The CNA web server provides phase transition points for four of the global indices: $P_{\infty,\text{ type 1 and 2}}$ and $H_{\text{type 1 and 2}}$. A smoothed spline fitted to the global indices is used to identify the phase transition points except for H_{type2}, for which a double sigmoid curve is fitted. The phase transition points are identified as the maximum in the first derivative in the case of the smoothed spline and the maximum of the differences in the asymptote pairs in the case of the double sigmoid curve. The phase transition points can be further exploited to identify structural weak spots in the network from where an unfolding begins: these are residues that are part of the largest rigid cluster before the phase transition and become flexible afterwards. Weak spots provide hints as to where introducing mutations in a structure may improve a protein's (thermo-)stability.

Simulating thermal unfolding of a network ensemble

Results from rigidity analyses are in general sensitive to the input structure in that small changes in the conformation can lead to a different rigid cluster decomposition (28,29). To overcome this drawback, an ensemble-based variant of CNA has been developed that makes use of an ensemble of structures derived from MD simulations (20,28). Structural ensembles from any other sources can be used as input, too. As yet another alternative, the CNA web server provides an option to create an ensemble of networks from a single input structure by using fuzzy non-covalent constraints (Pfleger, C., Gohlke, H., unpublished data). Here, the number and distribution of non-covalent constraints are modulated by random components within certain ranges, that way simulating thermal fluctuations of a biomolecule without actually moving atoms. This approach avoids the use of computationally expensive MD simulations. For ensemble-based CNA, averages are computed over the entire ensemble for phase transition points and local indices. In the case of weak spots, the frequencies of occurrence across the entire ensemble are reported.

DESCRIPTION OF THE WEB SERVER

Input

The submission page to the CNA web server is shown in Figure 1D. The CNA web server requires either a single structure of a protein provided as a PDB file or a PDB-ID (in which case the PDB file will be downloaded from the Research Collaboratory for Structural Bioinformatics repository), or multiple protein structures provided as PDB files in a compressed folder (allowed file formats are *.tgz, *.tar.gz or *.zip) or as a 'multi-PDB file' using MODEL/ENDMDL cards. Furthermore, the web server requires selection of an analysis type and input of a given security code to prevent misuse. Analyses can be performed on a single constraint network derived from a single structure, on an ensemble of networks derived from a single structure or on an ensemble of networks derived from a structural ensemble. According to the selected analysis type, default parameters will be provided that have been successfully used in previous studies (17–20). In addition, the user can request the time-consuming computation of a stability map. An email address can be provided in which case a link to the results will be sent to that address. In either case, a link to a results page is provided after job submission for monitoring the progress of the computations and viewing the results in the web browser. The results will also be stored on the server for 10 days and can be accessed via the provided link.

If the user provides a PDB ID, missing sidechain and hydrogen atoms will be automatically added by the leap program of the AMBER suite (30). For amino acid sidechains, a standard protonation state is assumed, i.e. Asp and Glu are treated as deprotonated, and Arg and Lys as protonated. By default, His is singly protonated with the hydrogen on the epsilon nitrogen. Alternate sidechain conformations are allowed, but only the first conformation is considered. Water molecules and ligands will be removed.

To have complete control on the input structure(s), a user can upload a/multiple PDB file(s). In that case, water molecules and ligands are considered, and correct bond orders for ligands are identified automatically using the Open Babel wrapper (Pybel) for Python (31,32). The correct identification of ligand bond orders is important to accurately set up covalent constraints for the network. In this case (i) hydrogen atoms must have been added already, (ii) alternate side-chain conformations are not allowed; and (iii) the protein structure must be complete, i.e. no missing sidechains are allowed. Finally, if the analysis is based on an ensemble of structures, all protein structures must be identical except for the coordinates. Otherwise, an error message will be issued.

Output and representation of results

A typical analysis run of the CNA web server for a constraint network of a single structure takes a few minutes; for the analysis of an ensemble, several hours of computing time may be required. After the start, the progression of the analysis is reported on the results page. On completion of a job, the results are presented in the web browser and, if an address is provided by the user, an email is sent with a link to the results, too.

The first part of the results page contains a summary of input parameters that were provided by the user and a download link to the modified input file. The modified

Downloaded from http://nar.oxfordjournals.org/ at Universitaetsbibliothek Duesseldorf on August 20, 2013

input file contains all changes made during preparation of the uploaded structure.

The second part of the results page contains a table that provides information about identified phase transition points. In the case of a single-network analysis, six plots of the global flexibility indices are depicted: floppy mode density Φ (Figure 2A), mean rigid cluster size S (Figure 2B), rigidity order parameter $P_{\infty,\ \text{type 1 and 2}}$ (Figure 2C and D) and cluster configuration entropy $H_{\text{type 1 and 2}}$ (Figure 2E and F). Identified phase transition points are marked by red vertical lines. In the case of an ensemble-based analysis, a summary of the identified phase transition points is presented instead. In any case, a file with the raw data can be downloaded by a link given next to the headline of the chapter for further evaluation.

The third part contains two plots of the local flexibility indices for each residue: percolation index p_i (Figure 3A) and rigidity index r_i (Figure 3B). In the case of an ensemble-based analysis, the standard errors are depicted in gray in addition to the mean values. Furthermore, both indices are mapped onto the input structure in a color-coded fashion and shown in JmolApplets. A file with the raw data can be downloaded by a link given next to the headline of the chapter for further evaluation as can be two PDB files. These PDB

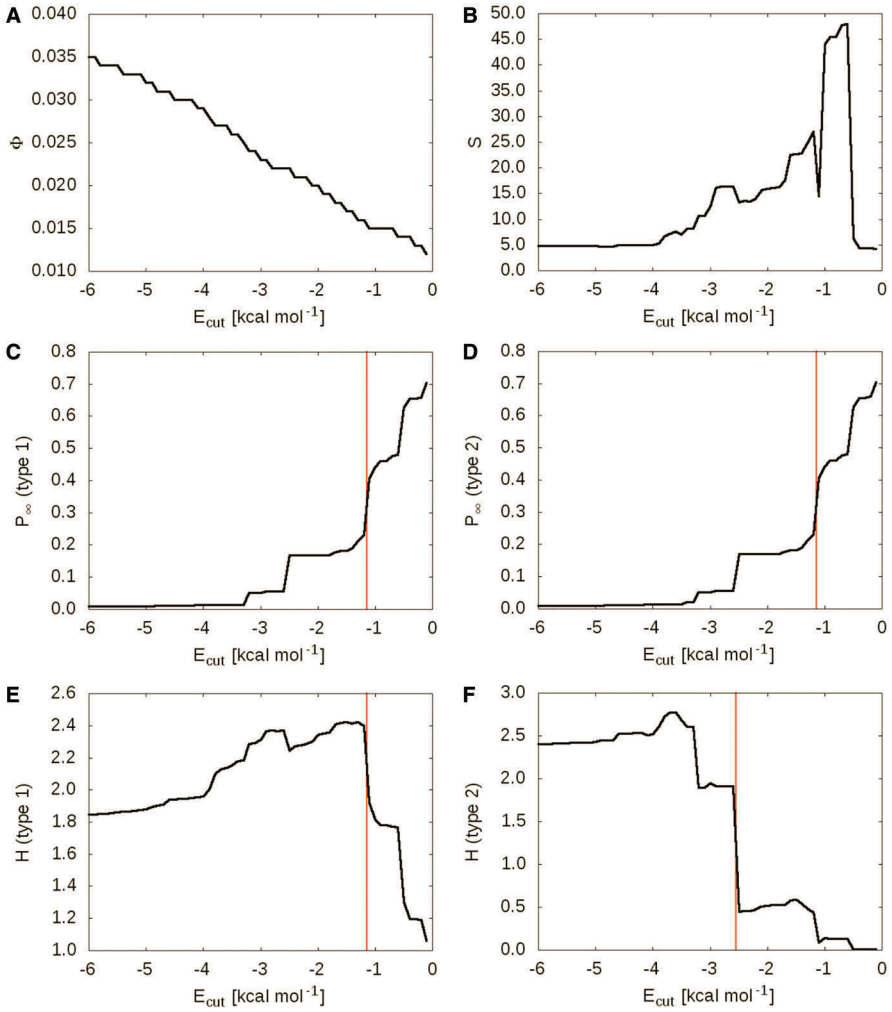


Figure 2. Global indices for the thermal unfolding of TLP as a function of the hydrogen bonding energy cutoff $E_{cut,hb}$: (**A**) floppy mode density Φ; (**B**) mean rigid cluster size S; rigidity order parameter P_∞ (**C**) type 1 and (**D**) type 2; cluster configuration entropy H (**E**) type 1 and (**F**) type 2. The red vertical lines (C–F) correspond to the identified phase transitions.

Downloaded from http://nar.oxfordjournals.org/ at UniversitaetsbibliothekDuesseldorf on August 20, 2013

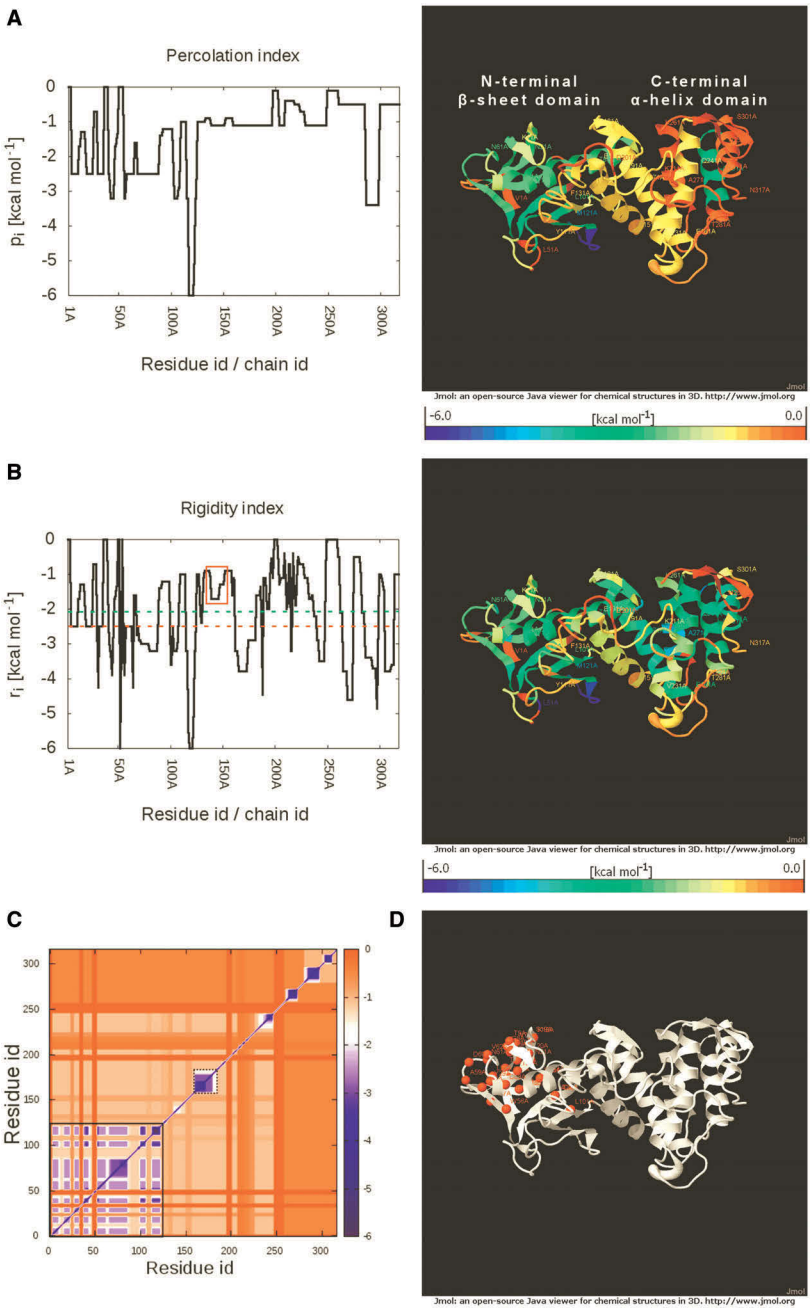


Figure 3. (**A**) Percolation index p_i for TLP. The lower p_i the longer is a residue part of the giant percolating cluster during the thermal unfolding simulation. (**B**) Rigidity index r_i for TLP. The lower the r_i, the longer is a residue part of a rigid cluster during the thermal unfolding simulation. Red- and green-dashed horizontal lines represent the identified phase transition point and the working temperature of TLP, respectively. The central α-helix and two preceding Gly residues (residues 136–154) residues are enclosed in a red rectangle. On the right, the respective indices are mapped onto the input structure in a color-coded manner. (**C**) Stability map rc_{ij} for TLP. Red colors indicate pairs of residues where no or only a weak rigid contact exists. In contrast, blue colors indicate strong rigid contacts. The black box with a continuous line covers the N-terminal giant rigid cluster, whereas the box with the broken line indicates a rigid cluster in the C-terminal domain. (**D**) Weak spots in the TLP structure are represented by red spheres.

Downloaded from http://nar.oxfordjournals.org/ at Universitaetsbibliothek Duesseldorf on August 20, 2013

files contain the modified input structure of the protein with the respective index values of C_α atoms in the B-factor column.

The fourth part contains a stability map rc_{ij} (Figure 3C) (if requested) and information about weak spots identified in the protein (Figure 3D). For rc_{ij}, the color code reveals for each residue pair if the rigid contact is weak (red) or strong (blue). Information about weak spots is mapped onto the input structure and displayed in a JmolApplet: in the case of a single-network analysis, identified weak spots are marked by red spheres; in the case of an ensemble-based analysis, the frequency of a residue for being identifed as a weak spot across the entire ensemble is depicted. This information is provided as a plot, too. Again, two files can be downloaded by links given next to the headline of the chapter for further evaluation: a file with the raw data and a PDB file containing the modified input structure of the protein with the respective frequency values in the B-factor column.

Implementation

The CNA web server has been implemented in Python, as have been the underlying CNA routines (16). All plots are generated with Gnuplot. Given the low computational demand of our approach, up to 10 submitted jobs can be run in parallel at present.

APPLICATION TO TLP AS A TEST CASE

In a previous study from our group, the CNA approach has been shown to link protein rigidity and thermostability by correctly predicting which one of two proteins is more thermostable for two thirds of a data set of 19 pairs of proteins from mesophilic and thermophilic organisms (18). In the same study, we showed that structural weak spots identified by CNA agreed with the positions of thermostabilizing mutations for two systems investigated. Extending the application domain of CNA one step further, the local flexibility and rigidity distributions in the active sites of 3-isopropylmalate dehydrogenase and TLP were then linked to enzymatic activity at different temperatures (19). Recently, using CNA on citrate synthases with distinct thermostabilities, we showed that thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio (14).

Here, we demonstrate the application of the CNA web server to TLP from *Bacillus cereus* (PDB code: 1NPC) using a single-network approach with default settings. These results are also available as a sample run on the web server. Plots for global indices are shown in Figure 2. The decrease in network rigidity of TLP with increasing temperature (equivalent to a decreasing $E_{cut,hb}$) is evident from the plots of all six global indices. Interestingly, the decay of the giant rigid cluster occurs in a hierarchical fashion as reflected by the presence of multiple steps in the P_∞ and H profiles (Figure 2C–F). The reason for this percolation behavior is that the TLP structure is composed of multiple sub-domains (N-terminal β-sheet domain and C-terminal α-helical domain connected by a central helix) that segregate from the giant cluster independently from each other. Phase transition points at which the TLP structure sharply loses rigidity on thermal unfolding are then identified: an early phase transition point at $E_{cut,hb} = -1.14\,\mathrm{kcal\,mol^{-1}}$ is identified from the P_∞ and H_{type1} profiles when the C-terminal α-helical domain segregates from the giant cluster; in contrast, $H_{\mathrm{type\ 2}}$ fosters the identification of a late transition at $E_{cut,hb} = -2.55\,\mathrm{kcal\,mol^{-1}}$ [or, equivalently, 351 K (19)], which represents the final substantial decay of the rigid core. Such late transitions have been found to best relate to the melting of a protein. The temperatures of the transition points should be considered relative values only because the absolute values may depend on the size and architecture of the analyzed protein (19). Still, the temperatures are helpful, e.g. when it comes to comparing the thermostability of two or more homologous proteins or the stability of a wild-type with its mutant (18–20). Accordingly, the $H_{\mathrm{type\ 2}}$ phase transition point for a thermophilic homologue of TLP, thermolysin from *Bacillus thermoproteolyticus*, was found at 373 K (24), ~20 K higher than for the mesophilic homologue investigated here.

The percolation index plot depicts the percolation behavior of TLP at a residue level (Figure 3A). The C-terminal α-helical domain segregates from the giant cluster early during the thermal unfolding simulation at $E_{cut,hb} \approx -1.1\,\mathrm{kcal\,mol^{-1}}$. The giant percolating cluster, which then consists of mainly the β-sheet region and an α-helix in the N-terminal domain, disintegrates into smaller clusters at the later phase transition point ($E_{cut,hb} = -2.55\,\mathrm{kcal\,mol^{-1}}$). Unsurprisingly, the rigidity index, which monitors the flexibility at a residue level, shows secondary structure elements to be more rigid than loops. As an exception, residues 117–120 forming a loop in the N-terminal domain are always part of a rigid cluster throughout the thermal unfolding simulation owing to a network of hydrophobic tethers. For TLPs, the central α-helix (residues 139–154) and the preceding Gly136 and Gly137 are important with respect to a postulated hinge bending motion (33,34), suggesting that these residues should be flexible at the working temperature of TLP [342 K, equivalent to $E_{cut,HB} = -2.1\,\mathrm{kcal\,mol^{-1}}$ (19)]. This is also found in the rigidity index profile (Figure 3B). Furthermore, contacts of these residues with other residues of TLP are less stable than contacts between residues of the giant cluster (black box with continuous line) and contacts between residues of a large rigid cluster in the C-terminal domain (black box with broken line) as identified by the stability map (Figure 3C). This is again in line with the hinge character of the central α-helix. Finally, several of the weak spots identified by CNA in the N-terminal β-sheet domain of TLP (Figure 3D) have been shown to improve the protein's thermostability on mutation in previous studies (35–38).

CONCLUSIONS

Increasing evidence of the importance of protein flexibility has warranted the development of efficient and accurate computational tools for characterizing protein flexibility

Downloaded from http://nar.oxfordjournals.org/ at Universitaetsbibliothek Duesseldorf on August 20, 2013

at global and local levels. To this end, the CNA approach has been developed for deriving maximal advantage from information on biomolecular flexibility and rigidity by linking results from rigidity analyses to biologically relevant characteristics of a structure, such as (thermo-) stability and function (16). In particular, CNA provides a refined modeling of thermal unfolding simulations that also considers the temperature dependence of hydrophobic tethers, allows performing rigidity analyses on ensembles of network topologies and computes a set of global and local indices for quantifying biomacromolecular stability. Furthermore, CNA robustly handles small-molecule ligands in general. To make these computations available for users even with only minimal or no prior knowledge of structural bioinformatics techniques, we developed the CNA web server. It provides a user-friendly interface, requires minimal input and displays the results intuitively both as plots and mappings onto the protein structure via JmolApplets. As a typical analysis run of the CNA web server for a constraint network of a single structure only takes a few minutes, we strongly believe that the CNA web server will be a valuable (interactive) tool for data-driven protein engineering and estimating the influence of ligand molecules on biomolecular stability.

ACKNOWLEDGEMENTS

The authors are grateful to Prof. Michael F. Thorpe (Arizona State University) for granting a FIRST license. They thank Daniel B. Ciupka, Doris L. Klein, Tobias Kröger and Giulia Pagani (Heinrich Heine University, Düsseldorf) for testing the web server and helpful suggestions.

FUNDING

Ministry of Innovation, Science, and Research of North Rhine-Westphalia and Heinrich Heine University, Düsseldorf, via a scholarship to P.C.R. within the CLIB-Graduate Cluster Industrial Biotechnology. Funding for open access charge: Institutional funding from Heinrich-Heine-University.

Conflict of interest statement. None declared.

REFERENCES

1. Schellman,J.A. (1975) Macromolecular binding. *Biopolymers*, **14**, 999–1018.
2. Daniel,R.M., Dunn,R.V., Finney,J.L. and Smith,J.C. (2003) The role of dynamics in enzyme activity. *Annu. Rev. Biophys. Biomol. Struct.*, **32**, 69–92.
3. Vihinen,M. (1987) Relationship of protein flexibility to thermostability. *Protein Eng.*, **1**, 477–480.
4. Závodszky,P., Kardos,J., Svingor,Á. and Petsko,G.A. (1998) Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Natl Acad. Sci. U.S.A.*, **95**, 7406–7411.
5. Cozzini,P., Kellogg,G.E., Spyrakis,F., Abraham,D.J., Costantino,G., Emerson,A., Fanelli,F., Gohlke,H., Kuhn,L.A., Morris,G.M. *et al.* (2008) Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.*, **51**, 6237–6255.
6. Frank,J. and Agrawal,R.K. (2000) A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature*, **406**, 318–322.
7. Ishima,R. and Torchia,D.A. (2000) Protein dynamics from NMR. *Nat. Struct. Mol. Biol.*, **7**, 740–743.
8. Weiss,S. (1999) Fluorescence spectroscopy of single biomolecules. *Science*, **283**, 1676–1683.
9. Zhang,X.J., Wozniak,J.A. and Matthews,B.W. (1995) Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozymes. *J. Mol. Biol.*, **250**, 527–552.
10. Karplus,M. and McCammon,J.A. (2002) Molecular dynamics simulations of biomolecules. *Nat. Struct. Mol. Biol.*, **9**, 646–652.
11. Case,D.A. (1994) Normal mode analysis of protein dynamics. *Curr. Opin. Struct. Biol.*, **4**, 285–290.
12. Jacobs,D.J., Rader,A.J., Kuhn,L.A. and Thorpe,M.F. (2001) Protein flexibility predictions using graph theory. *Proteins*, **44**, 150–165.
13. Jacobs,D.J. and Hendrickson,B. (1997) An algorithm for two-dimensional rigidity percolation: the pebble game. *J. Comp. Phys.*, **137**, 346–365.
14. Jacobs,D.J. and Thorpe,M.F. (1995) Generic rigidity percolation: the pebble game. *Phys. Rev. Lett.*, **75**, 4051–4054.
15. Rader,A.J., Hespenheide,B.M., Kuhn,L.A. and Thorpe,M.F. (2002) Protein unfolding: rigidity lost. *Proc. Natl Acad. Sci. U.S.A.*, **99**, 3540–3545.
16. Pfleger,C., Rathi,P.C., Klein,D.L., Radestock,S. and Gohlke,H. (2013) Constraint Network Analysis (CNA): A Python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. *J. Chem. Inf. Model*, **53**, 1007–1015.
17. Pfleger,C., Radestock,S., Schmidt,E. and Gohlke,H. (2013) Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comp. Chem.*, **34**, 220–233.
18. Radestock,S. and Gohlke,H. (2008) Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.*, **8**, 507–522.
19. Radestock,S. and Gohlke,H. (2011) Protein rigidity and thermophilic adaptation. *Proteins*, **79**, 1089–1108.
20. Rathi,P.C., Radestock,S. and Gohlke,H. (2012) Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.*, **159**, 135–144.
21. Fox,N., Jagodzinski,F., Li,Y. and Streinu,I. (2011) KINARI-Web: a server for protein rigidity analysis. *Nucleic Acids Res.*, **39**, W177–W183.
22. Krüger,D.M., Ahmed,A. and Gohlke,H. (2012) NMSim web server: integrated approach for normal mode-based geometric simulations of biologically relevant conformational transitions in proteins. *Nucleic Acids Res.*, **40**, W310–W316.
23. Whiteley,W. (2005) Counting out to the flexibility of molecules. *Phys. Biol.*, **2**, S116–S126.
24. Hespenheide,B.M., Jacobs,D.J. and Thorpe,M.F. (2004) Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. *J. Phys. Condens. Matter*, **16**, S5055–S5064.
25. Dahiyat,B.I., Gordon,D.B. and Mayo,S.L. (1997) Automated design of the surface positions of protein helices. *Protein Sci.*, **6**, 1333–1337.
26. Privalov,P.L. and Gill,S.J. (1988) Stability of protein structure and hydrophobic interaction. *Adv. Protein Chem.*, **39**, 191–234.
27. Schellman,J.A. (1997) Temperature, stability, and the hydrophobic interaction. *Biophys. J.*, **73**, 2960–2964.
28. Gohlke,H., Kuhn,L.A. and Case,D.A. (2004) Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins*, **56**, 322–337.
29. Mamonova,T., Hespenheide,B., Straub,R., Thorpe,M.F. and Kurnikova,M. (2005) Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.*, **2**, S137–S147.
30. Case,D.A., Cheatham,T.E. III, Darden,T., Gohlke,H., Luo,R., Merz,K.M. Jr, Onufriev,A., Simmerling,C., Wang,B. and Woods,R.J. (2005) The Amber biomolecular simulation programs. *J. Comp. Chem.*, **26**, 1668–1688.
31. O'Boyle,N.M., Banck,M., James,C.A., Morley,C., Vandermeersch,T. and Hutchison,G.R. (2011) Open Babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33.

Downloaded from http://nar.oxfordjournals.org/ at Universitaetsbibliothek Duesseldorf on August 20, 2013

32. O'Boyle,N.M., Morley,C. and Hutchison,G.R. (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.*, **2**, 5.
33. van Aalten,D.M., Amadei,A., Linssen,A.B., Eijsink,V.G., Vriend,G. and Berendsen,H.J. (1995) The essential dynamics of thermolysin: confirmation of the hinge-bending motion and comparison of simulations in vacuum and water. *Proteins*, **22**, 45–54.
34. Veltman,O.R., Eijsink,V.G., Vriend,G., de Kreij,A., Venema,G. and Van den Burg,B. (1998) Probing catalytic hinge bending motions in thermolysin-like proteases by glycine –> alanine mutations. *Biochemistry*, **37**, 5305–5311.
35. Imanaka,T., Shibazaki,M. and Takagi,M. (1986) A new way of enhancing the thermostability of proteases. *Nature*, **324**, 695–697.
36. Van den Burg,B., Dijkstra,B.W., Vriend,G., Vandervinne,B., Venema,G. and Eijsink,V.G.H. (1994) Protein stabilization by hydrophobic interactions at the surface. *Eur. J. Biochem.*, **220**, 981–985.
37. 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.
38. Veltman,O.R., Vriend,G., Middelhoven,P.J., van den Burg,B., Venema,G. and Eijsink,V.G. (1996) Analysis of structural determinants of the stability of thermolysin-like proteases by molecular modelling and site-directed mutagenesis. *Protein Eng.*, **9**, 1181–1189.

Downloaded from http://nar.oxfordjournals.org/ at UniversitaetsbibliothekDuesseldorf on August 20, 2013

Publication V

VisualCNA: A GUI for interactive Constraint Network Analysis and protein engineering for improving thermostability

Prakash Chandra Rathi[§], Daniel Mulnaes[§] and Holger Gohlke

Submitted manuscript, **2014.**
Contribution: 40%

[§] Both authors share first authorship

Structural bioinformatics

VisualCNA: A GUI for interactive Constraint Network Analysis and protein engineering for improving thermostability

Prakash Chandra Rathi†, Daniel Mulnaes† and Holger Gohlke*

Institute for Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich Heine University, Universitätsstr. 1, Düsseldorf, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Constraint Network Analysis (CNA) is a graph theorybased rigidity analysis approach for linking a biomolecule's structure, flexibility, (thermo)stability, and function. Results from CNA are highly information-rich and require intuitive, synchronized, and interactive visualization for a comprehensive analysis. We developed VisualCNA, an easy-to-use PyMOL plug-in that allows setup of CNA runs and analysis of CNA results linking plots with molecular graphics representations. From a practical viewpoint, the most striking feature of VisualCNA is that it facilitates interactive protein engineering aimed at improving thermostability.

Availability and Implementation: VisualCNA and its dependencies (CNA and FIRST software) are available free of charge under GPL and academic licenses, respectively. VisualCNA and CNA are available at http://cpclab.uni-duesseldorf.de/software; FIRST is available at http://flexweb.asu.edu.

Contact: Gohlke@uni-duesseldorf.de

1 INTRODUCTION

Structural flexibility (and rigidity) is important for a protein's function and stability. Being able to accurately predict protein flexibility is thus instrumental in protein-science and -engineering as well as drug design. Our group developed the Python-based software package Constraint Network Analysis (CNA) (Pfleger *et al.*, 2013) for characterizing biomolecular flexibility both at the global and residue level. CNA functions as a front- and back-end to the graph theory based rigidity analysis software FIRST (Jacobs *et al.*, 2001) and has been used to predict protein thermostability, identify structural weak spots (Radestock and Gohlke, 2008; Radestock and Gohlke, 2011; Rathi *et al.*, 2012), and link protein flexibility and function, including allosteric regulation (Pfleger and Gohlke, unpublished results). CNA models a protein as a body-and-bar constraint network in which bodies (atoms) are connected by bars (covalent and non-covalent interactions). A rigidity analysis is then performed using the *pebble game* algorithm (Jacobs and Thorpe, 1995), resulting in a decomposition of the network into rigid and flexible regions. By removing non-covalent constraints from the network in the order of increasing strength, CNA simulates thermal unfolding. From the unfolding trajectory, CNA calculates several global and local flexibility indices (Pfleger *et al.*, 2013). From the global indices, phase transition points are identified at which the network switches from rigid to flexible. These points are used to predict the thermostability of a protein and identify structural weak spots. Local indices describe intrinsic biomolecular flexibility at the level of bonds and can be compared with quantitative data from experiments on biomolecular mobility.

The output from CNA is highly information-rich. For a comprehensive analysis, the data needs to be visualized as plots (showing global and local flexibility indices) as well as 3D graphics representations of the biomolecule, the constraint network, and the decomposition into rigid and flexible regions. Furthermore, the speed of CNA allows performing real-time rigidity analyses on biomolecules. Thus, interactive structural modifications and/or editing of the constraint network followed by a re-analysis of the biomolecule's flexibility can be performed iteratively. With this in mind, we developed VisualCNA, an intuitive, easy-to-use graphical interface for CNA built as a plug-in for the molecular viewer PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.3 Schrödinger, LLC.). VisualCNA supports scientists interested in the rigidity analysis of biomolecules by a synchronized and interactive visualization of the CNA output. It also enables interactive protein engineering for improving thermostability by iteratively mutating identified weak spots, performing a subsequent rigidity analysis, and automatically comparing CNA output from wild type and mutant structures.

2 IMPLEMENTATION

VisualCNA is implemented in the Python programming language as a PyMOL plug-in for Linux operating systems. It uses the external modules NumPy, SciPy, Matplotlib, Biopython, tkintertable, and Open Babel. All modules and the PyMOL source code are packaged alongside VisualCNA for easy installation. VisualCNA requires CNA and FIRST for rigidity analysis, which are distributed independently. User manual and tutorial videos are also distributed with VisualCNA.

3 DESCRIPTION

The VisualCNA GUI consists of four panels: *Setup*, *Analyze*, *Modify*, and *Mutate*. The *Setup* panel allows preparing variations of thermal unfolding simulations, i.e., based on a single network

*To whom correspondence should be addressed.

†The authors wish it to be known that in their opinion, the first two authors should be regarded as joint First Authors.

© Oxford University Press 2005

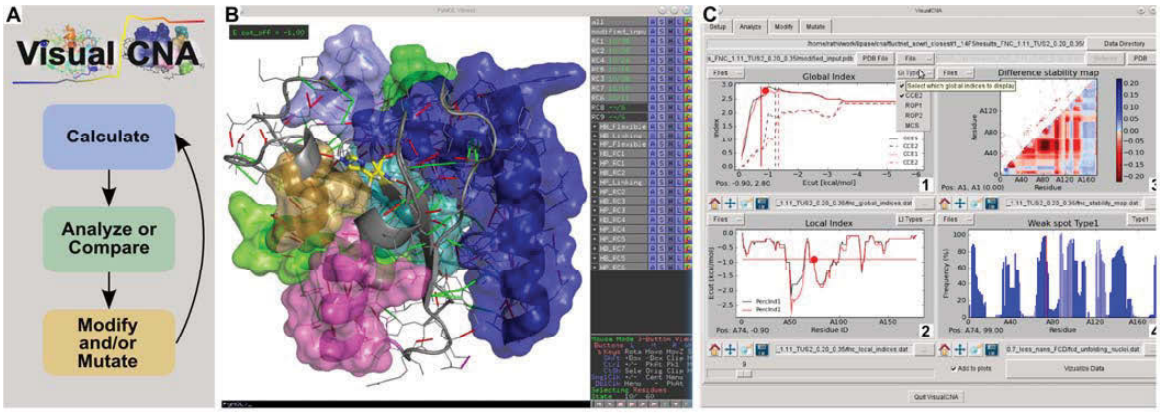


Fig. 1. A: Illustration of VisualCNA's iterative work flow for optimization of protein thermostability. B: PyMOL window showing the 3D protein structure at the melting point. Rigid clusters are shown as uniformly colored semi-transparent bodies. Constraints due to hydrogen bonds, salt bridges, and hydrophobic contacts are shown as red, magenta, and green sticks. A mutation is shown in yellow stick representation. Flexible regions are shown in grey. C: The VisualCNA *Analyze* panel showing a comparison of multiple graphs from wild-type (black) and mutant (red) analyses. 1: Global indices with transition points indicated as vertical lines. 2: Local index with a red circle indicating the mutation and a horizontal red line showing the unfolding state. 3: Difference stability map between wild-type and mutant. 4: Likelihood of a residue of being a structural weak spot with the mutant shown in red.

derived from a single input structure (Radestock and Gohlke, 2008), an ensemble of networks derived from a single structure using definitions of fuzzy non-covalent constraints (Pfleger and Gohlke, 2013), or a structural ensemble (Rathi *et al.*, 2012).

In the *Analyze* panel (Fig. 1C), CNA output from multiple thermal unfolding simulations can be shown simultaneously, which helps comparing wild type and mutants. CNA results are shown as interactive plots of global and local flexibility indices and weak spots. In parallel, an interactive 3D protein structure (Fig. 1B) is visualized in PyMOL in terms of states corresponding to the steps of the thermal unfolding trajectory. The trajectory can be played as an interactive movie and is linked with the flexibility indices by annotations in the plots. Clicking either the plots or the structure changes the appearance of both to focus on the selected residue and/or the corresponding unfolding state. Constraints are grouped by their associated rigid cluster or flexible region to aid visualization and selection. Constraints about to break in a given state are grouped, too, facilitating the identification of residues that could be mutated to stabilize these.

The *Modify* panel contains several ways to modify the constraint network of the protein, which is useful when modeling the effect of ligands, ions, or non-standard residues. An interactive table of constraints uses check boxes to enable or disable constraints, and a search box allows navigation. User defined constraints can be added by specifying atom pairs in text fields or the 3D structure.

The *Mutate* panel is central to the interactive protein engineering capability of VisualCNA. After loading an alignment of multiple sequences, the residue conservation and substitution frequencies are calculated for each residue. Clicking a bar in the conservation plot and a mutation in the substitution frequency plot then mutates the corresponding residue and updates the constraint network. The mutation is done using the PyMOL mutation tool, which allows the user to select an appropriate rotamer for the mutant. The new structure can be automatically submitted for unfolding simulation and compared to the wild type. In this way, the effect of point mutations can be iteratively analyzed to optimize the protein structure towards increased thermostability (Fig. 1A).

In summary, the CNA approach derives maximal advantage from information on biomolecular flexibility by linking results from rigidity analyses to relevant structural characteristics. With VisualCNA, an intuitive, easy-to-use graphical interface is available that makes CNA studies amenable to non-bioinformaticians interested in rigidity analysis of biomolecules and interactive protein engineering.

ACKNOWLEDGEMENTS

We are grateful to the Ministry of Innovation, Science, and Research of North Rhine-Westphalia and Heinrich Heine University for scholarships to PCR and DM within the CLIB-Graduate Cluster Industrial Biotechnology. We thank Christopher Pfleger, Daniel B. Ciupka, and Anuseema for testing and helpful suggestions.

REFERENCES

Jacobs,D.J., *et al.* (2001) Protein flexibility predictions using graph theory. *Proteins: Struct., Funct., Bioinf.*, **44**,150-165.

Jacobs,D.J. and Thorpe,M.F. (1995) Generic rigidity percolation: the pebble game. *Phys. Rev. Lett.*, **75**,4051-4054.

Pfleger,C. and Gohlke,H. (2013) Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. *Structure*, **21**,1725-1734.

Pfleger,C., *et al.* (2013) Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comp. Chem.*, **34**,220-233.

Pfleger,C., *et al.* (2013) Constraint Network Analysis (CNA): A Python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. *J. Chem. Inf. Model.*, **53**,1007-1015.

Radestock,S. and Gohlke,H. (2008) Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.*, **8**,507-522.

Radestock,S. and Gohlke,H. (2011) Protein rigidity and thermophilic adaptation. *Proteins: Struct., Funct., Bioinf.*, **79**,1089-1108.

Rathi,P.C., *et al.* (2012) Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.*, **159**,135-144.

Publication VI

Structural rigidity and protein thermostability

Prakash Chandra Rathi, Karl-Erich Jaeger and Holger Gohlke
Submitted manuscript, **2014.**
Contribution: 75%

Abstract

Understanding the origin of thermostability is of fundamental importance in protein biochemistry. Opposing views on increased or decreased structural rigidity of the folded state have been put forward in this context. They have been related to differences in the temporal resolution of experiments and computations that probe atomic mobility. Here, we find a significant and good correlation between the structural rigidity of a well-characterized set of 16 mutants of lipase A from *Bacillus subtilis* (*Bs*LipA) and their thermodynamic thermostability. We apply the rigidity theory-based Constraint Network Analysis (CNA) approach, analyzing directly and in a time-independent manner the statics of the *Bs*LipA mutants. We carefully validate the CNA results on macroscopic and microscopic experimental observables and probe for their sensitivity with respect to input structures. Furthermore, we introduce a robust, local stability measure for predicting thermodynamic thermostability. Our results complement work that showed for pairs of homologous proteins that raising the structural stability is the most common way to obtain a higher thermostability. Furthermore, they demonstrate that related series of mutants with only a small number of mutations can be successfully analyzed by CNA, which suggests that CNA can be applied prospectively in rational protein design aimed at higher thermodynamic thermostability.

Author summary

Protein stability is relevant for biological function, molecular evolution, and biotechnological applications. Hence, understanding the origin of protein thermostability is of exceptional interest. One of the factors frequently associated with elevated thermostability is an increase in the mechanical stability of the native, folded protein. However, opposing cases have also been reported. We applied an approach originating from structural engineering to analyze the statics of a set of 16 mutants of the biotechnologically important lipase A from *Bacillus subtilis*. Within this closely related series, we observe a good correlation between the mechanical stability and the thermostability of mutants that results from making the native, folded state more favorable over the unfolded state. This result complements earlier work that showed for many pairs of homologous proteins from mesophilic and thermophilic organisms that raising the structural stability is the most common way to obtain a higher thermostability. Our results also suggest that the applied approach can be used prospectively in rational protein design aimed at higher thermodynamic thermostability.

Introduction

Sufficiently high thermostability of proteins is important for both organisms living in high temperature environments and for biotechnological applications where enzymes are used as biocatalysts under often harsh reaction conditions [1,2]. From a mechanistic point of view, "protein thermostability" embraces at least two different meanings [3,4]: (1) *thermodynamic* thermostability describes the folded-unfolded equilibrium of a protein, and (2) *kinetic* thermostability refers to the length of time a protein remains active before undergoing irreversible denaturation at an elevated temperature. Several factors have been frequently attributed to elevated protein thermostability including improved hydrogen bonding [5], ion pair and salt bridge networks [6], better hydrophobic packing [7], shortened loops [8], and higher secondary structure content [9], in all favoring an increased structural rigidity of the folded state [10-13]. As an opposing view, proteins from thermophilic organisms have been reported to be as flexible as or even more flexible than homologs from mesophilic organisms [14-17].

These different views on the relation between protein thermostability and structural rigidity have been a matter of ongoing discussion [10,18-23]. In particular, it has been argued that atomic movements, which are the primary mobility data from which information on protein statics (rigidity and flexibility) is derived, cover a wide range of timescales within a protein [15,24,25]. Hence, depending on the temporal resolution of the experimental technique or computational analysis used to detect such movements, (parts of) a protein can come out as rigid or flexible [26-32]. Here, we address the question of the relation between protein thermostability and structural rigidity by analyzing *directly* the static properties of a well-characterized set of 16 mutants of lipase A from *Bacillus subtilis* (*Bs*LipA). We do so by applying the rigidity theory-based Constraint Network Analysis (CNA) approach developed by us [33-35], thereby considering the *Bs*LipA variants to be in static equilibrium. Therefore, the rigidity and flexibility characteristics derived that way are time-independent.

*Bs*LipA is an important member of the lipase class of enzymes and used in diverse biotechnological applications [36,37]. Owing to its importance, *Bs*LipA has been extensively studied with respect to structure [38-41] and thermostability [42-48]. As to the latter, Reetz *et al.* applied iterative saturation mutagenesis on the most flexible amino acids as identified by crystallographic B-factors, which resulted in *Bs*LipA mutants that were more thermostable than the wild type showing an increase in T_{50}^{60} (the temperature required to reduce the initial enzymatic activity by 50% within 60 min) of ≤ 45 K [42]. Subsequent biophysical characterization of the three most thermostable mutants revealed that the improved activity retention resulted from a reduced rate of protein unfolding and a reduced precipitation of the unfolding intermediates, i.e., due to kinetic reasons [49]. In contrast, Rao *et al.* sequentially developed several thermostable *Bs*LipA mutants using directed evolution assisted by structural information. These mutants were shown to be more thermostable than the wild type due to predominantly thermodynamic reasons [44-48,50]; the most thermostable mutant displayed an increase in the melting temperature T_m of ~22 K.

In the CNA approach, a protein is modeled as a constraint network where bodies (representing atoms) are connected by sets of bars (constraints, representing covalent and noncovalent interactions) [51]. A rigidity analysis performed on the network [52,53] results in a decomposition into rigid parts and flexible links in between. By analyzing a series of "perturbed" networks in which noncovalent interactions are included in a temperature-dependent manner [11,13,54], the thermal unfolding of a protein is simulated [12,13,54]. Results of these analyses can be linked to biologically relevant characteristics of a biomolecular structure by a set of global and local indices [55]. In particular, a phase transition point T_p can be identified during the thermal unfolding simulation at which a largely rigid network becomes almost flexible; this phase transition point has been related to the thermodynamic thermostability of a protein [11-13]. For improving the robustness of the analyses, the rigidity analyses are performed on ensembles of network topologies (ENT^{FNC}) [56]. That way, thermal fluctuations of a protein are considered without actually sampling conformations.

The main outcome of this work is the finding of a significant and good correlation between the structural rigidity of all *Bs*LipA variants and their thermodynamic thermostability. On the way, we carefully probed for the sensitivity of the results with respect to the input structures and developed an approach for detecting outliers based on differences in the pathways of thermal unfolding. We furthermore introduced a local stability measure for predicting thermodynamic thermostability, which complements the detection of the (global) phase transition point T_p. As the *Bs*LipA variants are sequentially closely related, these results have important implications for applying CNA in a prospective manner in rational protein design aimed at higher thermodynamic thermostability. Finally, we discuss our results in terms of potentially different mechanisms underlying the increased protein thermostabilities of mutants isolated by Reetz *et al.*and Rao *et al.*

Materials and methods

Data set

The wild type structure of *Bs*LipA with the highest resolution (PDB ID: 1ISP; resolution = 1.3 Å) was obtained from the Protein Data Bank (PDB; www.pdb.org) [57]. For probing the sensitivity of the CNA results on the conformation of the input structures, five additional crystal structures of wild type *Bs*LipA were analyzed (PDB IDs: 1I6W, 1R4Z, 1R50, 2QXT, 2QXU). We included in our study all mutants from Rao *et al.* for which T_m values were determined [44-48]. In addition, we included the three most thermostable mutants developed in the last rounds of iterative saturation mutagenesis by Reetz *et al.* [42]. Models of mutant structures for which crystal structures were not available in the PDB were generated with the SCWRL program [58], using the respective *Bs*LipA structure as a template that is closest in sequence to the mutant. SCWRL constructs mutant models by predicting backbone-dependent side chain conformations with the help of a rotamer library; coordinates of backbone atoms remain unchanged. Conformations of side chains of all

residues within 8 Å of a mutated residue were re-predicted in order to allow for a local structural relaxation. For all structures, hydrogen atoms were added using REDUCE [59]; side chains of Asn, Gln, and His were flipped in this stage if necessary to optimize the hydrogen bond network. All water molecules, buffer ions, and crystal solvents were removed from the structures. Finally, all structures were minimized by 5000 steps of conjugate gradient minimization (including an initial steepest descent minimization for 100 steps) or until the root mean-square gradient of the energy was $< 1.0 \cdot 10^{-4}$ kcal mol^{-1} Å^{-1}. The energy minimization was carried out with Amber11 [60] using the Cornell *et al.* force field [61] with modifications for proteins (ff99SB) [62] and the GB^{OBC} generalized Born model [63]. All variants of *Bs*LipA used in this study are summarized in Table 1.

Construction of the constraint network and rigidity analysis

As described in the previous section, only the protein part was considered for network construction, i.e., all non-protein molecules including water molecules were discarded. This was done based on previous findings that including water molecules does not significantly change the rigidity analysis results [64,65]. Proteins were modeled as constraint networks in a *body-and-bar* representation (see section "Body-and-bar networks" in the File S1) [66,67] using the CNA software [35] that acts as a front- and back-end to the Floppy Inclusion and Rigid Substructure Topography (FIRST) program [51]. Once the constraint network is built, rigidity analysis is carried out, which identifies (rigid) clusters of atoms with no internal motion and flexible links in between, using the pebble game algorithm [52,53] as implemented in the FIRST software [51].

Thermal unfolding simulation

By sequentially removing non-covalent constraints from a network, one can simulate a loss of structural rigidity due to a temperature rise. Specifically, hydrogen bonds were removed from the network in increasing order of their strength following the idea that stronger hydrogen bonds break at higher temperatures than weaker ones [69]. As such, only hydrogen bonds with an energy $E_{HB} \leq E_{cut}(\sigma)$ were included in the network of state σ. A thermal unfolding trajectory of 60 network states was generated for each input network by decreasing E_{cut} from −0.1 kcal mol^{-1} to −6.0 kcal mol^{-1} with a step size of 0.1 kcal mol^{-1}. According to the linear relationship between E_{cut} and the temperature T introduced by Radestock and Gohlke [12,13], the range of E_{cut} used in this study is equivalent to increasing the temperature of the system from 302 K to 420 K with a step size of 2 K. Because hydrophobic interactions remain constant or become even stronger as the temperature increases [70,71], the number of hydrophobic tethers were kept unchanged throughout the thermal unfolding simulation. Rigidity analysis was performed on all such generated network states, and then local and global rigidity characteristics were calculated (see section "Local and global rigidity indices" in the File S1). The setup of the thermal unfolding simulation and the subsequent rigidity analysis were performed using the CNA software [35], which is available from

http://cpclab.uni-duesseldorf.de/software. A web service for performing CNA analysis can be accessed via http://cpclab.uni-duesseldorf.de/cna [34].

Ensemble of networks generated by using fuzzy noncovalent constraints

For improving the robustness of rigidity analyses, CNA is generally carried out on an ensemble of structures (e.g., generated by molecular dynamics (MD) simulations), and then results are averaged [11,64]. The preceding MD simulation compromises the efficiency of the rigidity analysis, however. To overcome this drawback, Pfleger *et al.* [56] recently introduced an approach that performs rigidity analyses on an ensemble of network topologies (ENT^{FNC}) generated from a single input structure by using fuzzy noncovalent constraints. Here, the number and distribution of non-covalent constraints (hydrogen bonds and hydrophobic tethers) are modulated by random components within certain ranges, thus simulating thermal fluctuations of a biomacromolecule without actually moving atoms. An ensemble of 2000 network configurations was generated using these definitions of fuzzy noncovalent constraints for all *Bs*LipA variants, respectively. Finally, average local indices were calculated, as were average phase transition temperatures identified by the global index cluster configuration entropy H_{type2}. The index H_{type2} monitors the degree of disorder in the realization of a given network state σ: As long as a network is dominated by a very large rigid cluster, H_{type2} tends to be low because there are only a few configurations of a system with a large rigid cluster possible; H_{type2} increases when larger rigid clusters break down in smaller clusters (see section "Local and global rigidity indices" in the File S1 and ref. [55] for details).

Clustering of unfolding pathways

Recently, we showed that curves of the rigidity order parameter, which characterizes the general percolation behavior of a constraint network during thermal unfolding, for mesophilic proteins and their thermophilic counterparts are almost identical except for a shift of the curve of the thermophilic protein to higher temperatures [12]. This finding supported the hypothesis of corresponding states according to which mesophilic and thermophilic enzymes are in corresponding states of similar rigidity and flexibility at their respective optimal temperature [12]. The percolation index p_i is a local analog to the rigidity order parameter. It monitors for each bond when it segregates from the largest rigid cluster present at the beginning of a thermal unfolding simulation (see section "Local and global rigidity indices" in the File S1 and ref. [55] for details). That way, a residue-wise p_i profile of a protein, generated by taking the lower of the p_i values of the two backbone bonds for each residue, expresses the hierarchical break-down of the largest rigid cluster during a thermal unfolding simulation.

We thus reasoned that the (dis)similarity of unfolding pathways of *Bs*LipA variants can be measured by Manhattan distances between their respective p_i profiles. We used this distance measure for clustering the network topologies of all *Bs*LipA variants into 10 clusters using

the Partitioning Around Medoids algorithm [72] as implemented in the R program (http://www.r-project.org). This optimal number of clusters was chosen based on monitoring the change in the objective function of the clustering (the mean of the dissimilarities of all objects to their nearest medoids) as a function of the number of clusters (Figure S1 in the File S1) and visual inspection of cluster medoids for their dissimilarity to other medoids (residue-wise p_i profiles for medoids of the 10 clusters are shown in Figure S2 in the File S1). A clustering in more than 10 clusters essentially created additional clusters that were very similar to other clusters. From this, the cluster distribution (frequencies of network topologies in each of the 10 clusters out of in total 2000 network topologies) for each *Bs*LipA variant was calculated by counting the number of networks that belongs to each of the 10 clusters. A high (low) correlation between cluster distributions for two *Bs*LipA variants then indicates that both variants unfold in a similar (different) manner. Finally, a matrix of all pairwise correlations of cluster distributions of *Bs*LipA variants was generated.

Results

Data set

*Bs*LipA is a protein of 181 amino acids with a minimal α/β hydrolase fold; in this fold, a central parallel β-sheet of six β-strands is surrounded by six α-helices. Ser77, Asp133, and His156 constitute the catalytic triad (Fig. 1). Unlike other lipases, the catalytic site in *Bs*LipA is not covered with a lid. Hence, *Bs*LipA does not show interfacial activation [40]. The data set used in this study contains structures of the wild type *Bs*LipA, thirteen mutants from Rao *et al.* [44-48], and three mutants from Reetz *et al.* [42,49] (Table 1). The mutants differ from the wild type by three to twelve mutations, i.e., the sequence identity is > 93%. Models for the mutants for which X-ray structures were not available were built using the SCWRL program. As the number of mutations in the modeled variants is ≤ 7 with respect to the template structures (< 4% with respect to the sequence length) (Table 1), an overall similar backbone confirmation can be expected as can be an overall reliable modeling of side chain conformations by SCWRL. This was also evident from a very good structural alignment and low root-mean-square deviations (RMSD) between the wild type and those mutants for which crystal structures were available (C_α atom-based RMSD values between the wild type and the mutants < 0.38 Å). The high structural similarity allows a direct comparison of results from rigidity analyses for these structures [11-13].

The melting temperature T_m of the wild type is 329.15 K. The T_m values of the mutants of Rao *et al*. range from 334.35 to 351.35 K (Table 1). For the mutants of Reetz *et al.* no T_m values are available. Rather, unfolding initiation temperatures T_i were reported, which are lower by 2.5 to 6.2 K than that of the wild type. This suggests that mutants of Reetz *et al.* are thermodynamically less thermostable than the wild type [49], in contrast to mutants from Rao *et al.* [44-48]. However, we note that, while T_m reports on the temperature at which 50% of the protein is unfolded and, hence, properly describes the folded-unfolded equilibrium of a

protein, T_i only reports on the temperature at which the unfolding transition begins. Therefore, we will only consider relations within mutants of Rao *et al.* and to the wild type and distinguish those from relations within mutants of Reetz *et al.* and to the wild type. Finally, the T_{50}^{t} values of the mutants of Reetz *et al.* are higher than that of the wild type (Table 1), showing that these mutants more efficiently refold upon cooling after incubation at high temperatures than does the wild type. The location of mutations in all of the mutants investigated in this study is shown in Fig. 1; all mutations are located on the protein surface.

--- Table 1 ---

--- Fig. 1 ---

Thermal unfolding pathway of *Bs*LipA

From monitoring the loss in rigidity percolation during thermal unfolding simulations, major phase transitions in the protein can be identified that relate to the unfolding pathway [11-13,54,73]. Here, we describe the loss of rigidity percolation of the wild type *Bs*LipA (PDB ID 1ISP) as an example. Similarity or dissimilarity, respectively, of the unfolding pathways across all variants is described below. During the thermal unfolding, a giant rigid cluster that exists at low temperature (equivalent to a high E_{cut}) breaks down in smaller sub-clusters until, finally, the whole protein becomes flexible at a high temperature (Fig. 2; see also Video S1 showing the loss of rigidity percolation during the thermal unfolding of the wild type). As such, nearly the entire protein structure constitutes a single giant rigid cluster initially (at 302 K; Fig. 2). As the temperature increases, loops segregate first from the giant rigid cluster. Then, at 314 K, α-helix D (αD) and αE segregate to form individual small rigid clusters (Fig. 2), as do αA and αF at 318 K. The giant rigid cluster at this temperature is formed by the central β-sheet region and the two helices αB and αC (Fig. 2). Next, the β-sheet region becomes sequentially flexible, beginning with β4 and β8 at 320 K (Fig. 2). Then, the remaining β-strands become flexible in the order β3, β7, and β5−β6, leading to a completely flexible β-sheet region at 332 K (Fig. 2). The immediate next step at which αB and αC become two separate rigid clusters is identified as a phase transition point: Now most of the structure has become flexible. This transition is most prominent with respect to going from a structurally stable wild type *Bs*LipA to an unfolded one (Figure S3 in the File S1). After this phase transition point, the remaining rigidity is sequentially lost, and the structure finally becomes completely flexible at 374 K (Fig. 2).

During the thermal unfolding of *Bs*LipA, helices segregate from the giant rigid cluster as independent small rigid clusters. This is due to two reasons: First, in the *body-and-bar* network representation, a helix with a minimum of seven amino acids is already rigid by itself due to constraints arising from covalent and backbone hydrogen bonds [66]. Second, with the current energy function E_{HB} [69], all backbone hydrogen bonds are assigned a very similar strength, irrespective of their location along a helix. Thus, a helix will persist as an independent rigid cluster during the thermal unfolding simulation until all backbone

hydrogen bonds break almost simultaneously at a high temperature, which most likely represents an overstabilization of a helix [74]. Considering this behavior, the unfolding pathway identified for the wild type *Bs*LipA is in good agreement with respect to the early segregation of α-helices with experimental findings on the unfolding of proteins with an α/β hydrolase fold [75,76]. This indicates that side chain-mediated interactions between amino acids are well represented by the applied definitions of non-covalent constraints in the network. This is important as we want to detect effects of changes in such interactions due to mutations.

--- Fig. 2 ---

Prediction of thermodynamic thermostability of *Bs*LipA variants based on the global index H_{type2}

From the thermal unfolding simulations, the temperature of the phase transition point T_p was identified as described in the section "Local and global rigidity indices" in the File S1. Note that T_p values determined that way should be considered relative values only, as stated in previous studies [12,34,35]. Initially, we calculated phase transition points using single network topologies generated from the input structures of wild type *Bs*LipA and mutants of Rao *et al.*; however, this resulted in a very poor prediction of thermodynamic thermostability with a coefficient of determination (R^2) for a linear fit between experimental T_m and predicted T_p of 0.22 (data not shown). We anticipated that this result reflects the high sensitivity of CNA on the conformation of the input structures as also found previously [11,56,64,65]. We thus resorted to averaging T_p values over an ensemble of *Bs*LipA, applying the recently developed ENT^{FNC} approach. This approach generates an ensemble of network topologies from a single input structure and has been shown to yield results of rigidity analyses both at the local and global level that agree almost perfectly with those obtained from MD simulations-generated ensembles of structures [56]. However, this yielded a significant ($p = 0.002$) correlation between T_p and T_m with $R^2 = 0.58$ only if the two structures with the lowest (wild type) and highest (mutant 6B) T_m were considered outliers (Fig. 3A; see below for an explanation regarding the outliers; note that removing the two outliers in the case of using single network topologies only marginally improved R^2 from 0.22 to 0.29). The mutants IX, X and XI of Reetz *et al.* were predicted to be slightly less thermostable than the wild type (Fig. 3A). This is in line with experimental findings by Reetz *et al.* that suggest that these mutants are thermodynamically less stable than the wild type [49]. In summary, these results demonstrate that CNA coupled with the ENT^{FNC} approach can sense effects on the thermodynamic thermostability that arise from only a few sequence variations (pairwise sequence identity > 93%; pairwise RMSD < 0.38 Å). However, the false predictions for wild type *Bs*LipA and mutant 6B are dissatisfying.

--- Fig. 3 --

Difference in unfolding pathways explains outliers

Next, we investigated why the thermostabilities of the wild type and the mutant 6B were predicted falsely. Since the precision of the computations shown in Fig. 3A is high (the standard error in the mean is < 0.38 K in all cases), we reasoned that the false prediction must arise from a systematic difference between the wild type and 6B *versus* all other mutants of Rao *et al.* Thus, we mutually compared all unfolding pathways of the systems as described in "Materials and methods". After partitioning unfolding pathways of *Bs*LipA variants characterized on a residue basis by the percolation index p_i into 10 clusters (see Figure S2 in the File S1 for the p_i profiles of the 10 cluster medoids), we calculated correlation coefficients from the resulting cluster distributions for all pairs of variants (Fig. 4; Table S1 and S2 in the File S1).

These results revealed that the wild type enzyme shows an unfolding pathway distribution very distinct from other *Bs*LipA variants from Rao *et al.* with correlation coefficients r ranging from −0.69 to 0.54 (Fig. 4, Table S1 in the File S1). The average r value for the wild type against all other variants from Rao *et al.* is -0.06 ± 0.14 (mean ± SEM), which is lower than that of the other variants (≥ 0.16 except for the outlier 6B) (Table S1 in the File S1). The second outlier, mutant 6B, has an average r value of 0.12 ± 0.16 when comparing its unfolding pathway distribution to those of other variants from Rao *et al.* This average r value is lower than the corresponding average r values of all other mutants from Rao *et al.* (Table S1 in the File S1). The thermal unfolding pathway of 6B is shown in Figure S4 in the File S1. While the overall unfolding pathway of 6B is comparable with that of the wild type *Bs*LipA in that the helices segregate from the giant rigid cluster as individual rigid clusters in the early phase of unfolding, they do so in a different order (αD, αA−αF, αE, αB−αC; Figure S4 in the File S1). A probability density function (PDF) of r values of unfolding pathway distributions of the two outliers wild type and mutant 6B with all other variants shows a bimodal distribution and is shifted towards lower r values compared to the PDF of the r values of other mutants from Rao *et al.* Furthermore, about half of this distribution is related to negative r values (Figure S5 in the File S1). In all, this suggests that the two outliers have unfolding pathways different from all other mutants from Rao *et al.* for which the prediction of thermodynamic thermostability was successful. Finally, we note that the unfolding pathway distributions of the wild type and the three mutants from Reetz *et al.* are highly similar to each other ($r > 0.79$; $p \leq 0.001$; Table S2 in the File S1).

These findings have important implications: First, the results strongly suggest that the misprediction of the thermostabilities of the wild type and mutant 6B arises from them showing different unfolding pathways from all of the remaining mutants from Rao *et al.*. Apparently, the present approach of identifying phase transition points by monitoring the *global* index H_{type2} (see section "Local and global rigidity indices" in the File S1) is too sensitive with respect to the details of such pathways. Consequently, alternative methods should be explored (see section "Median stability of rigid contacts between residue neighbors as a new measure for predicting thermodynamic thermostability"). Second, the results

suggest that the history of the generation of the *Bs*LipA structures may play a role for the observed differences in the unfolding pathways: generally, the most similar unfolding pathways (Table S1 and Table S2 in the File S1) (and then the most coherent T_p predictions) are found for those variants that originate from a common structural "ancestor" (Table 1). Third, the results propose to apply the similarity/dissimilarity of unfolding pathway distributions as a measure to judge the reliability of thermostability predictions in future studies: the lower the similarity for two variants, the less confident should one be that relative thermostability predictions are correct. Finally, we cannot exclude at the present stage that thermostabilizing mutations lead to an unfolding pathway that is different from the one of the wild type. Considering that intrinsic and extrinsic modifications in other systems that led to thermostabilization have been shown to influence not just the folded state but the entire (un)folding free energy landscape [77,78], this possibility also exists for *Bs*LipA mutants [45,47].

Median stability of rigid contacts between residue neighbors as a new measure for predicting thermodynamic thermostability

The above findings called for predicting the thermodynamic thermostability in a way that is less sensitive to the details of the unfolding pathway than the present approach relying on the *global* index H_{type2}. The sensitivity arises here from the need to accurately identify the phase transition point from the percolation behavior of the constraint network as the most pronounced jump in H_{type2} during the unfolding (Figure S3 in the File S1). As shown previously, however, the percolation behavior of networks from protein structures is complex [13] (in contrast to that of network glasses [54,79]), reflecting that a protein structure is hierarchical and composed of modules. As a consequence, often more than one pronounced jump in H_{type2} is observed, which then makes it difficult to assign a phase transition point (Figure S3 in the File S1).

As an alternative, we set out to characterize thermodynamic thermostability at the *local* level [55], i.e., by monitoring residue pair-wise descriptors of local stability within a protein structure as a function of the temperature. The most comprehensive information in that direction is provided by stability maps [12], which depict when a rigid contact between two residues ceases to exist along a thermal unfolding trajectory. As such, a stability map denotes the distribution of flexibility and rigidity within the system, identifies regions that are flexibly or rigidly correlated across the structure, and provides information how these properties change during thermal unfolding [12,55]. To stress the locality of interactions within a protein, we focused on the stability of rigid contacts between structurally close residues only (i.e., those residues where at least one pair of respective atoms is within 5 Å distance). From this neighbor stability map, the median stability of rigid contacts $\tilde{rc}_{ij,neighbor}$ is defined as a new measure for predicting thermodynamic thermostability. With the ENT^{FNC} approach used here, $\tilde{rc}_{ij,neighbor}$ were finally averaged over the entire ensemble of 2000 constraint networks.

A significant and fair correlation of $\tilde{rc}_{ij,neighbor}$ with T_m values of the thermodynamically stable mutants from Rao *et al.* is obtained (R^2 = 0.46, p = 0.004; Fig. 3B). No outlier is observed now, indicating that our definition of an average local stability correctly reflects differences in the thermodynamic thermostability. As before, the mutants from Reetz *et al.* are found to have a lower thermodynamic thermostability than the wild type, in very good agreement with experimental findings (see above and Table 1) [49]. The $\tilde{rc}_{ij,neighbor}$-based measure is apparently less sensitive to differences in the unfolding pathway because the wild type and mutant 6B are now much better ranked. However, comparing the prediction of thermostabilities by $\tilde{rc}_{ij,neighbor}$ and H_{type2}, the latter yields a better correlation with T_m for mutants with similar unfolding pathways. From an application point of view, we thus recommend using H_{type2}-derived T_p values for comparing thermostabilities of variants of a protein unless the underlying unfolding pathways are dissimilar; in that case, we recommend using $\tilde{rc}_{ij,neighbor}$.

When applied to hen egg white lysozyme the ENT^{FNC} approach has been shown to significantly improve the robustness of rigidity analyses with respect to the conformation of the input structures [56]. To probe if this also holds for *Bs*LipA investigated here, we computed $\tilde{rc}_{ij,neighbor}$ using the ENT^{FNC} approach for five additional crystal structures of wild type *Bs*LipA (see section "Materials and methods"). The standard error of the mean in $\tilde{rc}_{ij,neighbor}$ over all six wild type *Bs*LipA structures is 0.57 K (Fig. 3B) including PDB ID 1ISP discussed so far. This error is likely within the experimental uncertainty, confirming our previous results of robust rigidity analyses with ENT^{FNC} [56]. Still, if the average $\tilde{rc}_{ij,neighbor}$ over all six crystal structures (315.9 K; see horizontal line in Fig. 3B; Table 1) is considered for the $\tilde{rc}_{ij,neighbor}$ *versus* T_m correlation, the quality of the correlation improves considerably to R^2 = 0.55 (p = 0.001) compared to if only $\tilde{rc}_{ij,neighbor}$ of PDB ID 1ISP is used (see above). This indicates that the use of multiple input structures in connection with the ENT^{FNC} approach further increases the accuracy of thermostability predictions.

Influence of mutations on local structural rigidity

Considering that the average local stability defined above correctly reflects differences in the (macroscopic) thermodynamic thermostability, we analyzed on a residue basis how changes in thermostability relate to changes in local structural stability (rigidity). First, we compared stability maps of variants from Rao *et al.* with distinct thermostabilities to analyze the effect of mutations on the local rigidity. In particular, we compared the wild type to a more thermostable variant 1-14F5 and the most thermostable variant 6B. We averaged stability maps of the six wild type structures (see above and Fig. 3B) and used this average for comparison against the thermostable variants of *Bs*LipA. Difference stability maps for 1-14F5/wild type (Fig. 5A) and 6B/wild type (Fig. 5B) pairs demonstrate that mutations in general improve the strength of rigid contacts to and in between neighboring residues of the mutations (lower triangles in Fig. 5A and B) but also in between residue pairs not in contact

distance (upper triangles in Fig. 5A and B). This effect is more pronounced for 6B/wild type than 1-14F5/wild type.

In more detail, the four mutations (indicated by arrows in Fig. 5A and shown in Fig. 5D) on 1-14F5 stabilize contacts of αD with its neighboring helix αC and contacts of αA with αF (Fig. 5A, D). More importantly, the contacts of helices αA and αF with their neighboring β-strands in the central β-sheet region are stabilized, which delays the early loss of these helices observed during the thermal unfolding of the wild type (Fig. 2). Similarly, the contacts between αB and the central β-sheet region also become stronger, which delays the decay of structural stability of the β-sheet during thermal unfolding. On average, contacts between all residue neighbors are ~−0.1 kcal mol^{-1} or ~2 K more stable in 1-14F5 than in the wild type.

Residues mutated in 6B (indicated by arrows in Fig. 5B and shown in Fig. 5E) include the mutations already found in 1-14F5. This explains a strengthening of inter-helical contacts and of the contacts between α helices and the central β-sheet region as discussed already for 1-14F5 (Fig. 5D, E). However, the additional mutations in 6B stabilize contacts between other α-helices (αD and αE) and the central β-sheet region and further reinforce those between αA or αF and the β-sheet. On average, contacts between all residue neighbors are ~−0.4 kcal mol^{-1} or ~8 K more stable in 6B than in the wild type (Fig. 5E).

Taken together, contacts between peripheral helices and the central β-sheet region are stronger in 6B than in 1-14F5. This delays the loss of α-helices during thermal unfolding (Fig. 2) to a larger extent in 6B than in 1-14F5, explaining at a structural level why 6B is more stable than 1-14F5. Remarkably, many of these stabilizations must arise from the long-range aspect of rigidity percolation [52,64,80,81], because almost all mutations in 6B are on the surface, i.e., far from the central β-sheet region. In contrast, inter-helical contacts of the αB/αC helix pair become weaker in the mutants than in wild type (Fig. 5D, E) indicating that the strengthened stability between these helices and the central β-sheet region is sufficient to keep the structure folded. At last, for all other thermodynamically more thermostable mutants, a similar profile of changes in contact stability between various secondary structure elements was observed (Figure S6 in the File S1). Not unexpected, the increase in contact stability compared to wild type (Figure S6 in the File S1) was generally the more pronounced the higher the thermodynamic thermostability is of the mutant (Table 1).

---Fig. 5---

Second, we compared the mutants from Reetz *et al.* to the wild type. Regarding mutant X, seven residues have been mutated (indicated by arrows in Fig. 5C and shown in Fig. 5G). In stark contrast to what was observed for the thermodynamically thermostabilized mutants, this mutant showed a destabilization of rigid contacts both locally and globally (Fig. 5C and F; see also Figure S7 in the File S1, where a similar finding is depicted for mutants IX and XI). For mutant X, the average decrease in stability over all residue neighbors is ~0.06 kcal mol^{-1} or ~1.2 K. The destabilization found on the local scale agrees with results of a lower T_p found

when analyzing the mutants globally. Furthermore, the results are in line with experimental findings which suggest that the mutants are thermodynamically less stable than the wild type (Table 1) [49]. Our findings are also in good agreement with results obtained by comparative crystal structure analysis of wild type and variant X [49]: Loop region 14-21, for which lower B-factors in X than in the wild type structure were observed, shows increased contact stabilities with its neighboring residues in X (Fig. 5C and F; Figure S8 in the SI). Likewise, regions 129-153 and 177-181, for which higher B-factors in X than in the wild type structure were observed, show decreased contact stabilities with their neighboring residues in X (Fig. 5C and F; Figure S8 in the SI). However, region 60-70 shows increased contact stabilities in X (Fig. 5C, F and Figure S8 in the SI) despite higher B-factors observed in the comparative crystal structure analysis. The latter may reflect increased motions of a stabilized region as a whole, taking into consideration that B-factors can report on rigid body motions of a structurally stable part [82].

Discussion

Understanding the origin of thermostability is of fundamental importance in protein biochemistry. Here, we have probed the relation between protein thermostability and structural rigidity by directly analyzing static properties of a well-characterized set of 16 *Bs*LipA mutants. The main outcome of this work is the finding of a good correlation between the structural rigidity of all *Bs*LipA variants and their thermodynamic thermostability. This finding of a quantitative relation between structural rigidity and thermodynamic thermostability within a series of closely related protein variants complements a previous study that showed for pairs of homologous proteins from thermophilic and mesophilc organisms that raising the structural stability is the most common way (~77% of all cases) to obtain a higher thermostability [84].

Intense discussions are ongoing regarding the question if elevated protein thermostability is related to increased or decreased structural rigidity of the folded state [10,18-23]. Part of this discussion is related to how information on structural rigidity is derived from information on mobility, in particular with respect to the temporal resolution of the experimental techniques and computational analysis [26-32]. In this context, the finding we describe here is highly relevant. As the rigidity theory-based CNA approach applied characterizes rigidity and flexibility of proteins directly, i.e., without the requirement of information on atomic movements, it does not suffer from such time dependence. Another part of the discussion is related to the fact that changes in the enthalpy, entropy and/or heat capacity can lead to thermodynamic stabilization; these changes can be linked to distinct effects on the structural stability of the folded state [19]. It was thus instructive to observe that the general increase in rigidity in the mutants of Rao *et al.* is accompanied by certain inter-helical contacts becoming weaker than in the wild type; these weakened contacts between the “modular” helices may increase the entropy of the folded state and so may further contribute to the overall stability of the systems [17,84,85]. This finding again calls attention to analyzing the origin of

thermostability with methods that cover a wide range of temporal and spatial resolution because otherwise one effect may be hidden beneath another.

Our results are backed up with a careful validation of the accuracy and robustness of the CNA approach on the data set both from a macroscopic and microscopic point of view. As to the former, good and statistically significant correlations between experimental melting temperatures (T_m) of mutants of Rao *et al.* and predicted thermodynamic thermostabilities have been found based on two independent measures (H_{type2} and $\widetilde{rc}_{ij,neighbor}$), as was correctly predicted that the thermodynamic thermostability of the mutants of Reetz *et al.* is lower than that of the wild type. Furthermore, $\widetilde{rc}_{ij,neighbor}$-based predictions of the thermodynamic thermostability on six crystal structures of wild type *Bs*LipA revealed a standard error of the mean likely within experimental error, confirming previous results of robust rigidity analyses when applying the ENT^{FNC} approach [56]. As to the latter, the detailed analysis of the unfolding pathway of wild type *Bs*LipA revealed a good agreement with respect to the early segregation of α-helices with experimental observations on other proteins with an α/β hydrolase fold. These findings are in line with previous successful applications of CNA in predicting melting temperatures and identifying structural weak spots [11-13].

From a methodological point of view, some additional comments are in order. First, in the present study we successfully predicted the thermodynamic thermostability for mutants that differ by as few as three to twelve mutations from the wild type. Compared to previous applications of CNA on either pairs of mesophilic and thermophilic homologues [12,13] or a series of homologous proteins from different organisms living at varying temperatures [11], this finding considerably broadens the application domain of CNA towards data-driven protein engineering: There, related series of mutants with only a small number of respective mutations will be the major focus of investigations. Second, we introduced a measure for the similarity/dissimilarity of unfolding pathways of mutants and used it for explaining false thermostability predictions. This suggests to use the measure in future studies as a significance criterion to judge the reliability of thermostability predictions from CNA. Third, we introduced the median stability of rigid contacts as a new local measure for predicting thermodynamic thermostability and showed that this measure is less sensitive to details of the unfolding pathway. The measure is thus recommended for comparing thermostabilities of mutants the underlying unfolding pathways of which are dissimilar.

Finally, regarding the subset of mutants of Reetz *et al.*, we find a decreased local rigidity compared to wild type, in line with findings of lower unfolding initiation temperatures, yet the mutants are more "thermostable" than the wild type in that they preserve enzymatic activity better after subjecting them to higher temperatures [42]. It would have been tempting to investigate how this relates to a potential kinetic stabilization of the mutants. However, we refrained from doing so due to the lack of direct experimental evidence for such a kinetic stabilization [49]. In turn, this finding draws attention to the fact that the term "protein thermostability" is often used in a non-discriminating sense, i.e., data reported in the

literature does not allow to establish whether a protein is thermodynamically or kinetically stable [49]. This adds another layer of complexity to the question of the relation between protein thermostability and structural rigidity as it may be required to decouple observations on "increased *vs*. decreased structural rigidity" from the general description of "protein thermostability" in future studies.

Acknowledgements

We are grateful to the Ministry of Innovation, Science, and Research of North Rhine-Westphalia and Heinrich-Heine-University Düsseldorf (HHU) for a scholarship to PCR within the CLIB^{2021} Graduate Cluster Industrial Biotechnology. We acknowledge the „Zentrum fuer Informations- und Medientechnologie" (ZIM) at HHU for computational support. We thank Anuseema Bhadauriya, HHU, for fruitful discussions on the experimental data of the investigated mutants.

Figure legends

Fig. 1. Cartoon representation of wild type *Bs*LipA with mutated residues indicated by spheres of their C_α atoms (mutations from Rao *et al*. [44-48]: magenta; Reetz *et al.* [42,49]: orange; mutations common in both data sets: cyan). The catalytic triad (Ser77-Asp133-His156) is shown in stick representation with yellow carbons. The protein is colored according to secondary structure (α-helices: red; β-sheets: yellow; loops: green). The right view (**B**) differs from the left (**A**) by an anti-clockwise rotation of ~90° about a horizontal axis. All figures of *Bs*LipA structures were generated with PyMOL (http://www.pymol.org).

Fig. 2. Average loss of structural rigidity of the wild type *Bs*LipA during a thermal unfolding simulation. Rigid clusters are depicted as uniformly colored bodies, with the largest rigid cluster shown in blue and smaller rigid clusters in the order of the colors green, magenta, cyan, orange, and violet. Temperatures are indicated for each depiction of a rigid cluster decomposition. At the beginning of the thermal unfolding simulation (302 K), almost the complete structure is part of the giant rigid cluster; in contrast, the structure becomes completely flexible at temperatures ≥ 374 K. The right views differ from the left ones by an anti-clockwise rotation of ~90° about a horizontal axis. Important secondary structure elements are labeled. Note that the unfolding pathway shown here represents an *average loss of rigidity percolation* calculated from a stability map (see section "Local and global rigidity indices" in the File S1) averaged over all unfolding trajectories obtained for the ensemble of 2000 network topologies. Hence, the temperature at the phase transition point identified that way (Figure S3) cannot be compared to the *average phase transition temperature*, which is obtained from 2000 individual T_p values and used for predicting the thermodynamic thermostability of *Bs*LipA variants (see section "Prediction of thermodynamic thermostability of *BsLipA* variants").

Fig. 3. Correlation between predicted and experimental thermostabilities (T_m values) of *Bs*LipA variants; for the predictions, the ENT^{FNC} approach was used. **A**: Correlation between T_p derived from the global index H_{type2} and T_m values for thermodynamically thermostabilized mutants from Rao *et al*. Data points colored red were considered outliers (see main text for explanation) and excluded when calculating R^2 values and the correlation lines. **B**: Correlation between $\widetilde{rc}_{ij,neighbor}$ and T_m values for thermodynamically thermostabilized mutants from Rao *et al*. Data points shown as empty squares represent $\widetilde{rc}_{ij,neighbor}$ values for five additional wild type crystal structures (see main text for details; two of the squares closely overlap; mean $\widetilde{rc}_{ij,neighbor}$ over all six data points for wild type structures is shown as a small horizontal line: 315.9 ± 0.6 K). A and B: Error bars represent the standard error in the mean. T_p and $\widetilde{rc}_{ij,neighbor}$ values for kinetically thermostabilized mutants from Reetz *et al*. are marked by arrows on the corresponding ordinates.

Fig. 4. Pairwise correlations of cluster distributions (using 10 clusters) of unfolding pathways of wild type *Bs*LipA and mutants from Rao *et al*. The upper triangle shows pairwise

correlation coefficients as dial plots where a filled portion of a pie indicates the magnitude of the correlation (*r*) and blue (red) color indicates a positive (negative) correlation. The lower triangle shows 68% data ellipses (depicting the bivariate mean ± 1 standard deviation) [86] and scatterplots of the respective cluster distributions (frequencies of network topologies in each of the 10 clusters) of the two *Bs*LipA variants as red lines smoothed by locally-weighted polynomial regression [87]. Axes for the plots in the lower triangle are omitted for clarity. The figure was generated using the "corrgram" package [88] of the R program (http://www.r-project.org).

Fig. 5. Differences in the stability of rigid contacts between wild type and mutants of *Bs*LipA. Maps depict differences between stability maps of the respective mutants and an average stability map of the six wild type structures (see the main text for explanation) for **A**: mutant 1-14F5, **B**: mutant 6B, and **C**: mutant X. A red (blue) color indicates that a rigid contact in the mutant is more (less) stable than in the wild type (see color scale at the bottom). The upper triangles show differences in the stability values for all residue pairs; the lower triangles show differences in the stability values only for residue pairs that are within 5 Å of each other, with values for all other residue pairs colored gray. Secondary structure elements as computed by the DSSP program [89,90] are indicated on both abscissa and ordinate and are labeled: α-helix (red rectangle), β-strands (green rectangle), loop (black line). Arrows represent the mutation positions with respect to the wild type sequence: Common mutations in 1-14F5 (A) and 6B (B) are shown in magenta, unique mutations in 6B (B) are shown in green, and mutations in X (C) are shown in orange. The differences in the stability of rigid contacts for residue neighbors is also displayed on the structures of the mutants by sticks connecting C_α atoms of residue pairs colored according to the color scale of the maps for **D**: 1-14F5, **E**: 6B, and **F**: X. Only those contacts that are stabilized by ≥ 4 K or destabilized by ≥ 3 K are shown for clarity; for the same reason, contacts between two residues of the same secondary structure element are not shown. Mutated residues are shown as sticks and a sphere at their C_α atoms (D, E, and F) in the same color used for arrows (A, B, and C).

Tables

Table 1: Summary of *Bs*LipA variants used in the study.

***Bs*LipA variant**[a]	**PDB ID**[b]	**Resolution**[c]	**Mutations**	T_m **(K)**	T_i **(K)** [d]	T_{50} **(K)**	$\widetilde{rc}_{ij,neighbor}$ **(K)** [e]	**Reference**
Wild type	1ISP	1.3	–	329.15	324.95	321.15[f]	317.4 (315.9)[h]	[39,42,44]
IX	1ISP*	–	K112D, M134D, Y139C, I157M	–	318.75	335.95[g]	313.5	[42,49]
X	1ISP*	–	R33Q, D34N, K35D, K112D, M134D, Y139C,	–	321.65	362.15[f]	314.5	[42,49]
XI	1ISP*	–	R33G, K112D, M134D, Y139C, I157M	–	322.45	366.15[f]	314.9	[42,49]
TM	1T2N	1.8	L114P, A132D, N166Y	334.35	–	–	317.6	[44]
1-14F5	1T2N*	-	TM + N89Y	336.15	–	–	319.5	[44]
1-17A4	3D2A	1.73	TM + I157M	336.55	–	–	319.9	[44]
1-8D5	1T2N*	–	TM + F17S	337.55	–	–	316.3	[44]
2D9	3D2B	1.95	TM + F17S, N89Y, I157M	340.55	–	–	319.7	[44]
3-18G4	3D2B*	–	2D9 + G111D	341.55	–	–	318.5	[44]
3-11G1	3D2B*	–	2D9 + A20E	341.75	–	–	319.6	[44]
3-3A9	3D2B*	–	2D9 + A15S	341.85	–	–	317.5	[44]
4D3	3D2C	2.18	2D9 + A15S, A20E,G111D	344.35	–	–	323.6	[44]
5-D	3D2C*	–	4D3 + S163P	345.35	–	–	320.0	[45]
5-A	3D2C*	–	4D3 + M134E	346.05	–	–	319.4	[45]
5-B	3D2C*	–	4D3 + M137P	347.25	–	–	320.5	[45]
6B	3QMM	1.89	4D3 + M134E, M137P, S163P	351.35	–	–	324.0	[47]

[a] Names of *Bs*LipA structures are taken from the respective references.
[b] A PDB ID marked with an asterisk indicates that the model of the corresponding variant was built using the structure with that PDB ID as a template.
[c] In Å.
[d] The temperature at which the unfolding transition begins.
[e] Median stability of rigid contacts between residue neighbors computed by applying the ENT^{FNC} approach (see section "Median stability of rigid contacts between residue neighbors as a new measure for predicting thermodynamic thermostability").
[f] T_{50}^{60} values, i.e., the temperature required to reduce the initial enzymatic activity by 50% within 60 min.
[g] T_{50}^{15} values, i.e., the temperature required to reduce the initial enzymatic activity by 50% within 15 min.
[h] Average $\widetilde{rc}_{ij,neighbor}$ over six wild type structures (see the main text for details).

Figures

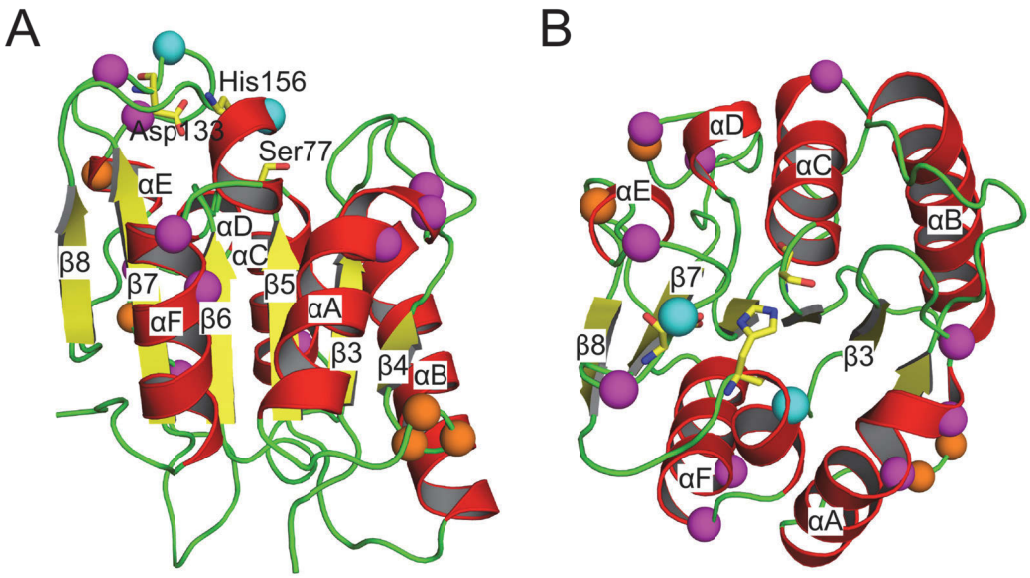


Fig. 1

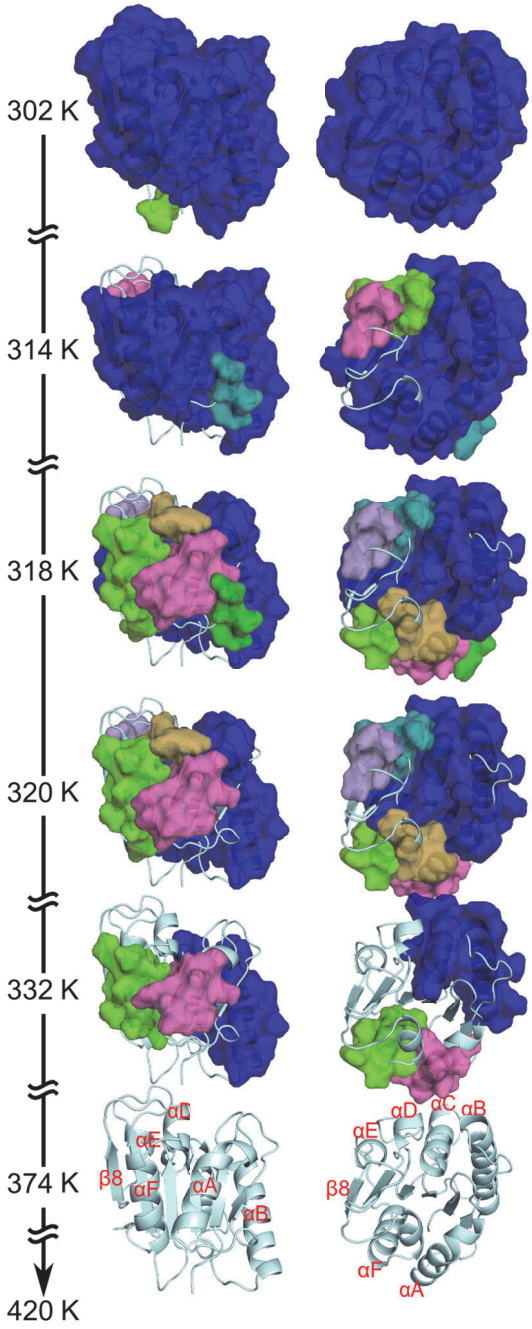

302 K
314 K
318 K
320 K
332 K
374 K
420 K
αD
αE
β8
αF
αA
αB
αD
αC
αB
αE
β8
αF
αA

Fig. 2

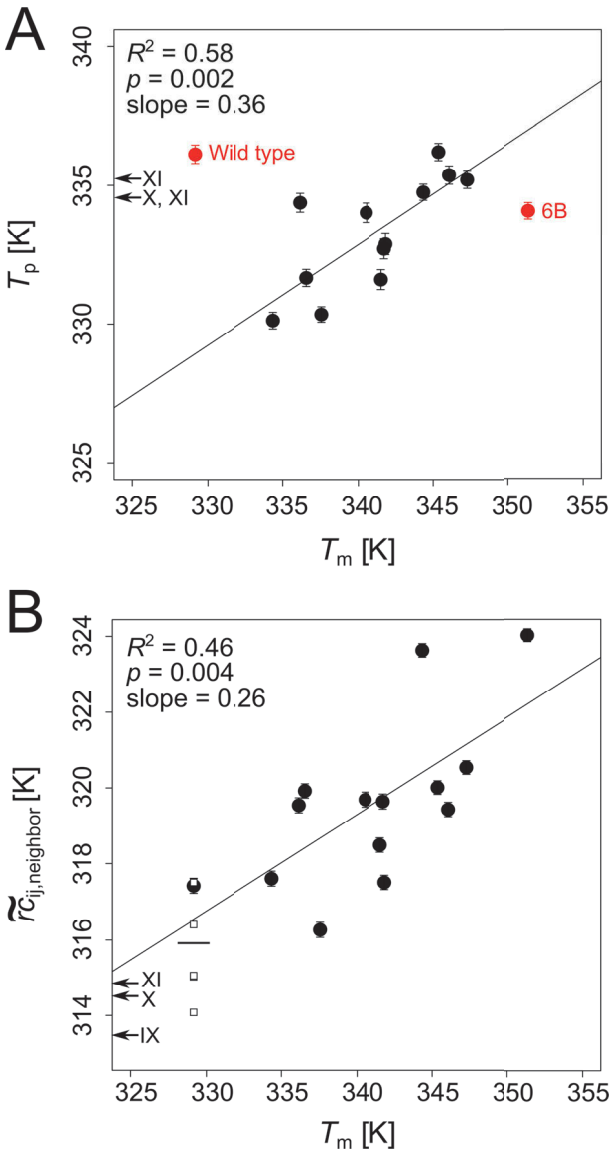

A
R^2 = 0.58
p = 0.002
slope = 0.36
Wild type
6B
XI
X, XI
T_p [K]
T_m [K]
B
R^2 = 0.46
p = 0.004
slope = 0.26
XI
X
IX
$\tilde{T}_{ij,neighbor}$ [K]
T_m [K]

Fig. 3.

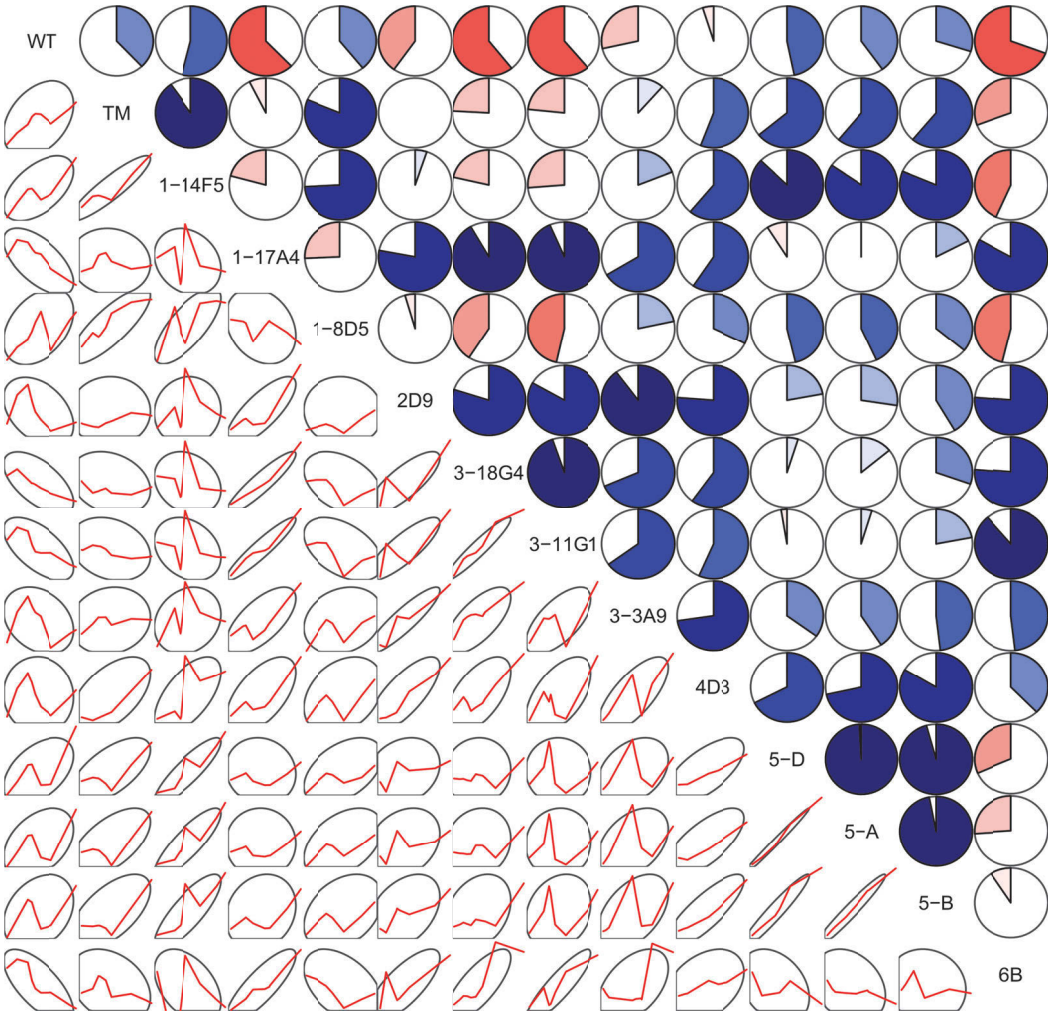


Fig. 4

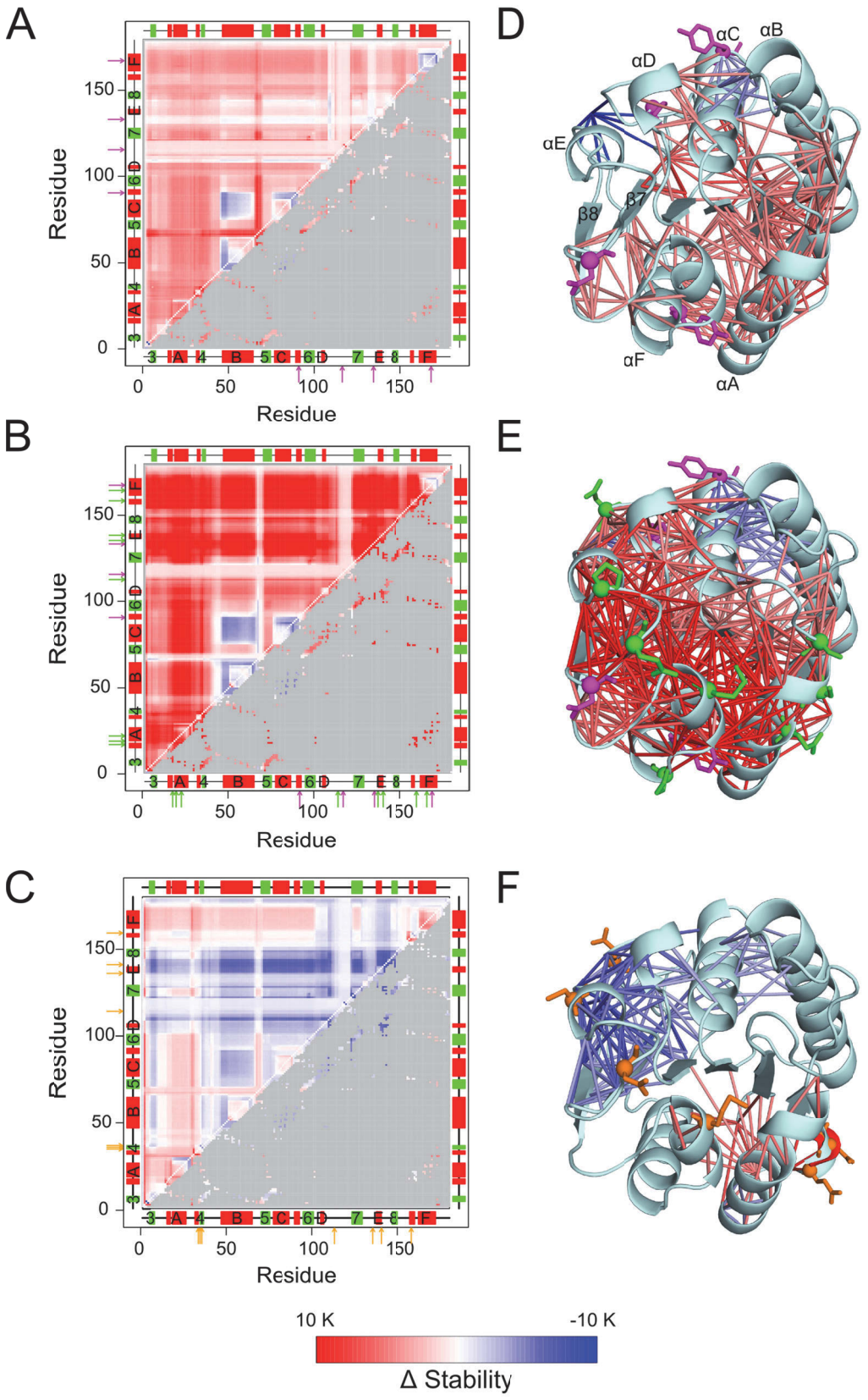

A
B
C
D
E
F
Residue
0
50
100
150
αA
αB
αC
αD
αE
αF
β7
β8
10 K
-10 K
Δ Stability

Fig. 5

References

1. Demirjian DC, Moris-Varas F, Cassidy CS (2001) Enzymes from extremophiles. Curr Opin Chem Biol 5: 144-151.
2. Van den Burg B (2003) Extremophiles as a source for novel enzymes. Curr Opin Microbiol 6: 213-218.
3. Ó'Fágáin C (2011) Engineering protein stability. In: Walls D, Loughran ST, editors. Protein Chromatography: Springer. pp. 103-136.
4. Polizzi KM, Bommarius AS, Broering JM, Chaparro-Riggers JF (2007) Stability of biocatalysts. Curr Opin Chem Biol 11: 220-225.
5. Vogt G, Woell S, Argos P (1997) Protein thermal stability, hydrogen bonds, and ion pairs. J Mol Biol 269: 631-643.
6. Kumar S, Tsai CJ, Nussinov R (2000) Factors enhancing protein thermostability. Protein Eng 13: 179-191.
7. Gromiha MM, Pathak MC, Saraboji K, Ortlund EA, Gaucher EA (2013) Hydrophobic environment is a key factor for the stability of thermophilic proteins. Proteins: Struct, Funct, Bioinf 81: 715-721.
8. Russell RJ, Hough DW, Danson MJ, Taylor GL (1994) The crystal structure of citrate synthase from the thermophilic archaeon, Thermoplasma acidophilum. Structure 2: 1157-1167.
9. Querol E, PerezPons JA, MozoVillarias A (1996) Analysis of protein conformational characteristics related to thermostability. Protein Eng 9: 265-271.
10. Vihinen M (1987) Relationship of protein flexibility to thermostability. Protein Eng 1: 477-480.
11. Rathi PC, Radestock S, Gohlke H (2012) Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. J Biotechnol 159: 135-144.
12. Radestock S, Gohlke H (2011) Protein rigidity and thermophilic adaptation. Proteins: Struct, Funct, Bioinf 79: 1089-1108.
13. Radestock S, Gohlke H (2008) Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. Eng Life Sci 8: 507-522.
14. Hernandez G, LeMaster DM (2001) Reduced temperature dependence of collective conformational opening in a hyperthermophile rubredoxin. Biochemistry 40: 14384-14391.
15. Hernandez G, Jenney FE, Adams MWW, LeMaster DM (2000) Millisecond time scale conformational flexibility in a hyperthermophile protein at ambient temperature. Proc Natl Acad Sci U S A 97: 3166-3170.
16. Fitter J, Heberle J (2000) Structural equilibrium fluctuations in mesophilic and thermophilic alpha-amylase. Biophys J 79: 1629-1636.
17. Danciulescu C, Ladenstein R, Nilsson L (2007) Dynamic arrangement of ion pairs and individual contributions to the thermal stability of the cofactor-binding domain of glutamate dehydrogenase from Thermotoga maritima. Biochemistry 46: 8537-8549.

18. Jaenicke R (2000) Do ultrastable proteins from hyperthermophiles have high or low conformational rigidity? Proceedings of the National Academy of Sciences 97: 2962-2964.
19. Jaenicke R, Böhm G (1998) The stability of proteins in extreme environments. Curr Opin Struct Biol 8: 738-748.
20. Kalimeri M, Rahaman O, Melchionna S, Sterpone F (2013) How Conformational Flexibility Stabilizes the Hyperthermophilic Elongation Factor G-Domain. J Phys Chem B 117: 13775-13785.
21. Basu S, Sen S (2013) Do Homologous Thermophilic-Mesophilic Proteins Exhibit Similar Structures and Dynamics at Optimal Growth Temperatures? A Molecular Dynamics Simulation Study. J Chem Inf Model 53: 423-434.
22. Oyeyemi OA, Sours KM, Lee T, Kohen A, Resing KA, et al. (2011) Comparative Hydrogen-Deuterium Exchange for a Mesophilic vs Thermophilic Dihydrofolate Reductase at 25 degrees C: Identification of a Single Active Site Region with Enhanced Flexibility in the Mesophilic Protein. Biochemistry 50: 8251-8260.
23. Marcos E, Jimenez A, Crehuet R (2012) Dynamic Fingerprints of Protein Thermostability Revealed by Long Molecular Dynamics. J Chem Theory Comput 8: 1129-1142.
24. Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. Nature 450: 964-972.
25. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, et al. (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature 450: 913-916.
26. Ishima R, Torchia DA (2000) Protein dynamics from NMR. Nat Struct Mol Biol 7: 740-743.
27. Englander SW, Kallenbach NR (1983) Hydrogen exchange and structural dynamics of proteins and nucleic acids. Q Rev Biophys 16: 521-655.
28. Weiss S (1999) Fluorescence spectroscopy of single biomolecules. Science 283: 1676-1683.
29. Zhang XJ, Wozniak JA, Matthews BW (1995) Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozymes. J Mol Biol 250: 527-552.
30. Frank J, Agrawal RK (2000) A ratchet-like inter-subunit reorganization of the ribosome during translocation. Nature 406: 318-322.
31. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. Nat Struct Mol Biol 9: 646-652.
32. Case DA (1994) Normal mode analysis of protein dynamics. Curr Opin Struct Biol 4: 285-290.
33. Rathi PC, Pfleger C, Fulle S, Klein DL, Gohlke H (2011) Statics of biomacromolecules. In: Comba P, editor. Molecular Modeling. Weinheim: Wiley-VCH. pp. 281-299.
34. Kruger DM, Rathi PC, Pfleger C, Gohlke H (2013) CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo-)stability, and function. Nucleic Acids Res 41: W340-W348.
35. Pfleger C, Rathi PC, Klein DL, Radestock S, Gohlke H (2013) Constraint Network Analysis (CNA): A Python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. J Chem Inf Model 53: 1007-1015.

36. Jaeger KE, Eggert T (2002) Lipases for biotechnology. Curr Opin Biotechnol 13: 390-397.
37. Jaeger KE, Ransac S, Dijkstra BW, Colson C, Vanheuvel M, et al. (1994) Bacterial Lipases. FEMS Microbiol Rev 15: 29-63.
38. Droge MJ, Boersma YL, van Pouderoyen G, Vrenken TE, Ruggeberg CJ, et al. (2006) Directed evolution of Bacillus subtilis lipase A by use of enantiomeric phosphonate inhibitors: crystal structures and phage display selection. Chembiochem 7: 149-157.
39. Kawasaki K, Kondo H, Suzuki M, Ohgiya S, Tsuda S (2002) Alternate conformations observed in catalytic serine of Bacillus subtilis lipase determined at 1.3 A resolution. Acta Crystallogr D Biol Crystallogr 58: 1168-1174.
40. van Pouderoyen G, Eggert T, Jaeger KE, Dijkstra BW (2001) The crystal structure of Bacillus subtilis lipase: a minimal alpha/beta hydrolase fold enzyme. J Mol Biol 309: 215-226.
41. Rajakumara E, Acharya P, Ahmad S, Sankaranaryanan R, Rao NM (2008) Structural basis for the remarkable stability of Bacillus subtilis lipase (Lip A) at low pH. Bba-Proteins Proteom 1784: 302-311.
42. Reetz MT, Carballeira JD, Vogel A (2006) Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. Angew Chem Int Ed Engl 45: 7745-7751.
43. Abraham T, Pack SP, Yoo YJ (2005) Stabilization of Bacillus subtilis Lipase A by increasing the residual packing. Biocatal Biotransfor 23: 217-224.
44. Ahmad S, Kamal MZ, Sankaranarayanan R, Rao NM (2008) Thermostable Bacillus subtilis lipases: In vitro evolution and structural insight. J Mol Biol 381: 324-340.
45. Ahmad S, Rao NM (2009) Thermally denatured state determines refolding in lipase: Mutational analysis. Protein Sci 18: 1183-1196.
46. Acharya P, Rajakumara E, Sankaranarayanan R, Rao NM (2004) Structural basis of selection and thermostability of laboratory evolved Bacillus subtilis lipase. J Mol Biol 341: 1271-1281.
47. Kamal MZ, Ahmad S, Molugu TR, Vijayalakshmi A, Deshmukh MV, et al. (2011) In Vitro Evolved Non-Aggregating and Thermostable Lipase: Structural and Thermodynamic Investigation. J Mol Biol 413: 726-741.
48. Kamal MZ, Ahmad S, Yedavalli P, Rao NM (2010) Stability curves of laboratory evolved thermostable mutants of a Bacillus subtilis lipase. Bba-Proteins Proteom 1804: 1850-1856.
49. Augustyniak W, Brzezinska AA, Pijning T, Wienk H, Boelens R, et al. (2012) Biophysical characterization of mutants of Bacillus subtilis lipase evolved for thermostability: factors contributing to increased activity retention. Protein Sci 21: 487-497.
50. Kamal MZ, Mohammad TAS, Krishnamoorthy G, Rao NM (2012) Role of Active Site Rigidity in Activity: MD Simulation and Fluorescence Study on a Lipase Mutant. PLoS ONE 7: e35188.
51. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF (2001) Protein flexibility predictions using graph theory. Proteins: Struct, Funct, Bioinf 44: 150-165.

52. Jacobs DJ, Thorpe MF (1995) Generic rigidity percolation: the pebble game. Phys Rev Lett 75: 4051-4054.
53. Jacobs DJ, Hendrickson B (1997) An algorithm for two-dimensional rigidity percolation: the pebble game. J Comp Phys 137: 346-365.
54. Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF (2002) Protein unfolding: rigidity lost. Proc Natl Acad Sci U S A 99: 3540-3545.
55. Pfleger C, Radestock S, Schmidt E, Gohlke H (2013) Global and local indices for characterizing biomolecular flexibility and rigidity. J Comp Chem 34: 220-233.
56. Pfleger C, Gohke H (2013) Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. Structure 21: 1725-1734.
57. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235-242.
58. Krivov GG, Shapovalov MV, Dunbrack RL (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins: Struct, Funct, Bioinf 77: 778-795.
59. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation1. J Mol Biol 285: 1735-1747.
60. Case DA, Cheatham III TE, Darden T, Gohlke H, Luo R, et al. (2005) The Amber biomolecular simulation programs. J Comp Chem 26: 1668-1688.
61. Wang JM, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J Comp Chem 21: 1049-1074.
62. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, et al. (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins: Struct, Funct, Bioinf 65: 712-725.
63. Onufriev A, Bashford D, Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins: Struct, Funct, Bioinf 55: 383-394.
64. Gohlke H, Kuhn LA, Case DA (2004) Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. Proteins: Struct, Funct, Bioinf 56: 322-337.
65. Mamonova T, Hespenheide B, Straub R, Thorpe MF, Kurnikova M (2005) Protein flexibility using constraints from molecular dynamics simulations. Phys Biol 2: S137-S147.
66. Whiteley W (2005) Counting out to the flexibility of molecules. Phys Biol 2: S116-S126.
67. Hespenheide BM, Jacobs DJ, Thorpe MF (2004) Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. J Phys Condens Matter 16: S5055-S5064.
68. FIRST, a program for analysing flexibility of networks. http://flexweb.asu.edu/ (accessed January 17, 2014).
69. Dahiyat BI, Gordon DB, Mayo SL (1997) Automated design of the surface positions of protein helices. Protein Sci 6: 1333-1337.

70. Folch B, Rooman M, Dehouck Y (2008) Thermostability of salt bridges versus hydrophobic interactions in proteins probed by statistical potentials. J Chem Inf Model 48: 119-127.
71. Privalov PL, Gill SJ (1988) Stability of protein structure and hydrophobic interaction. Adv Protein Chem 39: 191-234.
72. Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ (2006) Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. J Math Model Algorithms 5: 475-504.
73. Hespenheide BM, Rader AJ, Thorpe MF, Kuhn LA (2002) Identifying protein folding cores from the evolution of flexible regions during unfolding. J Mol Graph Model 21: 195-207.
74. Graf J, Nguyen PH, Stock G, Schwalbe H (2007) Structure and Dynamics of the Homologous Series of Alanine Peptides: A Joint Molecular Dynamics/NMR Study. J Am Chem Soc 129: 1179-1189.
75. Beermann B, Guddorf J, Boehm K, Albers A, Kolkenbrock S, et al. (2007) Stability, unfolding, and structural changes of cofactor-free 1H-3-hydroxy-4-oxoquinaldine 2,4-dioxygenase. Biochemistry 46: 4241-4249.
76. Hung HC, Chang GG (2001) Multiple unfolding intermediates of human placental alkaline phosphatase in equilibrium urea denaturation. Biophys J 81: 3456-3471.
77. Sapra KT, Balasubramanian GP, Labudde D, Bowie JU, Muller DJ (2008) Point mutations in membrane proteins reshape energy landscape and populate different unfolding pathways. J Mol Biol 376: 1076-1090.
78. Kayatekin C, Zitzewitz JA, Matthews CR (2008) Zinc binding modulates the entire folding free energy surface of human Cu,Zn superoxide dismutase. J Mol Biol 384: 540-555.
79. Thorpe MF, Jacobs DJ, Chubynsky MV, Phillips JC (2000) Self-organization in network glasses. J Non Cryst Solids 266-269: 859-866.
80. Yilmaz LS, Atilgan AR (2000) Identifying the adaptive mechanism in globular proteins: Fluctuations in densely packed regions manipulate flexible parts. J Chem Phys 113: 4454-4464.
81. Rader AJ, Yennamalli RM, Harter AK, Sen TZ (2012) A rigid network of long-range contacts increases thermostability in a mutant endoglucanase. J Biomol Struct Dyn 30: 628-637.
82. Fulle S, Gohlke H (2008) Analyzing the flexibility of RNA structures by constraint counting. Biophys J 94: 4202-4219.
83. Razvi A, Scholtz JM (2006) Lessons in stability from thermophilic proteins. Protein Sci 15: 1569-1578.
84. Seewald MJ, Pichumani K, Stowell C, Tibbals BV, Regan L, et al. (2000) The role of backbone conformational heat capacity in protein stability: Temperature dependent dynamics of the B1 domain of Streptococcal protein G. Protein Sci 9: 1177-1193.
85. Stone MJ, Gupta S, Snyder N, Regan L (2001) Comparison of protein backbone entropy and beta-sheet stability: NMR-derived dynamics of protein G B1 domain mutants. J Am Chem Soc 123: 185-186.

86. Friendly M, Monette G, Fox J (2013) Elliptical insights: understanding statistical methods through elliptical geometry. Statistical Science 28: 1-39.
87. Cleveland WS (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. J Am Stat Assoc 74: 829-836.
88. Friendly M (2002) Corrgrams: Exploratory displays for correlation matrices. Am Stat 56: 316-324.
89. Kabsch W, Sander C (1983) Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. Biopolymers 22: 2577-2637.
90. Joosten RP, Beek TAHT, Krieger E, Hekkelman ML, Hooft RWW, et al. (2011) A series of PDB related databases for everyday needs. Nucleic Acids Res 39: D411-D419.

Publication VI – Supplementary Information

Structural rigidity and protein thermostability

Prakash Chandra Rathi, Karl-Erich Jaeger and Holger Gohlke
Submitted manuscript, **2014.**
Contribution: 75%

Supplemental Methods

Body-and-bar networks

In *body-and-bar* networks, atoms are considered bodies with six degrees of freedom, and each bar between two bodies removes one degree of freedom. Depending on the strength of an interaction between two atoms, a constraint can be modeled as any number of bars between one and six with six bars completely freezing the motion between two atoms [1,2]. As done previously [3-6], covalent bonds were modeled with five (single bonds) and six (peptide and double bonds) bars, whereas hydrophobic tethers and hydrogen bonds (including salt bridges; together referred to as hydrogen bonds here) were modeled with two and five bars, respectively. A modified version of the potential by Mayo and coworkers [7] as described in ref. [8] was used to calculate hydrogen bond energies E_{HB}; hydrogen bonds with energies lower than a certain cutoff E_{cut} were included in the network (see "Thermal unfolding simulation" section in the main text for details). Hydrophobic constraints were considered between pairs of carbon and/or sulfur atoms according to a Gaussian probability function depending on the distances between the atoms (d_{ij}), the sum of their van der Waals (d_{vdw}) radii (C: 1.7 Å; S: 1.8 Å), and the full width at half maximum D_{cut} (eq. S1; see ref. [3] for details).

$$p(d_{ij}) = e^{-\frac{1}{2}\left(\frac{\left(d_{ij}-d_{vdW}\right)^2}{D_{cut}^2}\right)^2} \tag{S1}$$

Local and global rigidity indices

From a thermal unfolding trajectory, one can calculate both residue-level (local) and overall (global) rigidity characteristics of a protein [9]. Local indices can be used to investigate specific questions regarding the stability and activity of a protein. In the present study, the stability map rc_{ij} introduced by us in ref. [5] was used to characterize the local rigidity of *Bs*LipA and to understand the influence of mutations. A stability map is derived by identifying "rigid contacts" between two residues i and j that are represented by their C_α atoms. A rigid contact exists if the two residues belong to the same rigid cluster. During a thermal unfolding simulation, stability maps are then constructed in that, for each residue pair, E_{cut} (or, equivalently, a temperature derived from the relationship $T = \mathrm{f}(E_{\mathrm{cut}})$ described in

refs. [4,5]) is identified at which a rigid contact between these residues is lost. In that respect, the stability map is a two-dimensional itemization of the local rigidity index as detailed in ref. [9]. When filtered such that only rigid contacts between residues that are 5 Å apart from each other (measured as the distance between the closest atom pair from the two residues) are considered, a neighbor stability map results. This map helps focusing on short-range residue contacts that can be directly modulated by mutagenesis with the aim to stabilize them for improving the overall stability of a protein.

In addition, the (local) percolation index p_i introduced by us in ref. [9] was used to characterize thermal unfolding pathways of the *Bs*LipA structures. The percolation index monitors the percolation behavior (i.e., the loss of rigidity when diluting the constraint network) of a biomolecule on a microscopic level and so allows identifying the hierarchical break-down of the giant percolating cluster during a thermal unfolding simulation. The giant percolating cluster is the largest rigid cluster present at the highest E_{cut} value (i.e., at the lowest temperature at the beginning of a thermal unfolding simulation) with all constraints in place. More technically, p_i monitors the E_{cut} at which a bond segregates from the giant percolating cluster during a thermal unfolding simulation. For a C_α atom-based representation, the lower of the p_i values of the two backbone bonds is considered.

Global indices help identifying phase transition temperatures T_p at which a network switches from being largely rigid to largely flexible. Previously, we showed that T_p identified by a modified cluster configuration entropy H_{type2} [4,9] can be used for predicting the thermodynamic thermostability of and identifying structural weak spots in a protein [4-6]. The cluster configuration entropy has originally been introduced by Andraud *et al.* [10] as a morphological descriptor for heterogeneous materials and is adapted from Shannon's information theory. H_{type2} monitors the degree of disorder in the realization of a given network state: As long as a network is dominated by a very large rigid cluster, H_{type2} tends to be low because there are only a few configurations of a system with a large rigid cluster possible; H_{type2} increases when larger rigid clusters break down in smaller clusters. The H_{type2} *versus* T curve obtained from a thermal unfolding simulation was fitted with a double sigmoid [11] as done previously [6], and the temperature T_p was identified as the inflection point of the sigmoid with the larger difference in the asymptote values. This way, in most cases, a late transition involving the final decay of the giant percolating cluster is identified as T_p [5].

Supplemental Tables

Table S1. Pairwise Pearson correlation coefficients *r* (upper triangle) and corresponding *p* values (lower triangle) between cluster distributions of unfolding pathways of wild type *Bs*LipA and mutants from Rao *et al*.

	Wild type	**TM**	**1-14F5**	**1-17A4**	**1-8D5**	**2D9**	**3-18G4**	**3-11G1**	**3-3A9**	**4D3**	**5-D**	**5-A**	**5-B**	**6B**
Wild type		0.38	0.54	-0.62	0.39	-0.40	-0.61	-0.61	-0.29	-0.05	0.47	0.40	0.30	-0.69
TM	0.280		0.90	-0.07	0.81	0.01	-0.24	-0.23	0.11	0.56	0.64	0.61	0.61	-0.31
1-14F5	0.108	< 0.001		-0.21	0.74	0.05	-0.21	-0.26	0.19	0.61	0.87	0.84	0.81	-0.43
1-17A4	0.055	0.839	0.565		-0.26	0.78	0.92	0.93	0.66	0.59	-0.09	-0.01	0.17	0.83
1-8D5	0.264	0.004	0.014	0.475		-0.04	-0.41	-0.46	0.22	0.32	0.46	0.43	0.36	-0.46
2D9	0.250	0.989	0.887	0.008	0.902		0.80	0.83	0.90	0.76	0.22	0.27	0.41	0.76
3-18G4	0.063	0.502	0.553	< 0.001	0.241	0.006		0.95	0.69	0.60	0.05	0.14	0.30	0.76
3-11G1	0.060	0.516	0.463	< 0.001	0.176	0.003	< 0.001		0.65	0.57	-0.03	0.05	0.22	0.89
3-3A9	0.425	0.752	0.600	0.037	0.549	< 0.001	0.029	0.042		0.73	0.35	0.41	0.48	0.48
4D3	0.894	0.094	0.061	0.070	0.366	0.010	0.067	0.088	0.017		0.68	0.72	0.83	0.38
5-D	0.173	0.046	0.001	0.805	0.179	0.539	0.892	0.943	0.322	0.032		0.99	0.96	-0.31
5-A	0.251	0.061	0.002	0.981	0.214	0.444	0.698	0.896	0.243	0.020	< 0.001		0.97	-0.26
5-B	0.402	0.060	0.004	0.631	0.309	0.234	0.394	0.541	0.159	0.003	< 0.001	< 0.001		-0.09
6B	0.026	0.388	0.209	0.003	0.178	0.010	0.010	0.001	0.159	0.282	0.375	0.468	0.802	
Mean *r*	-0.06	0.29	0.34	0.28	0.16	0.41	0.29	0.27	0.43	0.56	0.40	0.43	0.49	0.12
SEM[a]	0.14	0.12	0.14	0.15	0.12	0.12	0.15	0.16	0.09	0.06	0.12	0.11	0.09	0.16

[a] Standard error of the mean.

Table S2. Pairwise Pearson correlation coefficients *r* (upper triangle) and corresponding *p* values (lower triangle) between cluster distributions of unfolding pathways of wild type *Bs*LipA and mutants from Reetz *et al*.

	Wild type	**IX**	**X**	**XI**
Wild type		0.86	0.91	0.87
IX	0.001		0.87	0.79
X	< 0.001	0.001		0.97
XI	0.001	0.007	< 0.001	

Supplemental Figures

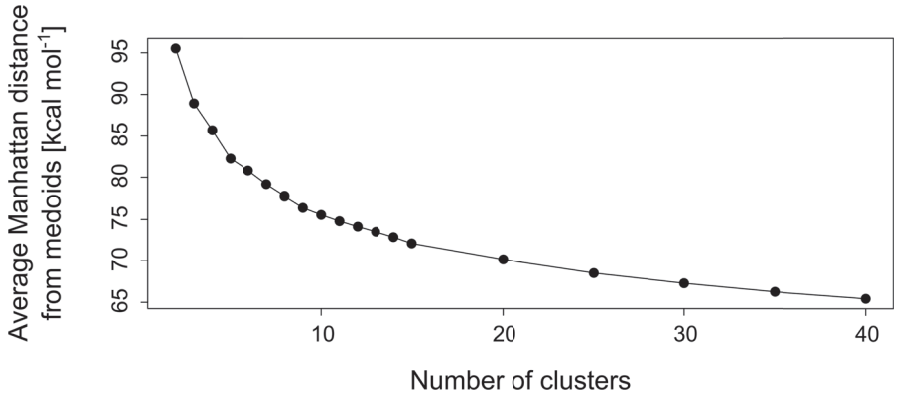


Figure S1. Objective function of the clustering (the mean of the dissimilarities of all objects to their nearest medoids) *vs*. the number of clusters of p_i profiles of all *Bs*LipA variants.

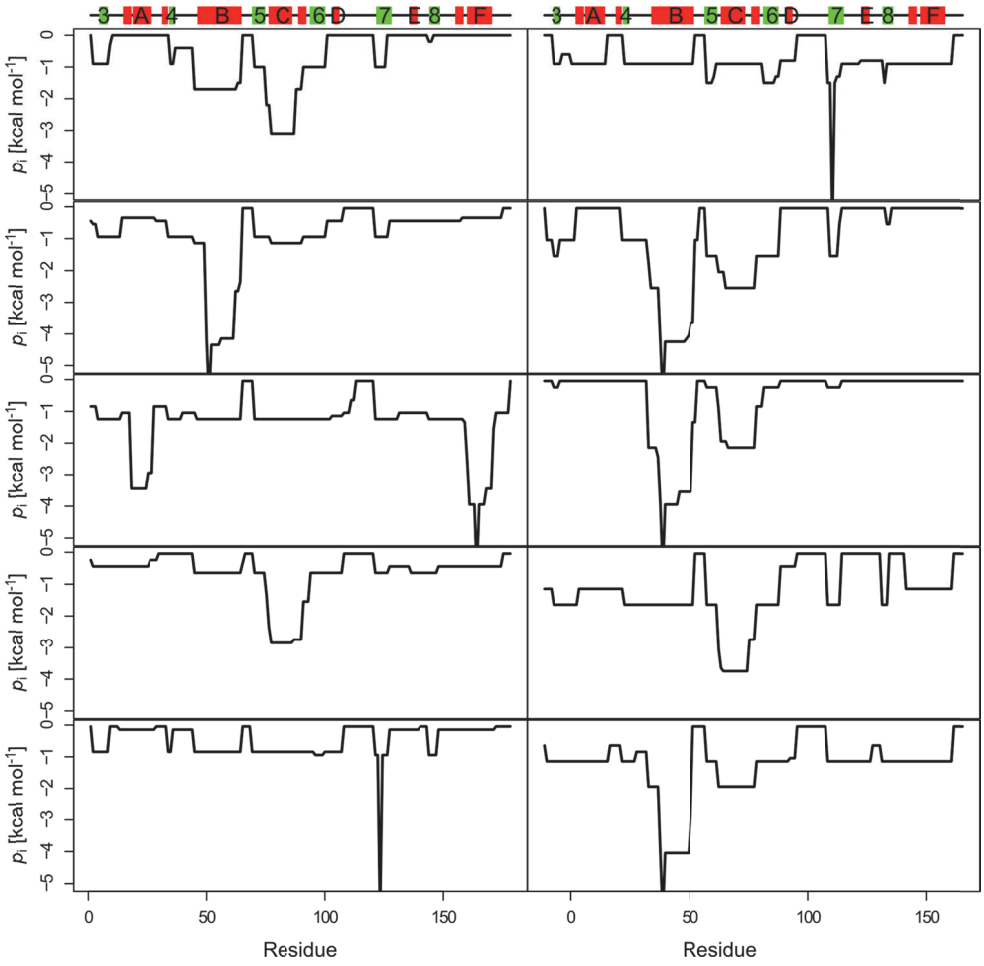


Figure S2. Residue-wise p_i plots for medoids of the 10 clusters. Secondary structure elements as computed by the DSSP program [12,13] are indicated on the top of the plots and are labeled: α-helix (red rectangle), β-strands (green rectangle), loop (black line).

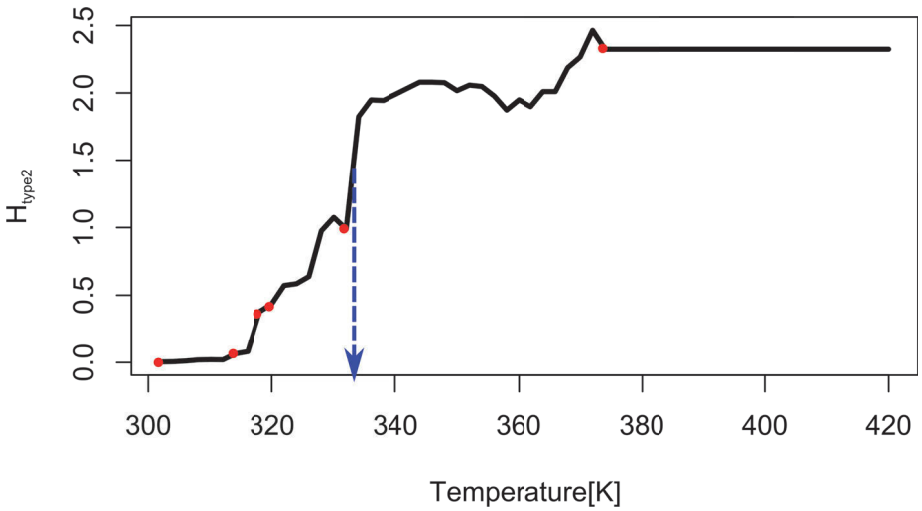


Figure S3. Cluster configuration entropy H_{type2} *vs*. temperature obtained from the *average loss of rigidity percolation* over the ensemble of 2000 network topologies of wild type *Bs*LipA; this average loss of rigidity percolation is calculated from a stability map averaged over all 2000 unfolding trajectories. Steps that involve a loss of secondary structure elements during the thermal unfolding (shown in Fig. 2 in the main text) are indicated with red points. The blue arrow indicates the phase transition point on the unfolding pathway.

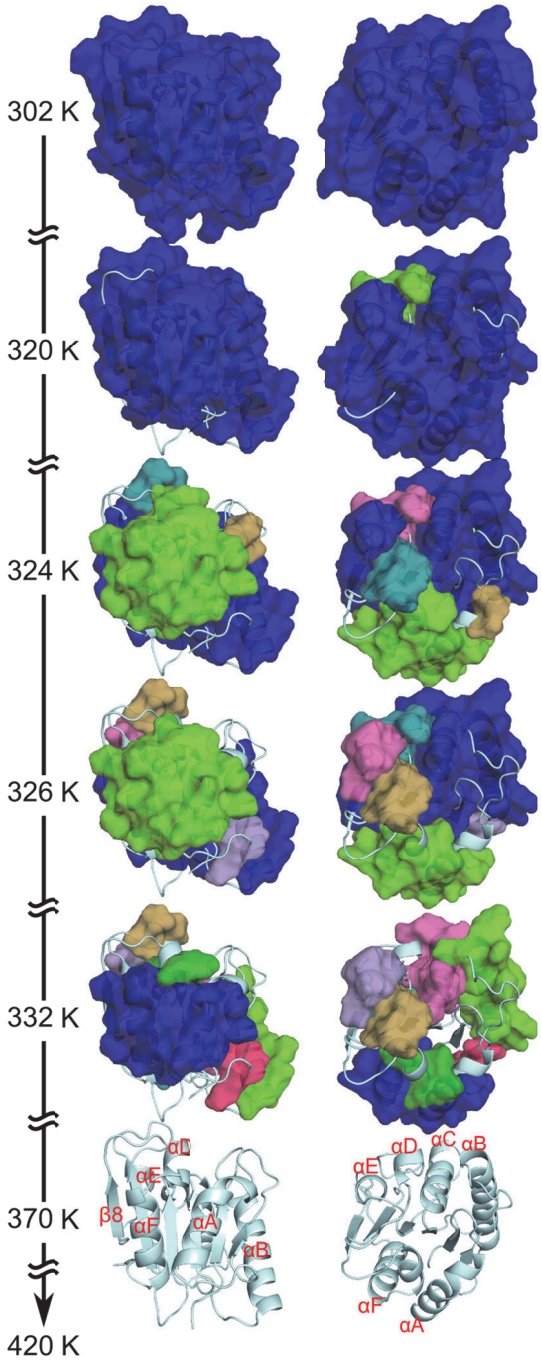


Figure S4. Average loss of structural rigidity of mutant 6B during a thermal unfolding simulation. Rigid clusters are depicted as uniformly colored bodies with the largest rigid cluster always shown in blue. Temperatures are indicated for each rigid cluster decomposition depiction. At the beginning of the thermal unfolding simulation (302 K) almost the complete structure is part of the giant rigid cluster; the structure becomes completely flexible at temperatures ≥ 370 K. The right views differ from the left ones by an anti-clockwise rotation of ~90° about a horizontal axis. Important secondary structure elements are labeled.

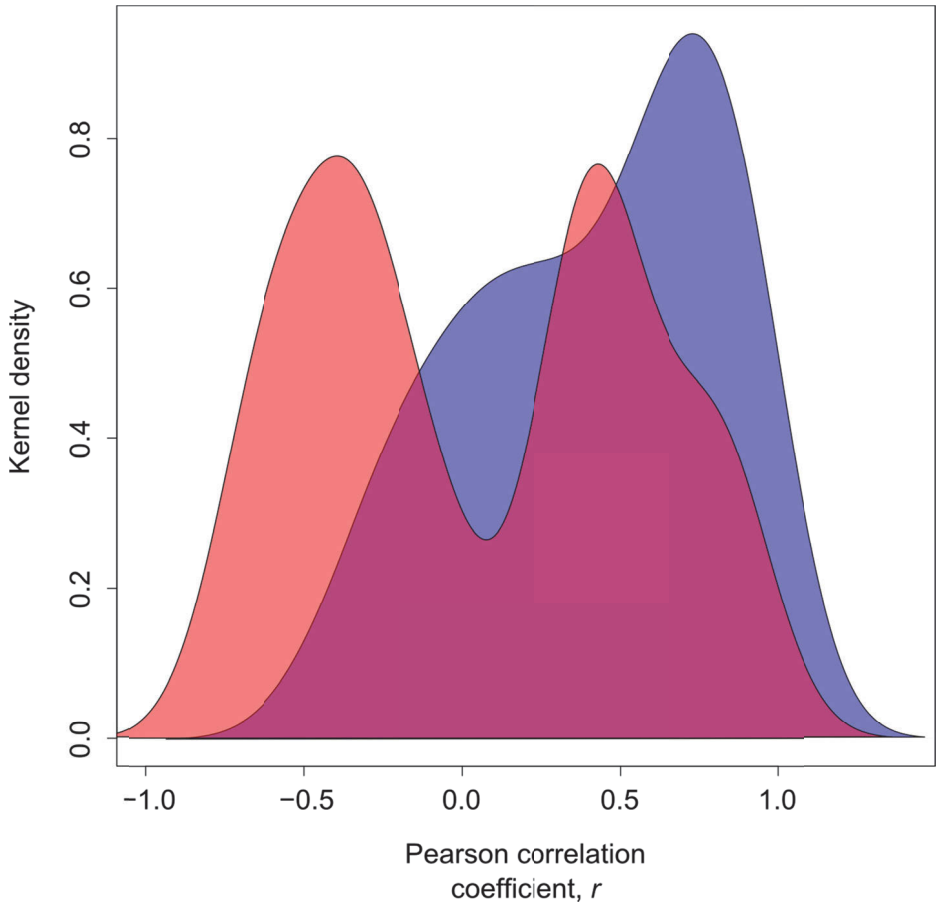


Figure S5. Probability density functions (PDFs) obtained by kernel density estimation of all pairwise Pearson correlation coefficients between cluster distributions (Table S1) I) of all *Bs*LipA variants but the two outliers, wild type and 6B, (blue) and II) of only the two outliers (red). A normal kernel function with an optimal smoothing parameter [14] at each data point was used for calculating the PDFs.

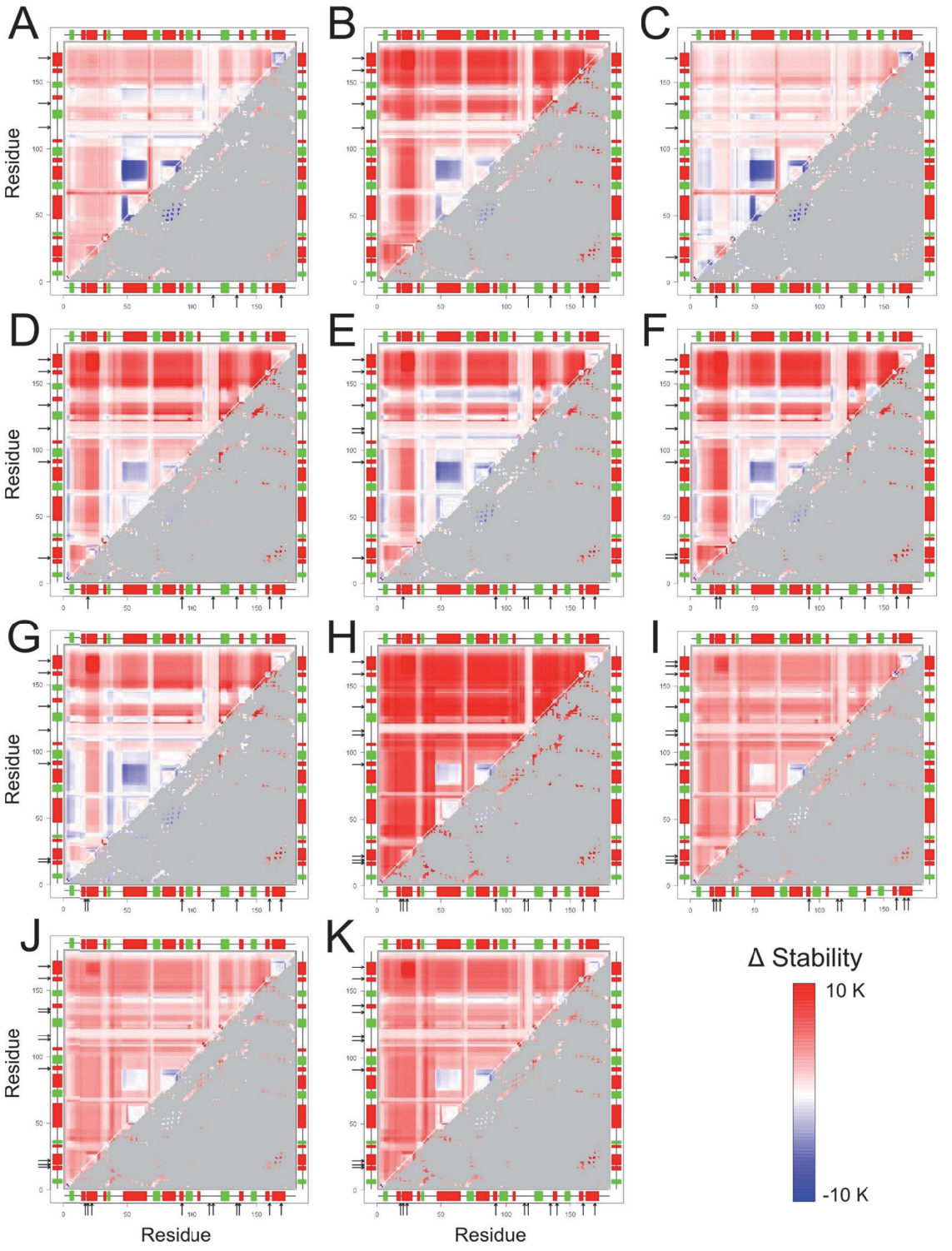


Figure S6. Differences in the stability of rigid contacts between variants of *Bs*LipA from Rao *et al.* Maps depict differences (against the wild type) for mutants TM (A), 1-17A4 (B), 1-8D5 (C), 2D9 (D), 3-18G4 (E), 3-11G1 (F), 3-3A9 (G), 4D3 (H), 5-D (I), 5-A (J), and 5B (K). Secondary structure elements as computed by the DSSP program [12,13] are indicated on both abscissa and ordinate: α-helix (red rectangle), β-strands (green rectangle), loop (black line). Arrows represent the mutated residue positions.

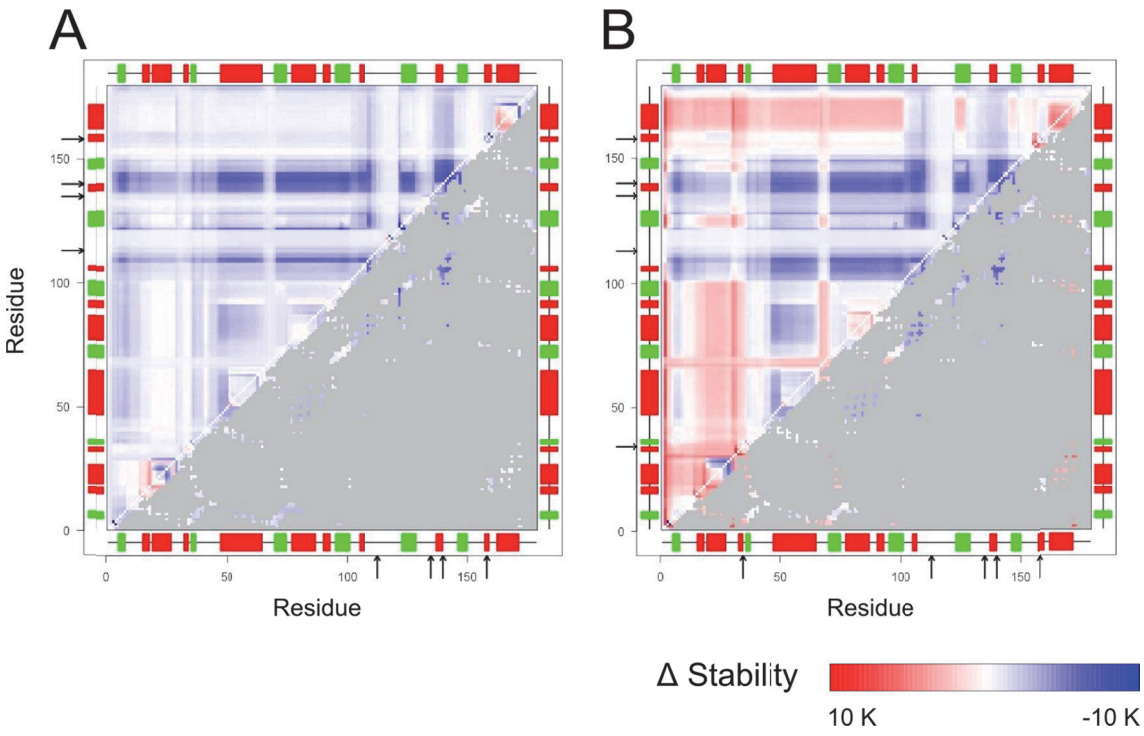


Figure S7. Differences in the stability of rigid contacts between mutants of *Bs*LipA from Reetz *et al*. Maps depict differences for mutants IX (A) and XI (B) against the wild type. Secondary structure elements as computed by the DSSP program [12,13] are indicated on both abscissa and ordinate: α-helix (red rectangle), β-strands (green rectangle), loop (black line). Arrows represent the mutated residue positions of the mutants with respect to the wild type.

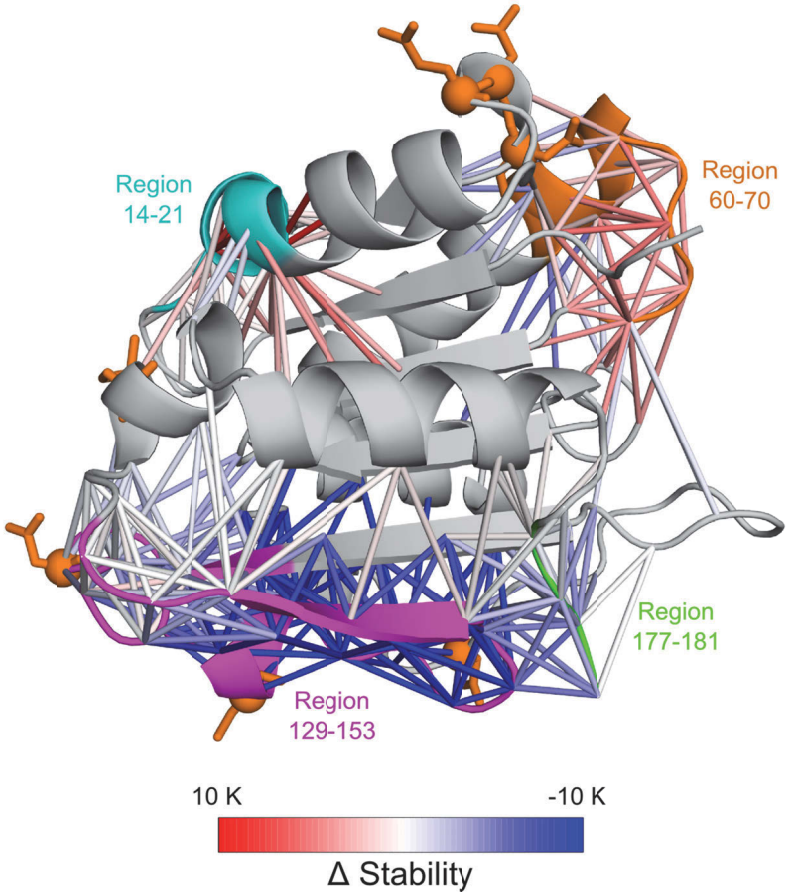


Figure S8. Differences in the stability of rigid contacts between variant X and wild type for residue neighbors shown on the structure of variant X. The sticks connecting C_α atoms of residue pairs are colored according to the color scale on the bottom. A contact in red (blue) is more (less) stable in variant X than in the wild type. Only those contacts involving residues in regions discussed in the main text are shown for clarity. Mutated residues are shown as sticks and a sphere at their C_α atoms.

Supplemental References

1. Whiteley W (2005) Counting out to the flexibility of molecules. Phys Biol 2: S116-S126.
2. Hespenheide BM, Jacobs DJ, Thorpe MF (2004) Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. J Phys Condens Matter 16: S5055-S5064.
3. Pfleger C, Gohke H (2013) Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. Structure 21: 1725-1734.
4. Radestock S, Gohlke H (2008) Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. Eng Life Sci 8: 507-522.
5. Radestock S, Gohlke H (2011) Protein rigidity and thermophilic adaptation. Proteins: Struct, Funct, Bioinf 79: 1089-1108.
6. Rathi PC, Radestock S, Gohlke H (2012) Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. J Biotechnol 159: 135-144.
7. Dahiyat BI, Gordon DB, Mayo SL (1997) Automated design of the surface positions of protein helices. Protein Sci 6: 1333-1337.
8. Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF (2002) Protein unfolding: rigidity lost. Proc Natl Acad Sci U S A 99: 3540-3545.
9. Pfleger C, Radestock S, Schmidt E, Gohlke H (2013) Global and local indices for characterizing biomolecular flexibility and rigidity. J Comp Chem 34: 220-233.
10. Andraud C, Beghdadi A, Lafait J (1994) Entropic analysis of random morphologies. Physica A 207: 208-212.
11. Cairns SP, Robinson DM, Loiselle DS (2008) Double-sigmoid model for fitting fatigue profiles in mouse fast- and slow-twitch muscle. Exp Physiol 93: 851-862.
12. Kabsch W, Sander C (1983) Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. Biopolymers 22: 2577-2637.
13. Joosten RP, Beek TAHT, Krieger E, Hekkelman ML, Hooft RWW, et al. (2011) A series of PDB related databases for everyday needs. Nucleic Acids Res 39: D411-D419.
14. Silverman BW (1998) Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC.

Publication VII

Application of rigidity theory to the thermostabilization of proteins

Prakash Chandra Rathi, Alexander Fulton, Karl-Erich Jaeger, and Holger Gohlke
Submitted manuscript, **2014.**
Contribution: 40%

Abstract

Protein thermostability is a crucial factor for biotechnological enzyme applications. Protein engineering studies aimed at improving thermostability have successfully applied both directed evolution and rational design. However, for rational approaches, the major challenge remains the prediction of mutation sites and optimal amino acid substitutions. Recently, we showed that such mutation sites can be identified as structural weak spots by rigidity theory-based thermal unfolding simulations of proteins. Here, we describe and validate a novel, unique, ensemble-based, yet highly efficient strategy to predict optimal amino acid substitutions at structural weak spots for improving a protein's thermostability. For this, we exploit the fact that in the majority of cases an increased structural rigidity of the folded state has been found as the cause for thermostability. When applied prospectively to the lipase A from *Bacillus subtilis*, we achieved both a high success rate in predicting thermostabilized lipase variants based on a single amino acid mutation and a remarkably large increase in those variants' thermostability. The results suggest that our strategy is a valuable complement to existing methods for rational protein design aimed at improving thermostability.

Introduction

Thermostability is a crucial factor for a wealth of biotechnological enzyme applications in chemical, environmental, cleaning, sensory, and pharmaceutical processes.[1, 2] Protein engineering aimed at improving thermostability is thus an important field of research in biotechnology. [3, 4] In this context, methods of directed evolution are usually applied, which mimic natural evolution by employing cycles of random mutagenesis/expression/selection leading to variants with desired properties.[5-13] However, directed evolution is limited by the fact that out of the extraordinarily large number of possible variant proteins, usually only a small subset can be experimentally tested for thermostability.[14] Alternatively, rational approaches have been successfully pursued,[15-18] but the major challenge here remains the prediction of mutation sites and the optimal amino acid substitution.[19, 20]

Regarding the prediction of mutation sites, we developed the rigidity theory-based Constraint Network Analysis (CNA) approach and implemented it as a web service (http://cpclab.uni-duesseldorf.de/cna/),[21-26] which allows identifying residues in a protein that are structural "weak spots". For this, a protein is modeled as a network of sites (atoms) and constraints (covalent and noncovalent interactions),[27] and then rigid atom clusters and flexible regions in between are rigorously determined by rigidity analysis.[28-30] By successively removing non-covalent constraints from the network, the thermal unfolding of the protein is simulated.[21-23, 31] From the thermal unfolding trajectory, a phase transition temperature T_p is identified, which relates to the (thermodynamic) thermostability, as are the weak spots. Mutating such weak spots should highly likely improve a protein's thermostability.[21-23]

Here, we describe and validate a novel and unique strategy based on the CNA approach to predict optimal amino acid substitutions at these weak spots for improving a protein's thermostability. For this, we exploit the fact that in the majority of cases an increased structural rigidity of the folded state has been identified as the underlying cause for thermostability.[32] We do so by adding a highly efficient, ensemble-based second step consisting of the generation of structural models of single-point site-saturation mutations at identified weak spots, the filtering of the models with respect to their structural quality, and the screening for variants with increased structural rigidity. Using a recently developed approach[33] that performs rigidity analyses on an ensemble of network topologies generated from a single input structure, rather than a structural ensemble, this second step only takes about 1 h per variant and can be performed in parallel for multiple variants. We have applied this approach prospectively using as a model enzyme the lipase LipA from *Bacillus subtilis* (*Bs*LipA), which is the smallest lipase known (consisting of 181 amino acids) and has considerable biotechnological importance.[34, 35] Out of 589 *Bs*LipA variants screened *in silico*, twelve were suggested for experimental testing. Of these, three showed a significant increase of up to 6.6°C in thermostability with respect to the wild-type enzyme (WT). We thus achieved both a high success rate in predicting thermostabilized lipase variants and a remarkably large increase in the thermostability of such variants. This demonstrates the value of this innovative approach, which extends the existing portfolio of methods for rational protein design aimed at improving thermostability.

Materials and methods

Predicting protein thermostability by Constraint Network Analysis

Constraint Network Analysis (CNA) predicts rigid and flexible regions within a biomolecule, which allows linking these static characteristics to the molecule's stability and function.[25, 26] CNA has been described in detail in refs.[25, 26, 33, 36] The approach has been used previously to predict the (thermodynamic) thermostability of proteins and to identify weak spot residues that, when mutated, are likely to improve thermostability.[21-23] In CNA, a protein is modeled as a *body-and-bar* network of bodies (atoms) and bars (covalent and noncovalent interactions). Each atom has six degrees of freedom, and each bar removes one degree of freedom.[27] An interaction between two atoms can be modeled as any number of bars between one and six depending on the strength of the interaction. Here, single covalent bonds (double and peptide bonds) were modeled as five (six) bars, hydrogen bonds and salt bridges (together referred to as "hydrogen bonds") as five bars, and hydrophobic interactions as two bars. For hydrogen bonds a hydrogen bond energy E_{HB} is computed by a modified version of the potential by Mayo and coworkers [37] as described in ref. [31].

By successively removing noncovalent constraints from a network, a thermal unfolding of the protein is simulated.[21-23, 31] Hydrogen bonds are removed from the network in increasing order of their strength,[37] i.e., hydrogen bonds with an energy $E_{HB} > E_{cut}(\sigma)$ are discarded from the network of state σ. In the present study, E_{cut} values ranging from −0.1 kcal mol^{-1} to −6.0 kcal mol^{-1} with a step size of 0.1 kcal mol^{-1} were used. E_{cut} can be converted to a temperature using a linear relation introduced by Radestock and Gohlke,[21, 22] according to which the range of E_{cut} used in this study is equivalent to increasing the temperature of the system from 302 K to 420 K with a step size of 2 K. The rigidity of each network state σ during the thermal unfolding simulation is analyzed by the pebble game algorithm[28, 29] as implemented in the FIRST program.[30] From these analyses, the change in the global rigidity characteristics are monitored by the cluster configuration entropy H_{type2}.[36] Finally, a phase transition temperature T_p is identified as the temperature when a largely rigid network becomes largely flexible. We showed that T_p can be used for predicting the thermodynamic thermostability of and identifying structural weak spots in a protein.[21-23] Usually, multiple phase transitions occur during the thermal unfolding of a protein because of its modular architecture, i.e., secondary structure elements can segregate from the largest rigid cluster as a whole.[22] In contrast to global indices, local indices monitor rigidity at a residue level. One such index, the rigidity index $r_{i,}$ is defined for each covalent bond i between two atoms as the E_{cut} value during the thermal unfolding simulation at which the bond changes from rigid to flexible.[36] For a C_α atom-based representation, the average of the two r_i values of the two backbone bonds is taken. As a two-dimensional itemization of r_i, a stability map rc_{ij} indicates for all residue pairs the E_{cut} value at which a rigid contact between the two residues i, j is lost, i.e., when the two residues stop belonging to the same rigid cluster. From rc_{ij}, a rigid cluster decomposition, i.e., a set of rigid clusters and flexible links in between, can be computed for each network state σ during the thermal unfolding simulation.

Rigidity analyses are sensitive with respect to the input structure.[38, 39] For improving the robustness, CNA is carried out on a structural ensemble derived from molecular dynamics (MD) simulations, and then results (T_p values and stability maps) are averaged.[23]

Generation of a structural ensemble of wild-type *Bs*LipA

MD simulations of WT *Bs*LipA were performed using the GPU accelerated version of PMEMD[40] of the AMBER 11 suite of programs[41, 42] together with the ff99SB force field.[43] The X-ray crystal structure of *Bs*LipA with the highest resolution (PDB ID: 1ISP; resolution 1.3 Å) was used as input structure.[44] Hydrogen atoms were added using the REDUCE program[45] during which side-chains of Asn, Gln, and His were flipped if necessary to optimize the hydrogen bond network. Then, the system, neutralized by addition of sodium counter-ions, was solvated by a truncated octahedral box of TIP3P[46] water such that a layer of water molecules of at least 11 Å width covers the protein surface. The particle mesh Ewald method[47] was used with a direct-space non-bonded cutoff of 8 Å. Bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm,[48] and the time step for the simulation was 2 fs. After equilibration, a production run of unrestrained MD in the canonical ensemble (NVT) was performed to generate a trajectory of 100 ns length, with conformations extracted every 40 ps from the last 80 ns resulting in a structural ensemble of 2000 conformations.

Weak spot identification and prioritization

The structural ensemble of 2000 conformations of WT *Bs*LipA (see above) was submitted to CNA (Figure 1-I). A thermal unfolding trajectory showing average rigid cluster decompositions during the thermal unfolding simulation was reconstructed from the average stability map (Figure 1-II). The thermal unfolding trajectory was visually inspected for identifying transitions at which the rigidity of WT *Bs*LipA is substantially reduced. For the inspection, the VisualCNA software, a graphical user interface for CNA, (P. C. Rathi, D. Mulnaes, H. Gohlke, unpublished results) was applied. The inspection was done with a view that rigidifying contacts between the largest rigid cluster and residues that segregate at these substantial phase transitions should improve the thermostability of the protein by delaying the disintegration of the largest rigid cluster. Accordingly, at every transition, residues that are in the neighborhood of, and whose side-chains point towards the largest rigid cluster from which they segregated, were identified as potential weak spots (Figure 1-III). Weak spot residues that showed a high sequence conservation (≥ 80% identity) in a multiple sequence alignment of 296 lipase class 2 sequences obtained from the Pfam database[49] were not considered any further (Figure 1-IV).

--- Figure 1---

Modeling of single-point site-saturation mutations

Structures of all possible mutations at each weak spot residue were generated by the SCWRL program[50] using WT *Bs*LipA (PDB ID: 1ISP) as a template (Figure 1-V). SCWRL constructs variant models by predicting backbone-dependent side-chain conformations with the help of

a rotamer library; coordinates of backbone atoms remain unchanged. Conformations of side-chains of all residues within 8 Å of a mutated residue were re-predicted in order to allow for a local structural relaxation. The goodness of fit of the mutated side-chain in its environment was assessed using the ANOLEA server,[51] which provides residue-wise non-local (with respect to sequence space) interaction energies using a knowledge-based potential of mean force.[52] A variant was discarded if its average ANOLEA energy of the neighboring residues (≤ 5 Å of the mutation) is higher than the average energy of the same residues in WT by ≥ 2 kcal mol^{-1}. For all variant structures, hydrogen atoms were added using REDUCE[45] in an identical way as done for WT *Bs*LipA. Finally, the structures were minimized by 2000 steps of conjugate gradient minimization (including an initial steepest descent minimization for 100 steps) or until the root mean-square gradient of the energy was < $1.0 \cdot 10^{-4}$ kcal mol^{-1} $Å^{-1}$. The energy minimization was carried out with Amber11[41] using the ff99SB force field[43] and the GB^{OBC} generalized Born model [53] for modeling solvation effects.

Thermostability prediction and prioritization of variants

In order to circumvent compute-intensive MD simulations for generating *structural ensembles* of each of the BsLipA variants, the more efficient ENT^{FNC} approach introduced recently by Pfleger *et al.*[33] was used in connection with thermal unfolding simulations as implemented in the CNA software[25]. Here, rigidity analyses are performed on an *ensemble of network topologies* generated from a single input structure by using fuzzy noncovalent constraints. Ensembles of 1000 network topologies of all single point variants of *Bs*LipA were analyzed; for consistency, the WT *Bs*LipA structure was treated in the same way (including an energy minimization as described above). For each variant and WT, T_p was automatically identified as the inflection point of the sigmoid with the larger change in H_{type2} using a double sigmoid function[23] fitted to H_{type2} *vs*. T curves. That way, in most cases, a late transition involving the final decay of the largest rigid cluster is identified as T_p [22] except when a very large loss of rigidity occurs during an early transition. Based on ensemble-averaged T_p (Figure 1-VI), variants were selected for experimental characterization of their thermostability. Table S1 in the Supporting Information (SI) summarizes the computing times required for the weak spot identification, site saturation mutagenesis, and screening for increased structural rigidity.

Materials

Phusion high fidelity polymerase, dNTPs, and the BCA protein quantification kit were obtained from Thermo Scientific (St. Leon Rot, Germany). Plasmid isolation and PCR cleanup was performed using the purification kits from Analytik Jena (Jena, Germany). Sequencing and synthesis of oligonucleotides was carried out by MWG eurofins (Ebersberg, Germany). The substrates for activity assays (*p*-nitrophenyl-palmitate (*p*NPP), *p*-nitrophenyl-decanoate (*p*NPD)) were purchased from Sigma Aldrich (Hamburg, Germany). Ni-NTA superflow resin and disposable PD-10 desalting columns were purchased from Macherey-Nagel (Düren, Germany).

Site directed mutagenesis

The *lipA* gene from *B. subtilis* encoding lipase without its signal sequence was cloned into pET22b+ (Novagen) with an N-terminal hexahistidine-tag fusion for purification. The site directed mutations (Table S2 in the SI) were introduced with a modified Quikchange® protocol.[54] The amplification was carried out in two separate 25 µl reactions with each 10-50 ng of template, 0.2 pM of either the forward or reverse primer (Table S2 in the SI), 0.2 mM dNTPs, 3% DMSO (v/v), and 1 U of Phusion polymerase in Phusion GC-buffer containing 7.5 mM $MgCl_2$. PCR conditions as follows: initial denaturation at 98°C for 10 min followed by 23 cycles of 98°C for 1 min, 55°C for 1 min, and 68°C for 3.5 min followed by final elongation at 68°C for 7 min. The PCR was paused after 5 cycles to combine the forward and reverse primer reaction and continued for the remaining 18 cycles. Template DNA was removed with 30U *Dpn*I at 37°C for 16 h. The hydrolysis reaction was stopped at 75°C for 20 min followed by a PCR clean up and eluted into 10 µl ddH_2O. An aliquot of 1 µl was transformed into *E. coli* DH5α electrocompetent cells and plated on selective LB plates. The plasmid DNA of the positive transformants was isolated and sequenced to ensure the successful site directed mutagenesis. Plasmid DNA with the desired mutation in *lipA* was stored at -20°C.

Cultivation and protein purification

The wild-type enzyme LipA and twelve variants were expressed in *E. coli* BL21(DE3) from a T7 promoter. An aliquot of 50 µl chemically competent *E. coli* BL21(DE3) was transformed with 1 µl of plasmid DNA, and cells were grown overnight in 10 ml selective LB media. This preculture was used to inoculate the main expression culture in 250 ml of TB autodinducing media in a 5 l shaking flask, shaken at 150 rpm for 3 h at 37°C followed by 72 h at 15°C. Cells were collected by centrifugation for 45 min at 5000 rpm and lysed by passing three times through a French pressure cell at 500 bar (lysis buffer: 50 mM NaH_2PO_4, 300 mM NaCl adjusted to pH 8). The soluble fraction was incubated 30 min under mild agitation with 1 ml Ni-NTA Superflow Resin (Qiagen). The resin was then washed with 50 ml of 20 mM imidazole on a gravity flow column after which the protein was eluted with 250 mM imidazole into fractions of 1.5 ml each until there was no absorption detectable at 280 nm. The fractions with the highest absorption at 280 nm were applied to a disposable PD-10 desalting column according to manufacturer's description to remove imidazole. The protein concentration of the desalted sample was measured using the Micro BCA protein assay reagent (Thermo scientific), with bovine serum albumin in concentrations of 25 – 2000 µg/ml as a standard.[55] The sample purity was analyzed by SDS-PAGE (Figure S1 in the SI).

Thermostability assay

The protein concentration was adjusted to 0.3 mg/ml in 10 mM glycine buffer pH 11 as determined by a BCA measurement. The enzyme stock was diluted 10-fold into 50 µl of 50 mM NaP_i/KP_i pH 7 and incubated at different temperatures between 40°C to 60°C in a PCR cycler. The lipase activity was measured with a *p*NPP substrate solution[56] (1.6 mM *p*NPP, 10% isopropanol, 50 mM NaP_i/KP_i pH 7, 1 mg/ml gum arabic, 2 mg/ml sodium

desoxycholate) warmed up in a separate PCR cycler 5 min prior to the measurement. For the measurement 50 µl of substrate solution were added to 50 µl of the sample, and the hydrolysis rate was quantified by the change in absorption at 410 nm for the duration of 5 min in a SpectraMax Plus plate reader (Molecular Devices, Sunnyvale, CA).

Michaelis-Menten kinetics

The initial rates of hydrolysis as a function of the substrate concentration were measured by following the change in absorption at 410 nm for 60 s with 10 s lag time in 1 ml disposable cuvettes using a SpectraMax Plus plate reader with a built in cuvette port (Molecular Devices, Sunnyvale, CA). The substrate *p*NPP is thermostable and was thus chosen for the thermostability assays; however, its low solubility at higher concentrations did not allow determining kinetic parameters. Therefore, kinetic constants were determined using the substrate *p*NPD, which is less stable but soluble also at higher concentrations. The measurement was started by mixing 990 µl of a 40°C substrate solution (6 – 1600 µM *p*NPD in 10% isopropanol, 50 mM NaPi/KPi pH 7, 1 mg/ml gum arabic, 2 mg/ml sodium desoxycholate) with 10 µl of room temperature enzyme solution (0.05 mg/ml in 50 mM NaPi/KPi pH 7, 1 mg/ml gum arabic, 2 mg/ml sodium desoxycholate). The K_M and k_{cat} values for each variant were derived by nonlinear fitting of the Michaelis–Menten curve using the software Graphpad PRISM (GraphPad Software, Inc., San Diego, CA) and the protein concentration from a BCA measurement as described earlier.

Results and discussion

Weak spot identification for *Bs*LipA

*Bs*LipA is a protein of 181 amino acids with a minimal α/β hydrolase fold; in this fold, a central parallel β-sheet of six β-strands is surrounded by six α-helices. Ser77, Asp133, and His156 constitute the catalytic triad.[57] For identifying weak spots on WT *Bs*LipA, initially, a thermal unfolding simulation was carried out by CNA on an ensemble of 2000 WT *Bs*LipA structures extracted from a MD trajectory of 100 ns length. The ensemble-based CNA was pursued to increase the robustness of the rigidity analyses.[23, 33, 38, 39] An average unfolding trajectory was then reconstructed from the average stability map (Figure 2). The unfolding involves early segregation of loops from the largest rigid cluster, followed by the segregation of α-helices and, finally, the segregation and disintegration of the β-sheet region. This order of segregation is in agreement with experimental findings on the unfolding of α/β hydrolase proteins:[58, 59] Unfolding studies of human placental alkaline phosphatase[59] and 1*H*-3-hydroxy-4-oxoquinaldine 2,4-dioxygenase[58] showed an early loss of α helices, but only very little change in β-strand content until the late stage of the unfolding. The apparently realistic description of WT *Bs*LipA thermal unfolding encouraged us to identify weak spots at major phase transitions along the unfolding trajectory. The procedure is at variance with previous studies in which weak spots were identified only at the last major phase transition related to the terminal loss of rigidity in the protein.[21-23] The procedure followed here has the advantage that it allows evaluating whether strengthening residues that segregate earlier from the largest rigid cluster will also lead to a protein's thermostabilization.

By visual inspection of the unfolding trajectory, we identified five major transitions (T1-T5) at which secondary structure elements αA, αF, αD and αE, αB, as well as αC and the central beta sheet segregate from the largest rigid cluster at temperatures 316, 318, 334, 336, and 338 K, respectively (Figure 2 and Table 1). Note that the reported temperatures should be considered relative values only as stated previously.[22, 24, 25] Weak spot residues were then identified as those residues that are in the neighborhood of the largest rigid cluster from which they segregate at the respective major transition. This follows the rationale that these residues are particularly promising for increasing *Bs*LipA's thermostability considering that their mutation can improve the interaction strength with the largest rigid cluster and, hence, delay the disintegration of that cluster with increasing temperature. In total, 36 weak spots were identified, which are located on α-helices and loops joining α-helices and β-strands (Figure 2). The weak spot residues are very diverse in size (ranging from Gly to Trp) and physicochemical properties (charged, uncharged polar, and hydrophobic) (Table 1). Finally, weak spot residues at highly conserved sequence positions were discarded (Table 1, Figure S2 in the SI). This was done because conserved residues are usually important for function and/or stability of a protein and, hence, should not be mutated.[60-63]

---Figure 2---

---Table 1---

Variant construction and prioritization

For each of the remaining 31 weak spots (~17% of all *Bs*LipA residues), computational site saturation mutagenesis was performed by generating structures of all possible single-point amino acid substitutions using the SCWRL program. This resulted in 589 single point variants. Structures with a single point mutation can be reliably modeled using SCWRL because a single-point mutation should not grossly change the conformation of the backbone as evidenced by a very low C_α atom root mean-square deviation (< 0.40 Å) between the crystal structure of WT *Bs*LipA and seven variants (incorporating ≤ 12 mutations) deposited in the Protein Data Bank (PDB codes: 1T2N, 1T4M, 3D2A, 3D2B, 3D2C, 3QMM, and 3QZU).[64] 67 variant structures with mutated residues unfavorably embedded in the surroundings of the protein as judged by the ANOELA energy (see above) were discarded (Figure S2 in the SI). In such structures, the mutation apparently does not fit into the environment of the other residues.

The remaining 522 variants were subjected to thermal unfolding simulations using the ENT^{FNC} approach[33] implemented in CNA. Differences in the phase transition temperatures $\Delta T_p = T_p$ (variant) − T_p (WT) were averaged over 1000 simulations started from different network topologies for each variant as described in the "Materials and methods" section. A map of ΔT_p values of all variants is shown in Figure S2 in the SI. Of the 239 single point mutations at the 13 weak spots identified from early transitions at low temperatures (T1 and T2), only four resulted in a higher T_p than WT *Bs*LipA (Figure S2 in the SI); two of these increases were statistically significant ($p < 0.05$ according to Welch's t-test[65]). This is in line with the fact that T_p is identified as the temperature of the late transition that involves the

final decay of the largest rigid cluster; strengthening interactions of residues involved in the disintegration of the largest rigid clusters at earlier transitions thus would only result in a profound increase in T_p if the strengthening also changed the order of disintegration steps. At five weak spots identified at transition T3, seven mutations resulted in higher T_p values than WT *Bs*LipA (Figure S2 in the SI); three of these increases were statistically significant. The most pronounced predicted thermostabilization both in terms of the number of variants showing increased T_p values (55, of which 27 were significant) and the magnitude of the T_p increase (Figure S2 in the SI) was observed for mutations at the nine weak spots identified at transition T4. Finally, nine mutations at four weak spot residues identified at the last transition T5 resulted in an increase in T_p compared to WT *Bs*LipA; six of these increases were significant. In total, this results in a predicted thermostabilization with respect to WT *Bs*LipA for 75 out of the 522 mutations (~14%) investigated.

In order to further reduce the number of mutations for experimental validation, variants were prioritized based on their ΔT_p values and the weak spot residue that is mutated: For each weak spot residue and all mutations with $\Delta T_p > 1$, the mutation with the highest ΔT_p was chosen for experimental validation. The sole exception is G104 located in the active site, for which two mutations were chosen. This resulted in twelve lipase variants of which the most are associated with weak spot residues on helix αB identified during the late transition T4 (Table 2 and Figure S2 in the SI).

---Table 2---

Thermostability of *Bs*LipA variants

Initially, specific activities of WT *Bs*LipA and the twelve variants (Table 2) for hydrolysis of *p*NPP were measured at temperatures between 40 and 60ºC after keeping them at the respective temperatures for 5 min. Under these conditions, WT *Bs*LipA showed the highest specific activity (246 U/mg) among all *Bs*LipA variants at the temperature of maximum activity T_{max} (40ºC) (Figure S3 in the SI). At temperatures above 55°C, the activity begins to drop, which is probably due to an unfolding already within 5 min of preincubation. Notably, two variants, F58I and V96S, showed higher activities than the WT at temperatures above 58ºC (Figure S3 in the SI), which may originate from them being more stable at high temperatures. Next, thermostability was assessed by measuring the activity of each *Bs*LipA variant at temperatures between 40 and 60ºC after incubating the respective variant at these temperatures for 30 min. Three variants, V54H, F58I, and V96S, were more thermostable than WT; they consistently showed higher activities than the WT at temperatures above 48ºC (Figure 3A and Figure S4 in the SI). Other variants, however, were found to be less thermostable than the WT (Figure S4 in the SI). The largest differences between thermostabilities of WT and variants of *Bs*LipA was observed at 53.5ºC where the activities of V54H and V96S were twice as a high as those of the WT, and the activity of F58I was four times higher (Figure 3B). At low temperatures, F58I and V96S showed similar activities as the WT, and V54H showed half the activity of the WT (Figure 3B). Finally, the kinetic constants of these variants were derived from initial rate measurements for hydrolysis of *p*NPD at 40°C (see section “Materials and methods”). No significant impact on the Michaelis

constants (K_M) was observed, and the turnover numbers (k_{cat}) were reduced by at most 25% (Table S3 in the SI). Thus, the thermostability of the variants has been increased without significantly influencing k_{cat} / K_M at 40°C.

The thermostability of *Bs*LipA variants was quantified by T'_{50} values; these values report on the temperature at which the fraction of the activity to the initial activity (at 40ºC) is 50% after incubation for 30 min. This is different from the T_{50} values normally used for characterizing the thermostability of proteins[20, 66, 67] in that the activity here is measured at the temperature of incubation, not at room temperature after cooling. T'_{50} thus reports on the thermo-tolerance of an enzyme during operational bioprocesses carried out at elevated temperatures for a longer duration of time, e.g., as done in the lipid processing industry.[68] The three variants V54H, F58I, and V96S showed T'_{50} values higher by 5.7, 6.6, and 3.6ºC, respectively, than WT *Bs*LipA (Figure 3C and Table 2). The predicted ΔT_p values for these variants were similar to each other, in agreement with the similar T'_{50} values found, but at the lower end of all predicted ΔT_p, suggesting that ΔT_p as computed in this study is more suitable for prioritizing variants than for ranking them (Table 2).

The three thermostable variants involve mutations at weak spots identified at later phase transitions T4 and T5 during the thermal unfolding simulation. This finding supports our previous reasoning that it is the late phase transition(s) involving the final decay of the rigid core during thermal unfolding that determine(s) the thermodynamic thermostability of a protein.[21, 22] Accordingly, mutations that strengthen connections of weak spot residues identified at late phase transitions, and, hence, increase the local stability of the folding core, should particularly improve thermostability. A sound discussion of this implication requires X-ray structural data of the variants, which is not yet available. Still, using the modeled variant structures, we observed that the three variants V54H, F58I, and V96S have in general stronger rigid contacts between neighboring residues than the WT: On average, the mutations V54H, F58I, and V96S increased the strength of rigid contacts of neighboring residues by 2.0, 1.2, and 0.4 K, respectively, compared to WT (Figure S5 in the SI; see section “Pairwise contact stability of neighboring residues” in the SI for an explanation how these values were calculated).

Considering the most thermostable variant F58I in more detail, the strengthening holds true for local contacts as well as contacts that arise from a long-range stabilization. As to local contacts, Ile at position 58 along with residues of the neighboring loop β4-αB (A38, V39, D40) are part of a rigid cluster, which persists to a temperature ~3 K higher than the rigid cluster formed by F58 of WT and the same loop residues (Figure 4A, B; Figure S5B and S6A in the SI). From the input structure for the thermal unfolding simulations using the ENT^{FNC} approach,[33] one can derive that the longer persistence of these loop residues in the rigid cluster in variant F58I results from their better side-chain packing than in WT. In particular, in F58I, V39 forms four hydrophobic contacts with three different residues (V7, S16, F41), whereas in WT it only forms two such hydrophobic contacts (Figure 4C). However, not all F58I mutation-induced changes lead to a stabilization: Contacts of the C-terminus of loop β4-αB with neighboring residues become less stable owing to a different conformation of W42

in F58I (Figure 4C); likewise, contacts between N-terminal residues of loop β4-αB and αB are weakened (Figure 4C). As to contacts that arise from a long-range stabilization, residues of several pairs of secondary structure (αA/β strands 3,4,5; αB/αC; loop αB-β5/loop αC-β6; loop αC-β6/loop αD-β7) remain part of one rigid cluster for temperatures 2-5 K higher in the variant F58I than in WT (Figure 4D; Figure S5B and S6 B-E in the SI). This demonstrates the inherent long-range aspect to rigidity percolation [43,55,70,71], i.e., a local change on one end of a network can affect the stability all across the network.

Five mutations at weak spots identified at transitions T4 and T5 resulted in lower T'_{50} values than that of WT *Bs*LipA (Table 2). This result appears to contradict our reasoning that mutations which strengthen connections of weak spot residues identified at late phase transitions should particularly improve thermostability. In each case, however, a small amino acid was substituted by a large amino acid, which likely could not be accommodated by the fold. This calls for improved modeling approaches for the variant construction in future studies, e.g., by applying comparative modeling rather than side-chain placement only. Along the same lines, the two variants G104I and G104L out of the three variants that showed a complete loss of activity after 30 min incubation at temperatures between 40-60°C involved a residue located in the active site. While at the opposite side of the catalytic triad, introducing larger residues there may occlude the substrate binding region. Such weak spots can be filtered out in future studies based on their location in the protein.[70]

Conclusions

We developed a novel rational approach based on increasing structural rigidity for improving a protein's thermostability and applied it prospectively to *Bs*LipA. The approach combines ensemble- and rigidity theory-based weak spot prediction by CNA, filtering of weak spots according to sequence conservation, computational site saturation mutagenesis, assessment of variant structures with respect to their structural quality, and screening of the variants for increased structural rigidity by ensemble-based CNA. Two reasons account for the high computational efficiency of our approach: In the first step, the number of potential mutation sites is dramatically reduced due to concentrating only on structural weak spots. In the second step, ensembles of network topologies, rather than structural ensembles, are employed alleviating the need for costly conformation sampling. As a result, about one mutation per hour can be processed once weak spots have been detected (Table S1 in the SI), and this task is furthermore trivially parallelizable for multiple single point mutations.

As to the application to *Bs*LipA, our approach resulted in three out of twelve experimentally tested single-point mutations with significantly increased thermostability with respect to WT, yielding a success rate of 25% and 6.6°C as the largest increase in thermostability. Due to the lack of appropriate data, the success rate for thermostabilization associated with random mutations on *Bs*LipA is unknown. However, it is instructive to compare our results to a complete site saturation mutagenesis of *Bs*LipA and subsequent testing of each possible variant for improved detergent stability: Over 3439 single point mutations, the success rate amounts to 3% there (A. Fulton, J. Frauenkron-Machedjou, P. Skoczinski, S. Wilhelm, U. Schwaneberg, K.-E. Jaeger, unpublished results). The effectiveness of our approach is also

demonstrated when comparing it to the study by Reetz and coworkers[20] applying iterative saturation mutagenesis to *Bs*LipA. The largest increase in T_{50} they have found for a single point variant in the first step was 4.3°C, and about 8000 colonies needed to be screened to result in two variants carrying five and seven mutations after five steps of optimization with an increase of T_{50} by 45°C. It should be noted that the study of Reetz *et al.* differs from ours in a fundamental aspect: Reetz *et al.* chose as weak spots those residues that showed the highest crystallographic B-factors, i.e., were most mobile. In contrast, weak spots in our study constitute residues that are rigid until shortly before the folding core ceases to exist. Subsequent work by the Reetz group[71] showed that their variants became more thermostable due to kinetic reasons. As our approach is based on increasing the structural rigidity of the folded state, we speculate that our variants are more thermostable due to thermodynamic reasons. In summary, these results suggest that our approach is a valuable complement to existing methods for rational protein design aimed at improving thermostability. The more thermostable variants can then serve as starting points for further engineering of substrate scope and/or enantioselectivity by directed evolution, exploiting that enhanced thermostability promotes the ease of evolvability.[72]

Acknowledgments

PCR and HG are grateful to the Ministry of Innovation, Science, and Research of North Rhine-Westphalia and Heinrich Heine University Düsseldorf for a scholarship to PCR within the CLIB-Graduate Cluster Industrial Biotechnology. AF and KEJ gratefully acknowledge support by the German Research Foundation (DFG) within the research training group 1166 “Biocatalysis using Non-Conventional Media-BioNoCo”.

Tables

Table 1. Phase transition points at which weak spot residues are identified during the thermal unfolding simulation of *Bs*LipA.

Phase transition	Temperature of the phase transition[a]	Major secondary structures segregating from the giant rigid cluster	Weak spot residues[b]
T1	314-316	αA	I22, L26, W31
T2	316-318	αF	**D133**, V136, G158, L159, L160, S163, V165, N166, I169, G172, L173
T3	332-334	αD and αE	G103, G104, A105, **N106**, T109, S141
T4	334-336	αB	N48, N51, G52, V54, L55, F58, V59, **V62**, L63, E65, **T66**, **V71**
T5	336-338	αC and central β sheet	T83, L84, I87, V96

[a] In K.
[b] Residues in bold are highly conserved in the multiple sequence alignment; see the main text for details.

Table 2. *Bs*LipA variants characterized experimentally.

*Bs*LipA variant[a]	Location of mutation on secondary structure element	Phase transition of weak spot identification	Predicted ΔT_p[b]	T'_{50}[c]
Wild type	-		-	49.10
I22W	αA	T1	2.80	44.89
N51F	αB	T4	4.30	46.05
G52M	αB	T4	16.47	49.59
V54H	αB	T4	2.09	54.80
L55F	αB	T4	3.48	47.62
F58I	αB	T4	2.27	55.65
V59F	αB	T4	11.95	49.44
I87W	αC	T5	4.91	-[d]
V96S	β6	T5	2.36	52.65
G104I	Loop β6- αD	T3	1.98	-[d]
G104L	Loop β6- αD	T3	5.07	-[d]
L160H	αF	T2	2.25	43.30

[a] Variants highlighted in bold show a significant increase in T'_{50} with respect to WT.
[b] Difference phase transition temperatures T_p (variant) − T_p (WT); in °C.
[c] The temperature at which the fraction of the activity to the initial activity (at 40ºC) is 50% after incubating for 30 min; in °C.
[d] No activity after 30 min incubation at temperatures of 40-60ºC.

Figure captions

Figure 1. Strategy to rationally predict mutations that increase structural rigidity and thermostability. A structural ensemble of the respective protein is generated by MD simulations (I). The average thermal unfolding trajectory depicting a decomposition into rigid clusters (in the order of decreasing size colored in blue, green, magenta, cyan, orange, and violet) for each step of the unfolding simulation is created by subjecting the structural ensemble to CNA (II). For every major transition during the thermal unfolding, weak spot residues (depicted as a sphere for the C_α atom and sticks for the side-chain) are identified as residues that segregate from the largest rigid cluster and can potentially interact with the then largest rigid cluster upon mutation (III). Weak spot residues identified in step III that are highly conserved in a multiple sequence alignment of the protein family (≥ 80% identity) are removed from the weak spot list (IV). For each remaining weak spot, structures of single-point variants involving mutations to all other 19 amino acids (termed M1-M19) are generated using the SCWRL program.[50] Mutations that lead to energetically unfavorable structures (indicated by red discs around the mutated residue in the case of M18) as calculated by the ANOLEA server[51] are not considered further (V). Finally, for each variant, the phase transition temperature T_p is computed using CNA; a higher T_p value than that of the WT protein indicates a thermostabilizing mutation (VI). All figures of *Bs*LipA structures in this publication were generated with PyMOL (http://www.pymol.org).

Figure 2. Thermal unfolding trajectory of WT *Bs*LipA showing transitions for which weak spot residues were identified. Uniformly colored bodies represent rigid clusters; for clarity, only the largest rigid cluster (blue) is shown for the first four transitions (T1-T4), and the two largest rigid clusters (blue and green) are shown for the last transition (T5). C_α atoms of the identified weak spot residues are shown as spheres, and side-chain atoms are shown in stick representation. Weak spot residues are colored according to the rigid cluster they are part of (rigid clusters are assigned blue, green, magenta, cyan, and orange colors in the descending order of their size in terms of the number of residues); a weak spot residue colored in gray indicates that it is part of a rigid cluster composed of less than three residues. Weak spot residues that are highly conserved in the multiple sequence alignment of the lipase family (see the main text) are not shown. Important helices that segregate from the largest rigid cluster at the respective transition are labeled.

Figure 3. Thermostability of WT (black) and variants V54H (blue), F58I (green), and V96S (red) shown as activity *vs*. temperature curves. The activity was measured at indicated temperatures after incubating for 30 min at these temperatures. Curves show absolute specific activity (A), activity normalized by the activity of WT (B), and the residual activity after 30 min compared to the initial activity after 5 min incubation (C).

Figure 4. Structural origin of differences in the thermostability of WT and F58I, shown by a rigid cluster decomposition of WT (A) and F58I (B) at 316 K. Rigid clusters are colored using the same color scheme as in Figure 2. The mutation site (residue 58), shown by a cyan (A) and a magenta (B) surface, is part of the largest rigid cluster (blue) in both WT and F58I. Hydrophobic contacts in the proximity of the mutation site between carbon atom pairs at

most 3.8 Å apart are sown as green (WT) and red (F58I) dashed lines (C). Residues involved in making such contacts are shown as cyan (WT) and magenta (F58I) sticks. Differences in the stability of "rigid contacts" between variant F58I and WT depicted on the variant structure (D). Two residues form a "rigid contact" if they belong to one rigid cluster. A red (blue) stick connecting C_α atoms of two residues indicates that a rigid contact in the variant is more (less) stable than in the WT (see color scale). Only those contacts of variant F58I that are stabilized or destabilized by ≥ 2 K are shown for clarity; for the same reason, contacts between two residues of the same secondary structure element are not shown. The mutated residue I58 is displayed by magenta sticks. Blow-ups of panel D showing the contact stability between secondary structure pairs mentioned in the main text can be found in Figure S6 in the SI.

Figures

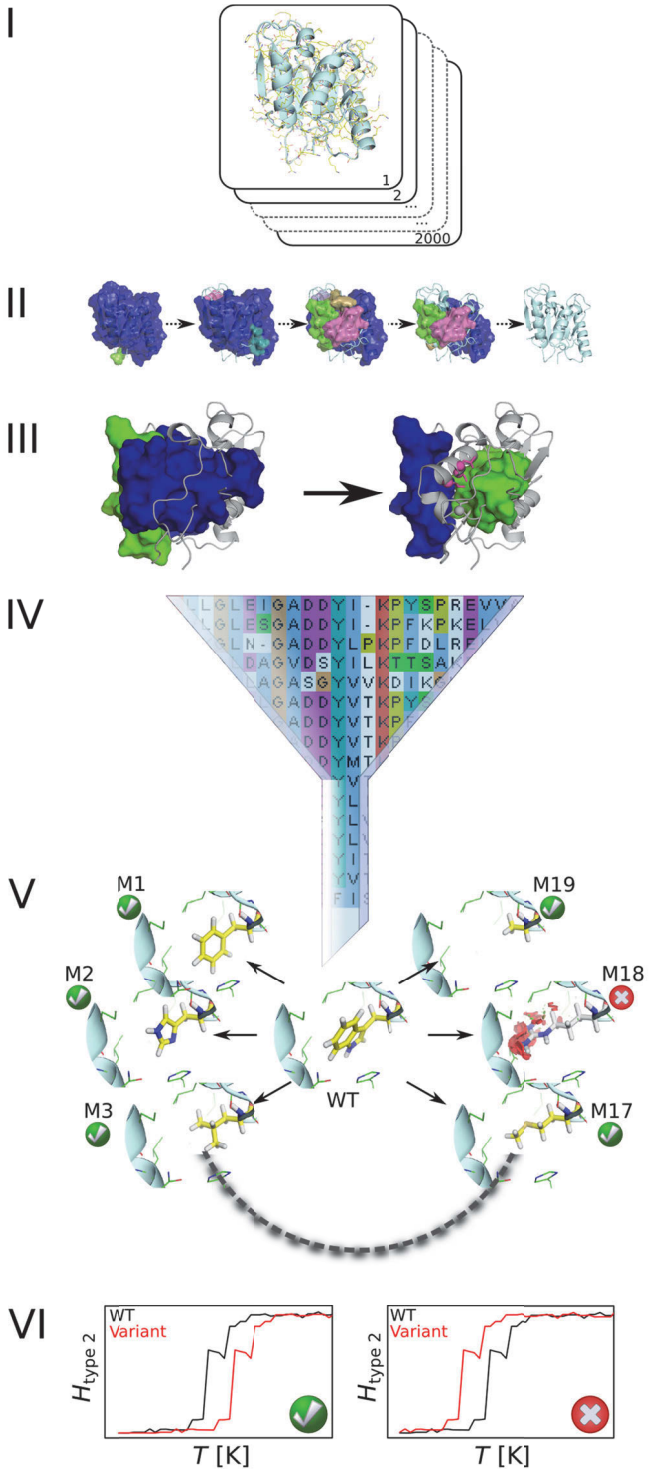


Figure 1

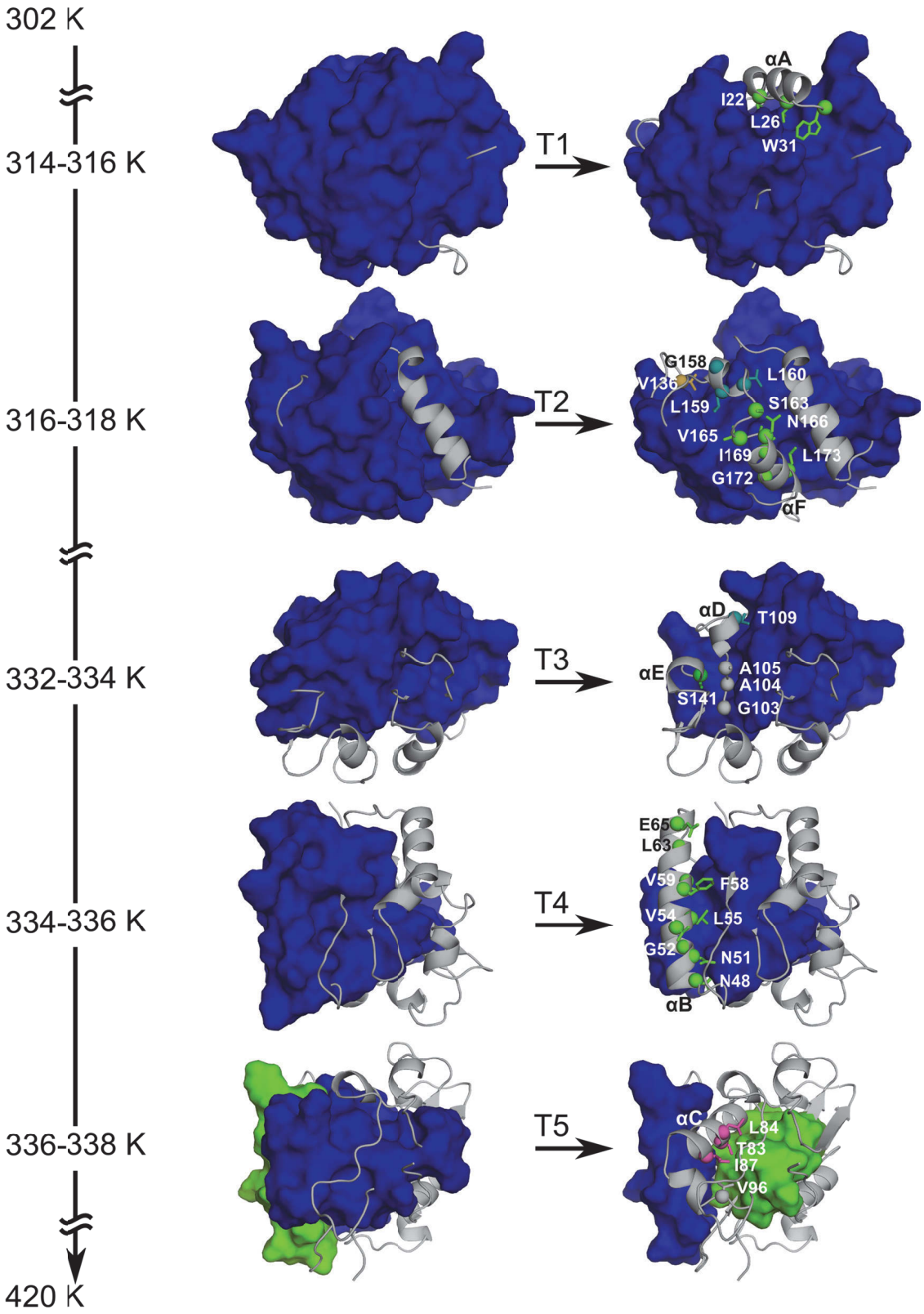

302 K
314-316 K
316-318 K
332-334 K
334-336 K
336-338 K
420 K
T1
T2
T3
T4
T5
αA
I22
L26
W31
G158
V136
L160
L159
S163
N166
V165
I169
L173
G172
αF
αD
T109
αE
A105
A104
S141
G103
E65
L63
V59
F58
V54
L55
G52
N51
N48
αB
αC
L84
T83
I87
V96

Figure 2

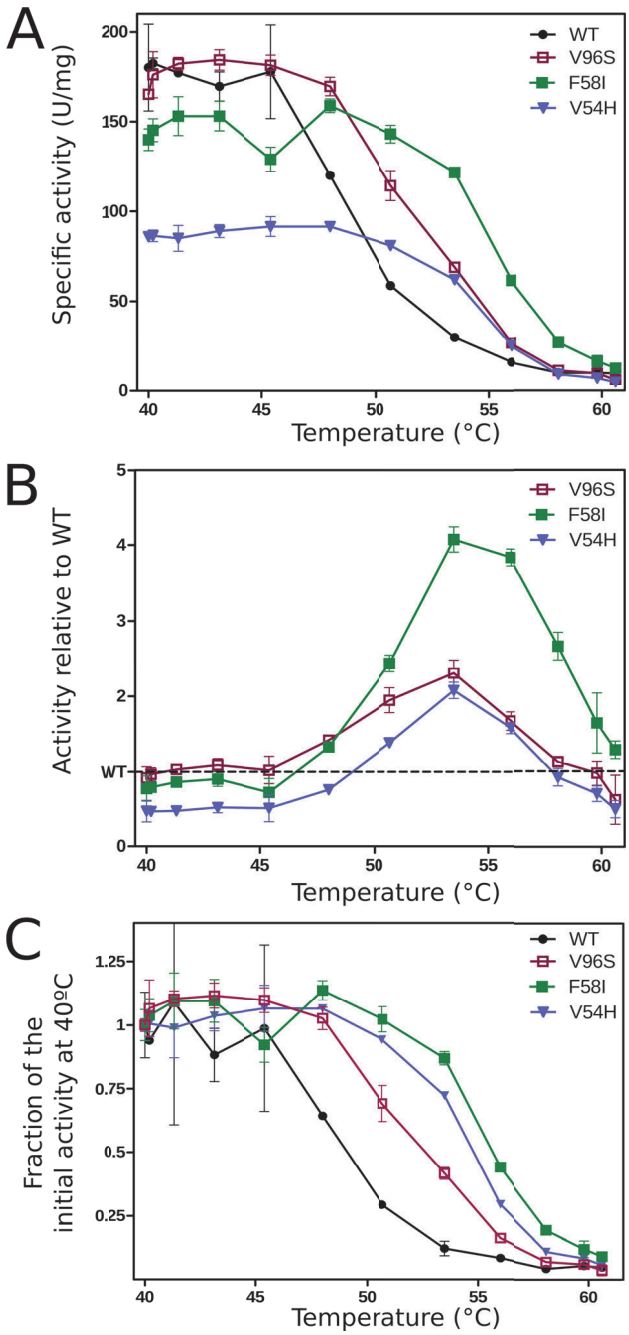

A
Specific activity (U/mg)
200
150
100
50
0
40
45
50
55
60
Temperature (°C)
WT
V96S
F58I
V54H
B
Activity relative to WT
5
4
3
2
WT
0
40
45
50
55
60
Temperature (°C)
V96S
F58I
V54H
C
Fraction of the initial activity at 40ºC
1.25
1
0.75
0.5
0.25
40
45
50
55
60
Temperature (°C)
WT
V96S
F58I
V54H

Figure 3

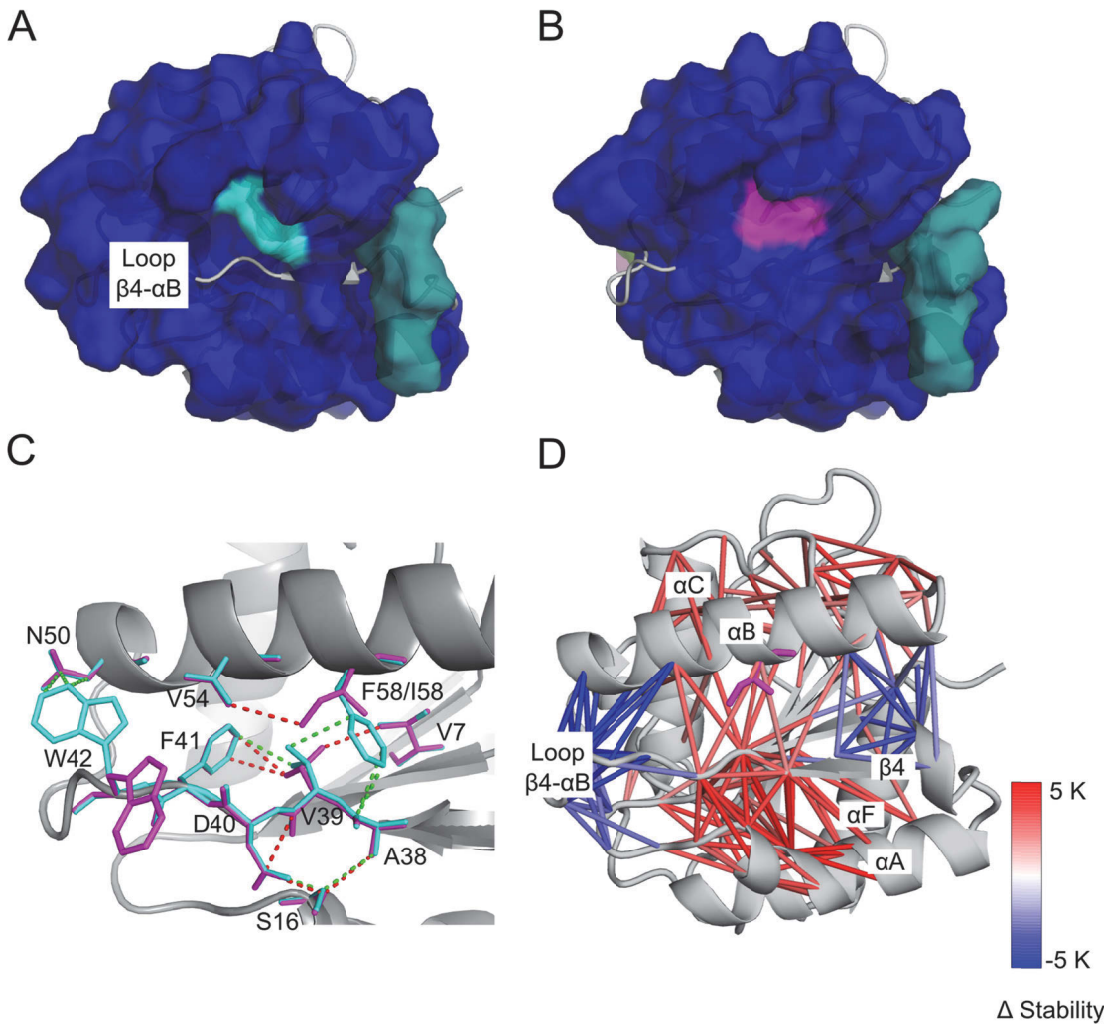

A
Loop
β4-αB
B
C
N50
V54
F58/I58
V7
F41
W42
D40
V39
A38
S16
D
αC
αB
Loop
β4-αB
β4
αF
αA
5 K
-5 K
Δ Stability

Figure 4

References

1. Demirjian, D. C.; Moris-Varas, F.; Cassidy, C. S. Enzymes from extremophiles. *Curr. Opin. Chem. Biol.* **2001,** 5, 144-151.
2. Van den Burg, B. Extremophiles as a source for novel enzymes. *Curr. Opin. Microbiol.* **2003,** 6, 213-218.
3. Polizzi, K. M.; Bommarius, A. S.; Broering, J. M.; Chaparro-Riggers, J. F. Stability of biocatalysts. *Curr. Opin. Chem. Biol.* **2007,** 11, 220-225.
4. Haki, G. D.; Rakshit, S. K. Developments in industrially important thermostable enzymes: a review. *Bioresour. Technol.* **2003,** 89, 17-34.
5. Eijsink, V. G. H.; Gaseidnes, S.; Borchert, T. V.; van den Burg, B. Directed evolution of enzyme stability. *Biomol. Eng.* **2005,** 22, 21-30.
6. Ahmad, S.; Kamal, M. Z.; Sankaranarayanan, R.; Rao, N. M. Thermostable Bacillus subtilis lipases: In vitro evolution and structural insight. *J. Mol. Biol.* **2008,** 381, 324-340.
7. Jochens, H.; Aerts, D.; Bornscheuer, U. T. Thermostabilization of an esterase by alignment-guided focussed directed evolution. *Protein Engineering Design & Selection* **2010,** 23, 903-909.
8. Zhang, Z. G.; Yi, Z. L.; Pei, X. Q.; Wu, Z. L. Improving the thermostability of Geobacillus stearothermophilus xylanase XT6 by directed evolution and site-directed mutagenesis. *Bioresour. Technol.* **2010,** 101, 9272-9278.
9. Chow, J. Y.; Xue, B.; Lee, K. H.; Tung, A.; Wu, L.; Robinson, R. C.; Yew, W. S. Directed Evolution of a Thermostable Quorum-quenching Lactonase from the Amidohydrolase Superfamily. *J. Biol. Chem.* **2010,** 285, 40911-40920.
10. Kotzia, G. A.; Labrou, N. E. Engineering thermal stability of L-asparaginase by in vitro directed evolution. *FEBS J.* **2009,** 276, 1750-1761.
11. Akbulut, N.; Ozturk, M. T.; Pijning, T.; Ozturk, S. I.; Gumusel, F. Improved activity and thermostability of Bacillus pumilus lipase by directed evolution. *J. Biotechnol.* **2013,** 164, 123-129.
12. Nakazawa, H.; Okada, K.; Onodera, T.; Ogasawara, W.; Okada, H.; Morikawa, Y. Directed evolution of endoglucanase III (Cel12A) from Trichoderma reesei. *Appl. Microbiol. Biotechnol.* **2009,** 83, 649-657.
13. Steffler, F.; Guterl, J. K.; Sieber, V. Improvement of thermostable aldehyde dehydrogenase by directed evolution for application in Synthetic Cascade Biomanufacturing. *Enzyme Microb. Technol.* **2013,** 53, 307-314.
14. Lehmann, M.; Wyss, M. Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Curr. Opin. Biotechnol.* **2001,** 12, 371-375.
15. Vazquez-Figueroa, E.; Chaparro-Riggers, J.; Bommarius, A. S. Development of a thermostable glucose dehydrogenase by a structure-guided consensus concept. *Chembiochem* **2007,** 8, 2295-2301.
16. Tian, J. A.; Wang, P.; Gao, S.; Chu, X. Y.; Wu, N. F.; Fan, Y. L. Enhanced thermostability of methyl parathion hydrolase from Ochrobactrum sp. M231 by rational engineering of a glycine to proline mutation. *FEBS J.* **2010,** 277, 4901-4908.
17. Leemhuis, H.; Rozeboom, H. J.; Dijkstra, B. W.; Dijkhuizen, L. Improved thermostability of Bacillus circulans cyclodextrin glycosyltransferase by the introduction of a salt bridge. *Proteins-Structure Function and Genetics* **2004,** 54, 128-134.
18. Kaneko, H.; Minagawa, H.; Shimada, J. Rational design of thermostable lactate oxidase by analyzing quaternary structure and prevention of deamidation. *Biotechnol Lett* **2005,** 27, 1777-1784.

19. Eijsink, V. G. H.; Bjørk, A.; Gåseidnes, S.; Sirevåg, R.; Synstad, B.; van den Burg, B.; Vriend, G. Rational engineering of enzyme stability. *J. Biotechnol.* **2004,** 113, 105-120.
20. Reetz, M. T.; Carballeira, J. D.; Vogel, A. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew. Chem. Int. Ed. Engl.* **2006,** 45, 7745-7751.
21. Radestock, S.; Gohlke, H. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.* **2008,** 8, 507-522.
22. Radestock, S.; Gohlke, H. Protein rigidity and thermophilic adaptation. *Proteins: Struct., Funct., Bioinf.* **2011,** 79, 1089-1108.
23. Rathi, P. C.; Radestock, S.; Gohlke, H. Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.* **2012,** 159, 135-144.
24. Kruger, D. M.; Rathi, P. C.; Pfleger, C.; Gohlke, H. CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo-)stability, and function. *Nucleic Acids Res.* **2013,** 41, W340-W348.
25. Pfleger, C.; Rathi, P. C.; Klein, D. L.; Radestock, S.; Gohlke, H. Constraint Network Analysis (CNA): A Python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. *J. Chem. Inf. Model.* **2013,** 53, 1007-1015.
26. Rathi, P. C.; Pfleger, C.; Fulle, S.; Klein, D. L.; Gohlke, H. Statics of biomacromolecules. In *Molecular Modeling*, Comba, P., Ed. Wiley-VCH: Weinheim, 2011; pp 281-299.
27. Hespenheide, B. M.; Jacobs, D. J.; Thorpe, M. F. Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. *J. Phys. Condens. Matter* **2004,** 16, S5055-S5064.
28. Jacobs, D. J.; Thorpe, M. F. Generic rigidity percolation: the pebble game. *Phys. Rev. Lett.* **1995,** 75, 4051-4054.
29. Jacobs, D. J.; Hendrickson, B. An algorithm for two-dimensional rigidity percolation: the pebble game. *J. Comp. Phys.* **1997,** 137, 346-365.
30. Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins: Struct., Funct., Bioinf.* **2001,** 44, 150-165.
31. Rader, A. J.; Hespenheide, B. M.; Kuhn, L. A.; Thorpe, M. F. Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci. U. S. A.* **2002,** 99, 3540-3545.
32. Razvi, A.; Scholtz, J. M. Lessons in stability from thermophilic proteins. *Protein Sci.* **2006,** 15, 1569-1578.
33. Pfleger, C.; Gohke, H. Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. *Structure* **2013,** 21, 1725-1734.
34. Jaeger, K. E.; Eggert, T. Lipases for biotechnology. *Curr. Opin. Biotechnol.* **2002,** 13, 390-397.
35. Jaeger, K. E.; Ransac, S.; Dijkstra, B. W.; Colson, C.; Vanheuvel, M.; Misset, O. Bacterial Lipases. *FEMS Microbiol. Rev.* **1994,** 15, 29-63.
36. Pfleger, C.; Radestock, S.; Schmidt, E.; Gohlke, H. Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comp. Chem.* **2013,** 34, 220-233.
37. Dahiyat, B. I.; Gordon, D. B.; Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci.* **1997,** 6, 1333-1337.
38. Mamonova, T.; Hespenheide, B.; Straub, R.; Thorpe, M. F.; Kurnikova, M. Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.* **2005,** 2, S137-S147.
39. Gohlke, H.; Kuhn, L. A.; Case, D. A. Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins: Struct., Funct., Bioinf.* **2004,** 56, 322-337.

40. Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* **2013,** 9, 3878-3888.
41. Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comp. Chem.* **2005,** 26, 1668-1688.
42. D.A. Case, T. A. D., T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, P.A. Kollman. *AMBER 11*, University of California, San Francisco., 2010.
43. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Bioinf.* **2006,** 65, 712-725.
44. Kawasaki, K.; Kondo, H.; Suzuki, M.; Ohgiya, S.; Tsuda, S. Alternate conformations observed in catalytic serine of Bacillus subtilis lipase determined at 1.3 A resolution. *Acta Crystallogr. D Biol. Crystallogr.* **2002,** 58, 1168-1174.
45. Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation1. *J. Mol. Biol.* **1999,** 285, 1735-1747.
46. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983,** 79, 926-935.
47. Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N.log (N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993,** 98, 10089-10092.
48. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comp. Phys.* **1977,** 23, 327-341.
49. Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; Heger, A.; Holm, L.; Sonnhammer, E. L. L.; Eddy, S. R.; Bateman, A.; Finn, R. D. The Pfam protein families database. *Nucleic Acids Res.* **2012,** 40, D290-D301.
50. Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Struct., Funct., Bioinf.* **2009,** 77, 778-795.
51. Melo, F.; Devos, D.; Depiereux, E.; Feytmans, E. ANOLEA: a www server to assess protein structures. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1997,** 5, 187-190.
52. Melo, F.; Feytmans, E. Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.* **1997,** 267, 207-222.
53. Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct., Funct., Bioinf.* **2004,** 55, 383-394.
54. Edelheit, O.; Hanukoglu, A.; Hanukoglu, I. Simple and efficient site-directed mutagenesis using two single-primer reactions in parallel to generate mutants for protein structure-function studies. *BMC Biotechnol.* **2009,** 9, 61.
55. Smith, P. K.; Krohn, R. I.; Hermanson, G. T.; Mallia, A. K.; Gartner, F. H.; Provenzano, M. D.; Fujimoto, E. K.; Goeke, N. M.; Olson, B. J.; Klenk, D. C. Measurement of protein using bicinchoninic acid. *Anal. Biochem.* **1985,** 150, 76-85.

56. Winkler, U. K.; Stuckmann, M. Glycogen, hyaluronate, and some other polysaccharides greatly enhance the formation of exolipase by Serratia marcescens. *J. Bacteriol.* **1979,** 138, 663-670.
57. Jaeger, K. E.; Dijkstra, B. W.; Reetz, M. T. Bacterial biocatalysts: Molecular biology, three-dimensional structures, and biotechnological applications of lipases. *Annu. Rev. Microbiol.* **1999,** 53, 315-+.
58. Beermann, B.; Guddorf, J.; Boehm, K.; Albers, A.; Kolkenbrock, S.; Fetzner, S.; Hinz, H. J. Stability, unfolding, and structural changes of cofactor-free 1H-3-hydroxy-4-oxoquinaldine 2,4-dioxygenase. *Biochemistry* **2007,** 46, 4241-4249.
59. Hung, H. C.; Chang, G. G. Multiple unfolding intermediates of human placental alkaline phosphatase in equilibrium urea denaturation. *Biophys. J.* **2001,** 81, 3456-3471.
60. Lehmann, M.; Pasamontes, L.; Lassen, S. F.; Wyss, M. The consensus concept for thermostability engineering of proteins. *Bba-Protein Struct M* **2000,** 1543, 408-415.
61. Lehmann, M.; Loch, C.; Middendorf, A.; Studer, D.; Lassen, S. F.; Pasamontes, L.; van Loon, A. P. G. M.; Wyss, M. The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng.* **2002,** 15, 403-411.
62. Steipe, B.; Schiller, B.; Pluckthun, A.; Steinbacher, S. Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. *J. Mol. Biol.* **1994,** 240, 188-192.
63. Vihinen, M.; Ollikka, P.; Niskanen, J.; Meyer, P.; Suominen, I.; Karp, M.; Holm, L.; Knowles, J.; Mantsala, P. Site-Directed Mutagenesis of a Thermostable Alpha-Amylase from Bacillus-Stearothermophilus - Putative Role of 3 Conserved Residues. *J. Biochem. (Tokyo)* **1990,** 107, 267-272.
64. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000,** 28, 235-242.
65. Welch, B. L. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika* **1947**, 28-35.
66. Tielen, P.; Kuhn, H.; Rosenau, F.; Jaeger, K. E.; Flemming, H. C.; Wingender, J. Interaction between extracellular lipase LipA and the polysaccharide alginate of Pseudomonas aeruginosa. *BMC Microbiol.* **2013,** 13.
67. Eijsink, V. G. H.; Vriend, G.; Vandervinne, B.; Hazes, B.; Vandenburg, B.; Venema, G. Effects of Changing the Interaction between Subdomains on the Thermostability of Bacillus Neutral Proteases. *Proteins-Structure Function and Genetics* **1992,** 14, 224-236.
68. Cho, A. R.; Yoo, S. K.; Kim, E. J. Cloning, sequencing and expression in Escherichia coli of a thermophilic lipase from Bacillus thermoleovorans ID-1. *FEMS Microbiol. Lett.* **2000,** 186, 235-238.
69. Jaenicke, R.; Böhm, G. The stability of proteins in extreme environments. *Curr. Opin. Struct. Biol.* **1998,** 8, 738-748.
70. Korkegian, A.; Black, M. E.; Baker, D.; Stoddard, B. L. Computational thermostabilization of an enzyme. *Science* **2005,** 308, 857-860.
71. Augustyniak, W.; Brzezinska, A. A.; Pijning, T.; Wienk, H.; Boelens, R.; Dijkstra, B. W.; Reetz, M. T. Biophysical characterization of mutants of Bacillus subtilis lipase evolved for thermostability: factors contributing to increased activity retention. *Protein Sci.* **2012,** 21, 487-497.
72. Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **2006,** 103, 5869-58674.

Publication VII – Supplementary Information

Application of rigidity theory to the thermostabilization of proteins

Prakash Chandra Rathi, Alexander Fulton, Karl-Erich Jaeger, and Holger Gohlke
Submitted manuscript, **2014.**
Contribution: 40%

Supplemental methods

Pairwise contact stability of neighboring residues

From a thermal unfolding trajectory, stability maps rc_{ij} can be calculated that characterize the local (residue-pairwise) rigidity of a protein by indicating for all residue pairs (*i*, *j*) the temperature at which the two residues stop being part of the same rigid cluster, i.e., the "rigid contact" between these two residues then breaks (see also section "Predicting protein thermostability by Constraint Network Analysis" in the main text and supplemental ref.[1] for details). When filtered such that only rigid contacts between residues that are at most 5 Å apart from each other (measured as the distance between the closest atom pair of the two residues) are considered, a neighbor stability map results. This map helps focusing on short-range rigid contacts that can be directly modulated by mutagenesis with the aim to stabilize them for improving the overall stability of a protein.

Here we use neighbor stability maps to analyze the (local) effect of mutations on the stability of rigid contacts of neighboring residues (Figure S5). The increase in the strength of rigid contacts is calculated as the average over differences in rc_{ij} of the variant *versus* WT for all neighboring residue pairs (lower triangles in **Figure S5**). The increase in the strength is measured in K.

Supplemental tables

Table S1. Computing times for weak spot identification, site saturation mutagenesis, and screening for increased structural rigidity

Step[a]	Time required	Comment
I) MD simulation	~78 h	100 ns long MD simulation on a single GPU
II) Thermal unfolding simulation	4 h and 35 min	Structural ensemble of 2000 structures run on one CPU core
III) Weak spot detection	2 h	Manual identification by visual inspection
IV) Filtering weak spots	Instantaneous	Highly conserved weak spots were discarded
V) Variant modeling by SCWRL	< 1 s	For a single mutation
VI) ENT^{FNC} run	~ 1h and 10 min	For a single mutation applying 1000 network topologies

[a] Steps are according to Figure 1 in the main text.

Table S2. *Bs*LipA variants and mutagenesis primer sequences

Mutation	Forward primer sequence	Reverse Primer sequence
G104I	GTCGTGACGCTTGGCatcGCGAACCGTTTGACG	CGTCAAACGGTTCGCgatGCCAAGCGTCACGAC
G104L	GTCGTGACGCTTGGCctgGCGAACCGTTTGACG	CGTCAAACGGTTCGCcagGCCAAGCGTCACGAC
L55F	AACAATGGACCGGTAttcTCACGATTTGTGCAA	TTGCACAAATCGTGAgaaTACCGGTCCATTGTT
V59F	GTATTATCACGATTTttcCAAAAGGTTTTAGAT	ATCTAAAACCTTTTGgaaAAATCGTGATAATAC
I122W	CAGATCCAAATCAAtggATTTTATACACATCC	GGATGTGTATAAAATccaTTGATTTGGATCTG
L160H	GGACACATCGGCCTTcatTACAGCAGCCAAGTC	GACTTGGCTGCTGTAatgAAGGCCGATGTGTCC
N51F	GGCACAAATTATAACttcGGACCGGTATTATCA	TGATAATACCGGTCCgaaGTTATAATTTGTGCC
G52M	CACAAATTATAACAATatgCCGGTATTATCACGA	TCGTGATAATACCGGcatATTGTTATAATTTGTG
V54H	TATAACAATGGACCGcatTTATCACGATTTGTG	CACAAATCGTGATAAatgCGGTCCATTGTTATA
F58I	CCGGTATTATCACGAatcGTGCAAAAGGTTTTA	TAAAACCTTTTGCACgatTCGTGATAATACCGG
I87W	GAACACACTTTACTACtggAAAAATCTGGACGGC	GCCGTCCAGATTTTTccaGTAGTAAAGTGTGTTC
V96S	GACGGCGGAAATAAAagcGCAAACGTCGTGACG	CGTCACGACGTTTGCgctTTTATTTCCGCCGTC

Table S3. Kinetic parameters of *Bs*LipA variants and wildtype

Variant	K_M[a] (μM)	k_{cat} (s^{-1})	k_{cat} / K_M ($\mu M^{-1} * s^{-1}$)
Wildtype	34.72 ± 6.49	926.40 ± 38.81	26.68 ± 6.10
V54H	40.02 ± 8.64	784.60 ± 38.97	19.51 ± 5.18
F58I	36.71 ± 7.83	690.50 ± 33.35	18.80 ± 4.91
V96S	32.30 ± 7.39	785.00 ± 39.73	24.30 ± 6.79

[a] Kinetic parameters were derived from experiments conducted at 40°C using *p*NPD as substrate.

Supplemental figure

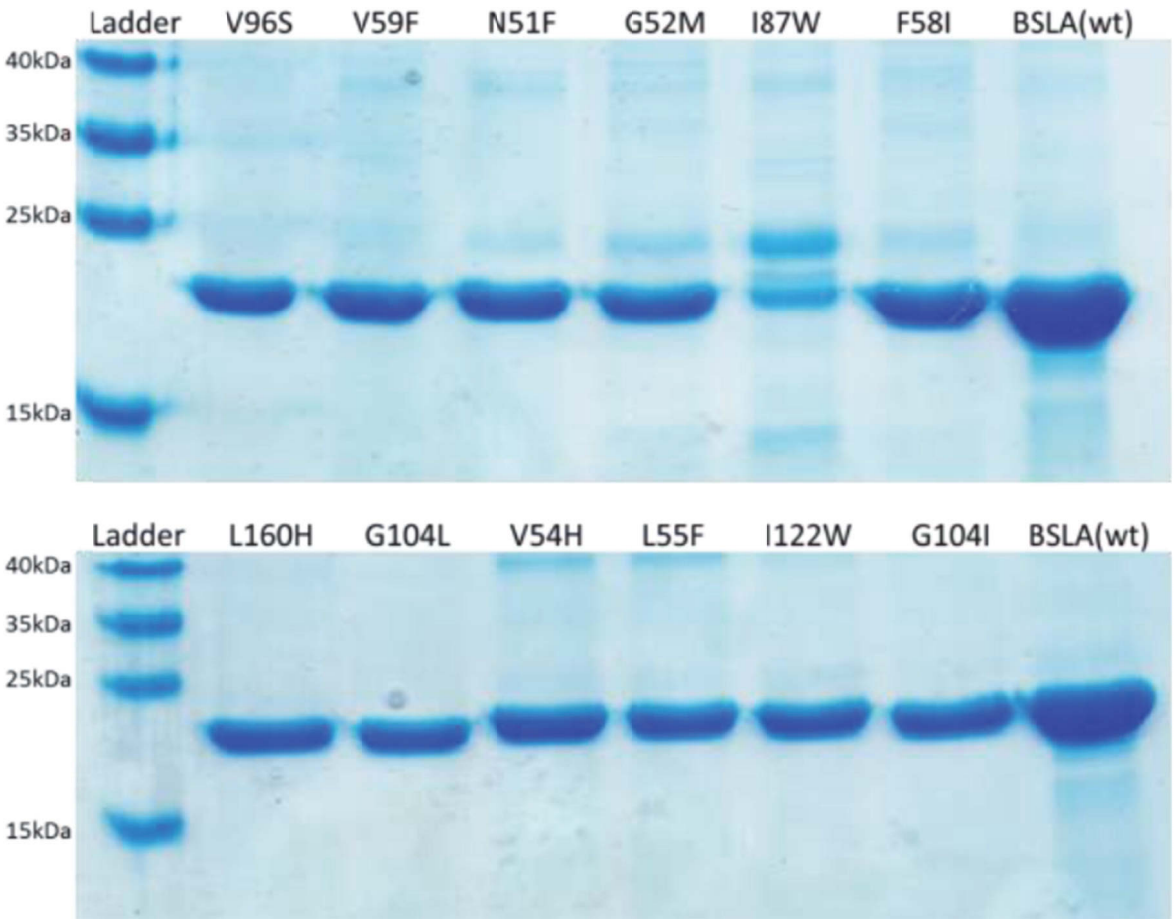


Figure S1. SDS-PAGE of all mutants and the WT that were purified using a N-terminal his-tag and the Ni-NTA purification method. After purification the samples were desalted and stored in 10mM Glycine buffer pH11. The mutant I87W was in all biological replicates not expressed properly and could only be purified in small amounts.

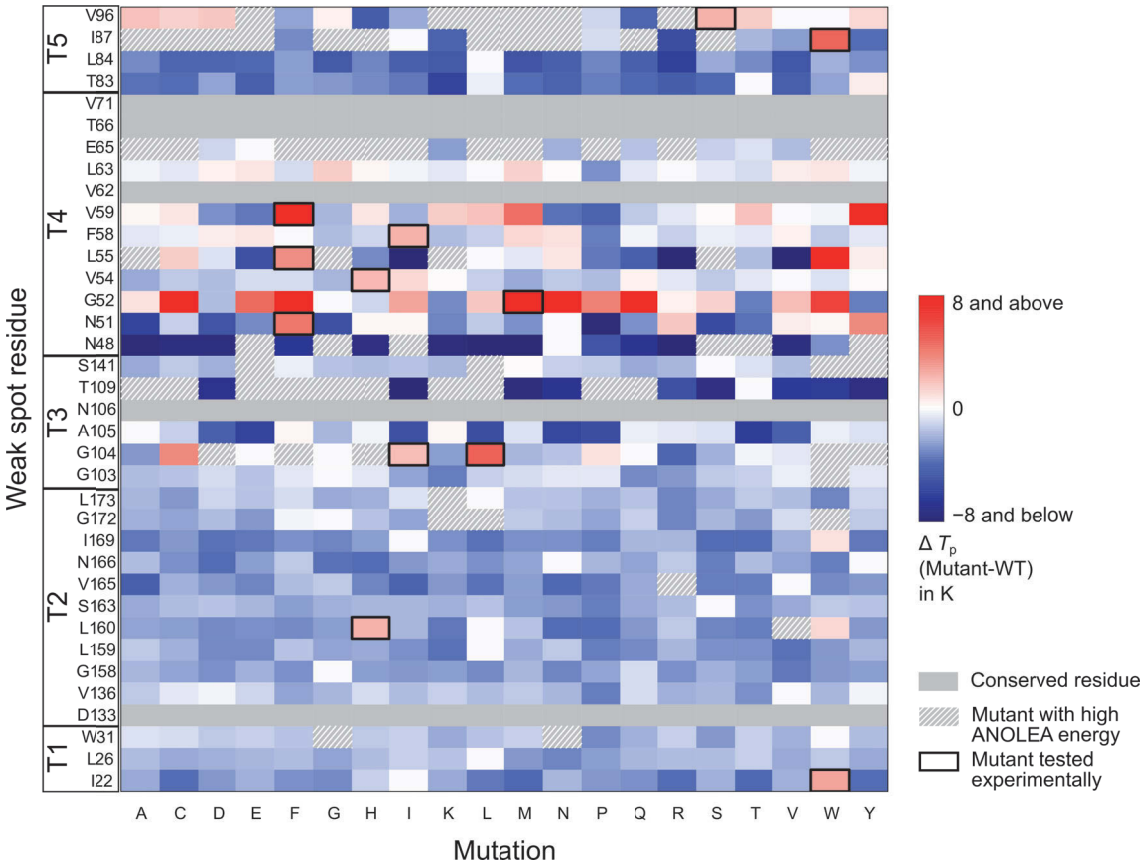


Figure S2. Map of $\Delta T_p = T_p$ (mutant) − T_p (wt) values for each mutation (abscissa) at each weak spot residue (ordinate) identified by CNA in WT *Bs*LipA. Weak spot residues are grouped by the major transition at which they are identified (Table 1, Figure 2 in the main text). Weak spot residues that are highly conserved in the multiple sequence alignment of the lipase family (see the main text) are shown in gray. Mutations are colored according to thermostabilizing (red) or thermodestabilizing (blue) effects. Mutations that led to energetically unfavorable structures as calculated by the ANOLEA server are shown as white stripes on gray color. Experimentally tested mutants are marked by a black box.

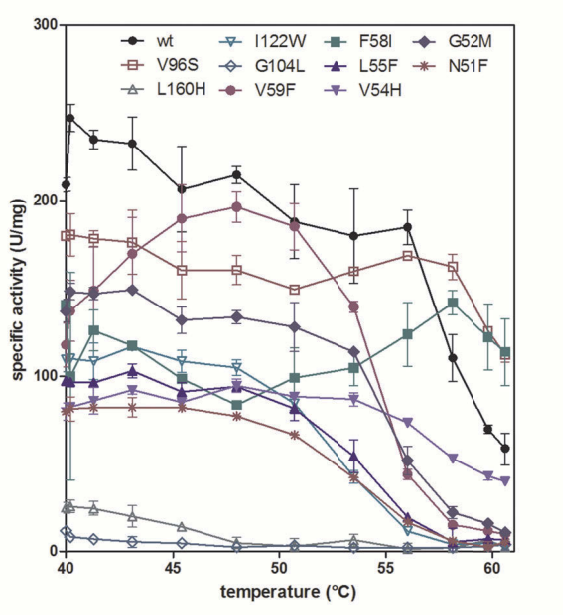


Figure S3. Specific activities of *Bs*LipA variants between temperatures 40 and 60ºC. The *Bs*LipA variants and the *p*NPP substrate solutions were incubated for 5 min at the indicated temperatures, and then the activity was measured at these temperatures. Mutants G104I and I87W were inactive at these temperatures.

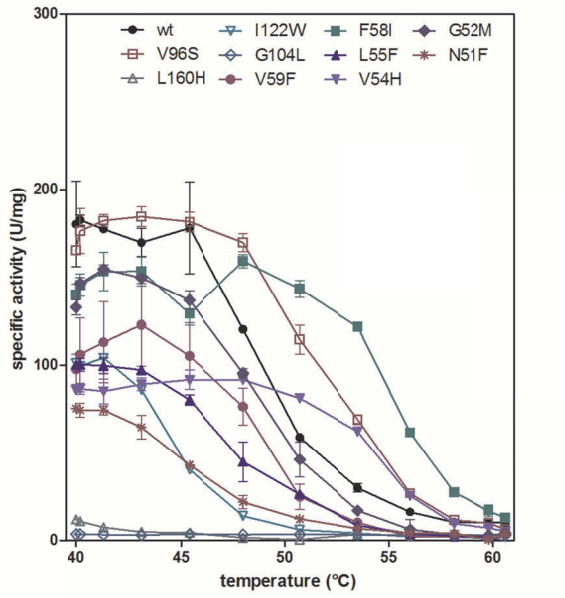


Figure S4. Specific activities of *Bs*LipA variants between temperatures 40 and 60ºC. The *Bs*LipA variants and the *p*NPP substrate solutions were incubated for 30 min at the indicated temperatures, and then the activity was measured at these temperatures. Mutants G104I, G104L, and I87W were inactive at these temperatures.

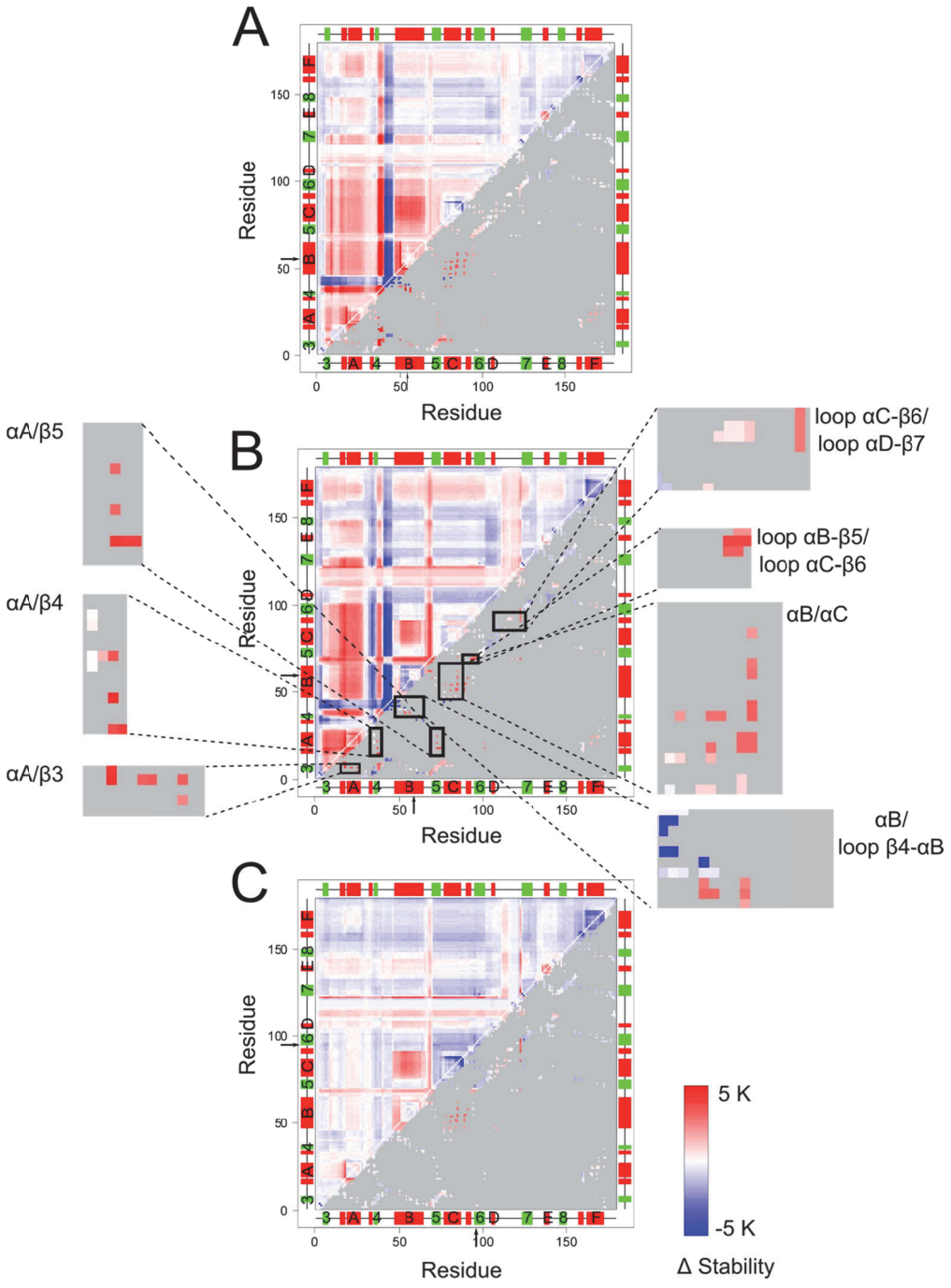


Figure S5. Differences in the stability of rigid contacts between wild type and variants of *Bs*LipA: V54H (A), F58I (B), V96S (C). In the map, a red (blue) color indicates that a rigid contact in the variant is more (less) stable than in the WT (see color scale). The upper triangle shows differences in the stability values for all residue pairs; the lower triangle shows differences in the stability values only for residue pairs that are within 5 Å of each other, with values for all other residue pairs colored grey. Secondary structure elements as computed by the DSSP program [2, 3] are indicated on both abscissa and ordinate and are labeled: α-helix (red rectangle), β-strand (green rectangle), loop (black line). Mutated residues are indicated by arrows. Blow-ups for secondary structure pairs of F58I described in the main text are shown.

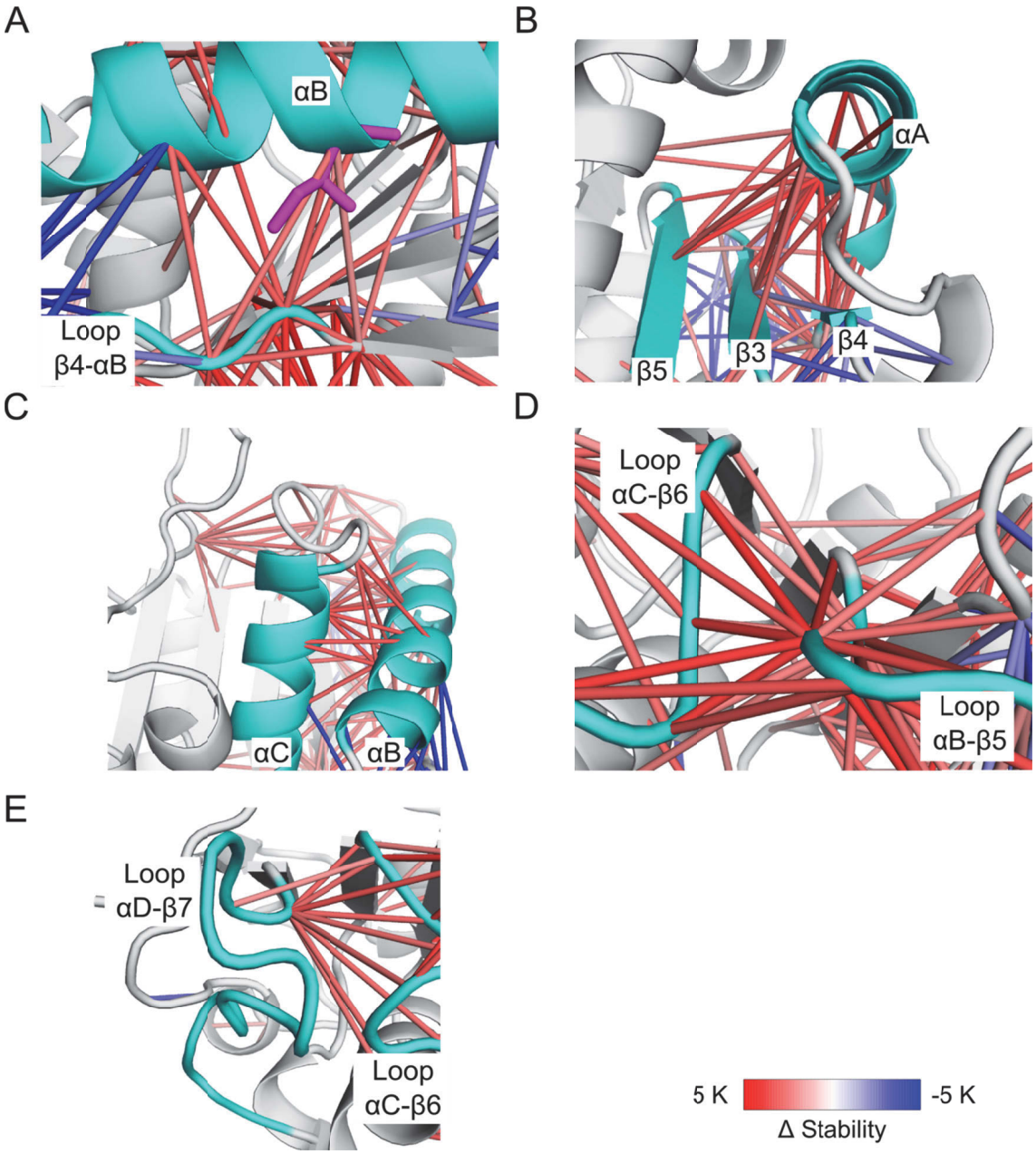


Figure S6. Blow-ups of Figure 4D in the main text showing differences in the stability of "rigid contacts" between variant F58I and WT depicted on the variant structure: αB/Loop β4-αB (A); αA/β strands 3,4,5 (B); αB/αC (C); loop αB-β5/loop αC-β6 (D); loop αC-β6/loop αD-β7 (E). Two residues form a "rigid contact" if they belong to one rigid cluster. A red (blue) stick connecting C_α atoms of two residues indicates that a rigid contact in the variant is more (less) stable than in the WT (see color scale). Only those contacts of variant F58I that are stabilized or destabilized by ≥ 2 K are shown for clarity; for the same reason, contacts between two residues of the same secondary structure element are not shown. Note that the blow-ups shown here are related to the blow-ups shown in Figure S5B.

Supplemental references

1. Pfleger, C.; Radestock, S.; Schmidt, E.; Gohlke, H. Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comp. Chem.* **2013,** 34, 220-233.
2. Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983,** 22, 2577-2637.
3. Joosten, R. P.; Beek, T. A. H. T.; Krieger, E.; Hekkelman, M. L.; Hooft, R. W. W.; Schneider, R.; Sander, C.; Vriend, G. A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **2011,** 39, D411-D419.

Publication VIII

Statics of biomacromolecules

Prakash Chandra Rathi, Christopher Pfleger, Simone Fulle, Doris L. Klein,
and Holger Gohlke
in: Modeling of Molecular Properties , P. Comba (ed.), S. 281-299,
Wiley-VCH, Weinheim, **2011.**
Contribution: 30%

18
Statics of Biomacromolecules

Prakash C. Rathi, Christopher Pfleger, Simone Fulle, Doris L. Klein, and Holger Gohlke

18.1
Introduction

Proteins, DNAs, and RNAs are the ultimate functional units that carry out biological functions by interacting with other biomacromolecules or small molecules [1]. Almost all of these interactions come along with a certain degree of conformational adaptation to attain complementarity of the binding partners. This structural flexibility of biomacromolecules has been associated with molecular recognition as well as with catalysis [2, 3]. The first binding model for enzymes, the "lock and key" model, assumed the enzymes to have a rigid catalytic site [4]; in contrast, the "induced fit" model highlighted the importance of flexibility in enzyme action [5]. An extension to the induced fit model came in the form of the "conformational selection and population shift" model, which states that biomacromolecules are in a state of continuous conformational fluctuation; a binding partner binds preferentially to one of the conformations, which shifts the conformational ensemble towards that state [6, 7]. Overall, binding to biomacromolecules may involve one or a combination of these phenomena. Apart from implications for function, flexibility is also linked to the structural stability of biomacromolecules [8]. In particular, it has been observed that thermophilic proteins are in general more rigid than their mesophilic homologs in order to preserve the structural integrity at higher temperature [9]. Hence, knowing *what can move, and how* in a biomacromolecule is instrumental in understanding the molecule's flexibility/stability and, thus, its function. The flexibility and mobility of biomacromolecules have been frequently investigated using X-ray crystallography, cryo-electron microscopy, single-molecule fluorescence, and nuclear magnetic resonance (NMR) spectroscopy [10–13]. Crystallographic B-factors, atomic fluctuations derived from NMR structural ensembles, NMR relaxation measurements, and residual dipolar couplings are the main source of information about flexibility and mobility of biomacromolecules [14, 15].

Alternatively, computational methods, such as molecular dynamics (MD) simulation or normal mode analysis, are widely used to obtain deeper insights into the dynamics of biomacromolecules. Nevertheless, MD simulation is still too

Modeling of Molecular Properties, First Edition. Edited by Peter Comba.
© 2011 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2011 by Wiley-VCH Verlag GmbH & Co. KGaA.

time-consuming to investigate conformational transitions that occur on a millisecond time scale on a routine basis [16, 17], while normal mode analysis generally tends to describe conformational changes in the vicinity of the starting structure only [18–20]. As yet another alternative, a fast, graph theory-based approach for characterizing the biomacromolecular flexibility and its opposite, rigidity, will be discussed in this chapter. This approach has been implemented into the FIRST program (*F*loppy *I*nclusions and *R*igid *S*ubstructure *T*opography), and allows the determination of biomacromolecular flexibility from a single input structure [21]. It should be noted that *flexibility* and *rigidity* are *static* properties – that is, a rigidity analysis determines those parts of a molecule that *can potentially move*, but says nothing about the direction or amplitude of a motion [22]. The approach has already been applied in several areas of computational biomacromolecular research, including the sampling of biomacromolecular conformational space [23–26], analyzing structural determinants of thermostability [27, 28], identifying folding cores of proteins [29, 30], assessing complex structural stability [31, 32], linking flexibility and function [33], finding putative binding sites [34], understanding allostery [35, 36], investigating large biomacromolecules such as the ribosome [35], and predicting thermodynamic properties [37].

18.2
Rigidity Theory and Analysis

18.2.1
Introduction to Rigidity Theory

The quest to identify rigid and flexible regions in networks (graphs) of sites (vertices) and constraints (edges) dates back many years. In 1864, Maxwell proposed an approximate method to calculate the number of floppy modes F in a d-dimensional *generic* network – that is, a network without any symmetries such as collinear constraints [38]. The term "floppy modes" denotes (independent) internal degrees of freedom in which the sites of the network can move without violating any of the constraints. For a network with N sites lacking any constraint, $F = dN - d(d+1)/2$, with the subtrahend denoting the global degrees of freedom (overall translation and rotation) of the d-dimensional network. Each added constraint, if independent of all other constraints, removes one floppy mode. Thus, if all constraints in the network were independent, as assumed by Maxwell, the number of floppy modes (F_{mxw}) in a network with N_c constraints can be calculated by Eq. (18.1):

$$F_{mxw} = dN - N_c - d(d+1)/2 \tag{18.1}$$

Usually, this underestimates F because, in reality, not all constraints are independent: if a constraint is placed between two already mutually rigid sites, it does not decrease the number of floppy modes any further and, thus, is a redundant constraint. Taking into account the number of redundant constraints N_r then leads to Eq. (18.2):

$$F = dN - (N_c - N_r) - d(d+1)/2 \tag{18.2}$$

Incorporating a redundant constraint introduces stress in the network; network regions with such constraints are thus called *over-constrained* or *stressed*. In contrast, a region with fewer constraints than internal degrees of freedom is called *under-constrained*. Finally, in a region with as many independent constraints as internal degrees of freedom, $F = 0$; this region is called *isostatically rigid*.

In 1970, a theorem by Laman [39] had a major impact in that it allowed the precise determination of the degrees of freedom in a two-dimensional (2-D) network, even in the presence of redundant constraints, by applying constraint counting to all subgraphs within the network. As such, a generic 2-D network does not have a redundant constraint if and only if for all subgraphs of size $n \geq 2$, the number of constraints in the subgraph $N_{cs} \leq 2n - 3$. By applying Laman's theorem, a network can be decomposed into rigid regions and flexible links in between. This constraint counting can be extended to a certain subtype of three-dimensional (3-D) networks with a molecule-like character – so-called "bond-bending networks" or "molecular frameworks" [40, 41]. In these networks, bond angles (distances between second-nearest neighbor sites) are constrained in addition to the bond lengths (distances between first-nearest neighbor sites), which makes them particularly applicable to biomacromolecules.

For both the 2-D and 3-D bond-bending networks, combinatorial algorithms – termed *pebble games* – were devised that allow the network flexibility and rigidity to be determined according to Eq. (18.2) [10, 42–44]. These algorithms have been implemented in ProFlex [45] and in early versions of the FIRST [46] software package. As an example, bond-bending networks of two molecules are depicted in Figure 18.1. In both networks, fixed bond lengths and angles are modeled as distance constraints between nearest and next-nearest neighbor atoms. Free rotation about the bond between atom *1* and atom *2* in molecule M1 results in one floppy mode and two rigid clusters of three atoms each (Figure 18.1a–c). A double bond is modeled by placing an additional distance constraint between third-nearest neighbors (Figure 18.1e), which results in molecule M2 being a single rigid cluster (Figure 18.1d–f).

A more recent implementation of FIRST uses a *body-and-bar* representation of 3-D networks, where every atom is considered as a rigid body having six degrees of freedom [47]. Any number of bars between one and six can be placed between two such bodies, and every such bar removes one degree of freedom. The number of floppy modes is then computed according to Eq. (18.3):

$$F = 6N - N_{ibar} - 6 \tag{18.3}$$

where N_{ibar} represents the total number of independent bars in the network. In the *body-and-bar* network representation, covalent single bonds are modeled as five bars between two atoms leaving one degree of freedom, the dihedral rotation (Figure 18.1c). Double bonds are modeled with six bars locking the rotation (Figure 18.1f). Apart from algorithmic advantages over the bond-bending representation, the body-and-bar representation also has a methodological advantage which lies in the fact that constraints can be modeled semi-quantitatively: strong bonds are modeled with more bars, whereas weaker bonds are modeled with fewer bars [47].

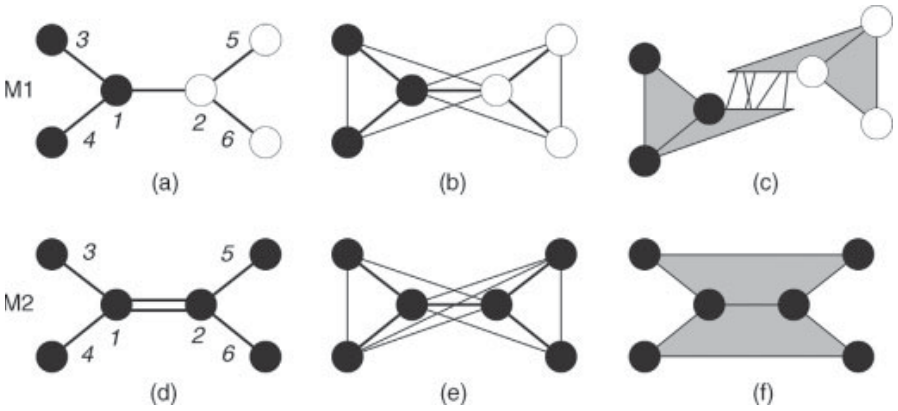


Figure 18.1 Network representations of molecules M1 (a) and M2 (d). In the *bond-bending* networks (b, e), the double bond in M2 is modeled by placing an additional constraint between atom *4* and *5*. In the *body-and-bar* networks (c, f), the bond between atom *1* and *2* in M1 is modeled by five bars, whereas six bars are used in M2 for locking the rotation. The atom colors represent the rigid clusters to which they belong: M1 has two rigid clusters and one flexible joint, whereas all atoms of M2 belong to a single rigid cluster. Figure adapted from Ref. [101].

18.2.2
Modeling Biomacromolecules as Constraint Networks

Biomacromolecules can be effectively represented either as *bond-bending* or *body-and-bar* networks. Here, it is described how biomacromolecules can be represented by the latter representation. The atoms of the biomacromolecules are modeled as bodies, while covalent and noncovalent bonds are modeled as bars. A covalent bond is generally modeled as five bars, allowing for the dihedral rotation about it. Peptide and double bonds are modeled with six bars, disallowing any bond rotation. Considering that the mechanical rigidity of a biomacromolecule is largely determined by noncovalent interactions, there is also a need to include hydrogen bonds, salt bridges, and hydrophobic interactions as constraints in the network. Stronger interactions such as hydrogen bonds (and salt bridges) and hydrophobic interactions are modeled as five bars and two bars, respectively [47]. Weaker interactions such as van der Waals interactions are not modeled as constraints. Figure 18.2 shows a network representation for hen egg-white lysozyme (PDB code: 1vb1), which is then decomposed into rigid clusters and flexible joints by rigidity analysis using the FIRST software.

18.2.3
Simulating Folded–Unfolded Transitions in Biomacromolecules

By consecutively removing constraints from a network, it is possible to simulate the melting of the network and to identify a phase transition where the network switches from an overall rigid state to a floppy one. Phrased differently, at this so-called rigidity

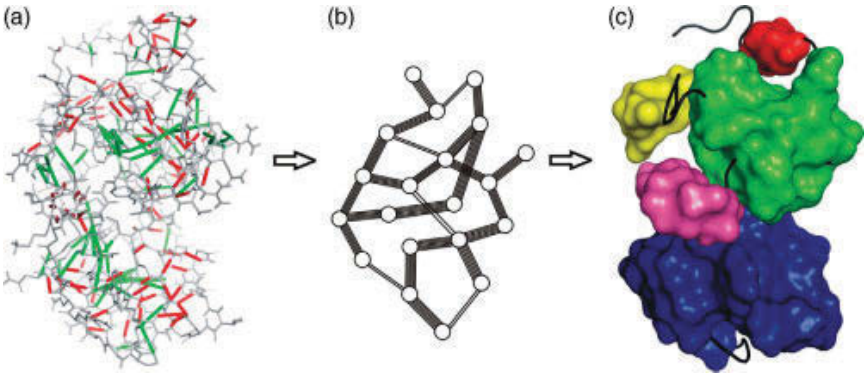


Figure 18.2 Workflow of rigidity analysis of biomacromolecules showing hen egg-white lysozyme as an example. A PDB structure with added hydrogen atoms is used as an input (a) from which a *body-and-bar* network is modeled (b). Covalent bonds are depicted in gray, hydrogen bonds in red, and hydrophobic interactions in green (a). Each bond is identified either as a part of rigid region or a flexible joint, resulting in a rigid cluster decomposition (c) where each rigid cluster has a unique color.

percolation threshold, the network loses its ability to transmit stress – that is, rigidity ceases to percolate through the network. Crosslinked covalently bonded 3-D network glasses have been thoroughly studied in that sense, both computationally and experimentally [48–50]. It has been observed that the phase transition for network glasses takes place at a mean coordination value of 2.385, and is continuous or of second order. However, it has also been found that the phase transition can become first order for self-organized networks where locally stressed regions or small rings of bonds are suppressed [49]. Biomacromolecular networks can be considered similar to network glasses, and the melting of the network can be realized by consecutively removing noncovalent bonds, which is equivalent to a thermal unfolding of the biomacromolecule. However, the percolation behavior of protein networks is usually more complex, and multiple transitions can be observed. This is due to the fact that protein structures are modular because they are assembled from secondary structure elements, subdomains, and domains. These modules often spontaneously break away from the giant cluster as a whole, giving rise to multiple transitions.

18.2.4
Constraint Network Analysis

The Constraint Network Analysis (CNA) program package has been developed by the present authors' group, with the aim of analyzing structural features of biomacromolecules that are important for the molecule's stability. CNA functions as a front-end to the FIRST software and allows: (i) the setting up of a variety of constraint network representations for rigidity analysis (see also below); (ii) processing of the results obtained from FIRST; and (iii) calculating the different indices for characterizing biomacromolecular stability, both globally and locally (see Section 18.2.5).

CNA can be used to carry out thermal unfolding simulations by gradually removing noncovalent constraints from the initial network representation (see above) [27, 29, 51–53]. That is, for a given network state $s = f(T)$, hydrogen bonds (including salt bridges) with an energy $E_{HB} > E_{cut,s}$ are removed from the network [54]. This follows the idea that stronger hydrogen bonds will break at higher temperatures than weaker ones. To convert the original, geometry-based hydrogen bond energy scale E_{HB} [54] into a temperature scale T, Radestock and Gohlke proposed a simple linear fit by comparing computed phase-transition temperatures for pairs of homologous mesophilic and thermophilic proteins with experimental melting temperatures [27]. The number of hydrophobic contacts is kept constant during the thermal unfolding, because the strength of hydrophobic interactions remains constant or even increases with increasing temperature [55]. Finally, a rigidity analysis is performed on each constraint network state s.

In principle, CNA can be performed on a single 3-D structure of a biomacromolecule. However, different conformations of a protein structure may lead to different results of the rigidity analysis, as observed by the present authors [32] and others [56]. This sensitivity arises from the facts that: (i) proteins are generally marginally stable [57]; and (ii) different protein conformations can lead to different numbers of constraints being included based on geometric criteria. Consequently, as the protein network is already close to the rigidity percolation threshold [due to point (i) above], a few constraints more or less [due to point (ii)] can result in the network being largely rigid or already floppy. To overcome this problem, CNA allows the use of an ensemble of constraint networks rather than a single structure. There are two ways in which these ensembles can be generated:

- Conformations extracted from a MD simulation-derived trajectory can be individually subjected to CNA, and the results are then averaged over the whole ensemble. This approach has the advantage that CNA is based on a thermodynamic *ensemble of conformations*. As a downside, a computationally expensive MD simulation is required to generate the input.
- An *ensemble of network topologies* can be generated by fluctuating noncovalent constraints in a network derived from a single structure. The fluctuating noncovalent constraints are realized by modulating the stability of the constraints by white noise. This is supposed to mimic variations in the constraint stability due to the wiggling of atoms. As an advantage, this approach does not require computationally expensive MD simulations. As a downside, the thus-generated networks might not be very different from the network of the input structure. Preliminary results have shown that CNA results derived in this way are more consistent with those obtained from ensembles of experimentally derived structures than if a single input structure is used instead (C. Pfleger, H. Gohlke, unpublished results). Finally, as a further advantage over analyzing a single structure, either approach allows to determine the significance of CNA results by means of statistical testing.

It should be noted that the distance constraint model (DCM) [28, 58] also relies on ensembles of constraint topologies, which are differently generated, however. Here,

mean-field probabilities of hydrogen bonds and torsion constraints are used for Monte Carlo sampling to generate such an ensemble, however, assuming that the atom positions of the input structure are unique.

18.2.5 Indices to Characterize Flexibility and Rigidity

CNA can be used to calculate several indices to characterize the global and local flexibility/rigidity of a biomacromolecule. All of these indices share the common feature that they are derived by analyzing a thermal unfolding simulation of a constraint network.

18.2.5.1 Global Indices

In order to describe the global percolation behavior of a network, the *microstructure* of the network – that is, the properties of the set of clusters generated by the bond dilution process – can be analyzed [52]. For the *rigidity order parameter* (P_∞), the fraction of the network belonging to the percolating (giant) rigid cluster is chosen as an order parameter. In other words, P_∞ denotes the probability that an atom belongs to the giant cluster and is zero in the floppy phase. Thus, monitoring the decay of the giant cluster by P_∞ provides a global and intuitive description of the rigidity within the protein structure during thermal unfolding (Figure 18.3a). Notably, P_∞ curves of proteins are similar to P_∞ curves observed for network models of glasses and amorphous solids [59, 60]. Likewise, homologous proteins have P_∞ curves of very similar shape (Figure 18.3a) [27, 60].

The *cluster configuration entropy* (H) is another global index, which has been introduced by Andraud *et al.* as a morphological descriptor for heterogeneous materials [61]. H has been adapted from Shannon's information theory and, thus, is a measure of the degree of disorder in the realization of a given state. As long as the giant cluster dominates the system, H is low because of the limited number of possible ways to configure a system with a very large cluster (Figure 18.3b). At the rigidity percolation threshold, H jumps as the network is now in a partially flexible state with many ways to configure a system consisting of (many) small clusters. H has already been successfully applied to analyze unfolding transitions in proteins [27, 51].

18.2.5.2 Local Indices

Local flexibility/rigidity indices characterize the network flexibility/rigidity down to the bond level. The *percolation index* (p) is a local analogon to the rigidity order parameter P_∞. As such, the index is derived for each covalent bond by monitoring the hydrogen bond energy cut-off E_{cut} during a thermal unfolding simulation at which this bond segregates from the *giant* cluster. Thus, the percolation index can be applied to locally monitor the percolation behavior of protein structures. The *rigidity index* (r) is a generalization of the percolation index p. It is derived for each covalent bond in the network by monitoring the hydrogen bond energy cut-off E_{cut} during a

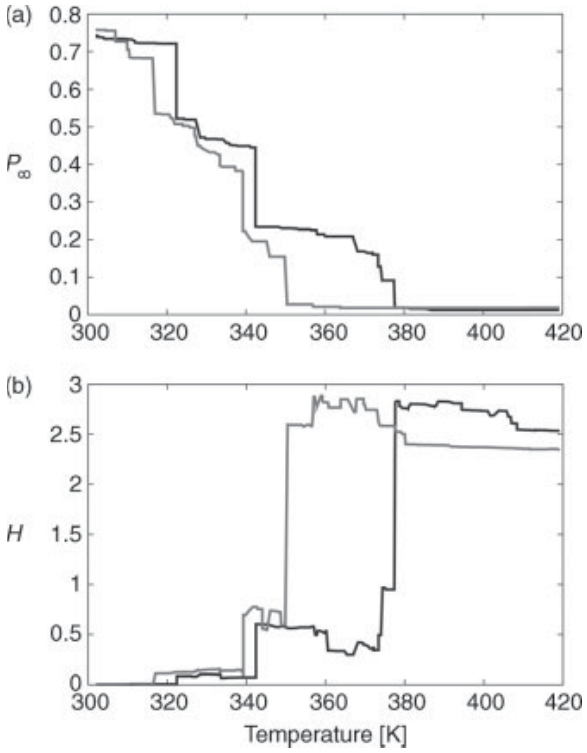


Figure 18.3 (a) Rigidity order parameter (P_∞) and (b) cluster configuration entropy (H) plotted versus temperature for thermolysin-like protease of the mesophilic organism *B. cereus* (gray line) and thermolysin of the thermophilic organism *B. thermoproteolyticus* (black line).

thermal unfolding simulation at which this bond switches from rigid to flexible. Phrased differently, this index monitors when a bond segregates from *any* rigid cluster.

Stability maps have been introduced as a third local index by Radestock and Gohlke [52]. A stability map is somewhat a 2-D generalization of the rigidity index. To derive a stability map, "rigid contacts" between two residues, represented by their C_α atoms, are identified. A rigid contact exists if two residues belong to the same rigid cluster. During a thermal unfolding simulation, stability maps are then constructed in that, for each residue pair, E_{cut} is identified at which a rigid contact between two residues is lost. In that way, a contact's stability relates to the microscopic stability in the network and, taken together, the microscopic stabilities of all residue–residue contacts result in a stability map. Thus, stability maps denote the distribution of rigidity and flexibility within the system, they identify regions that are flexibly or rigidly correlated across the structure, and they provide information on *how these properties change with temperature*. Stability maps are comparable to cooperativity correlation plots generated by DCM [28, 58], in that these plots identify regions that are correlated across the entire ensemble of constraint topologies generated *at a fixed temperature*.

18.3
Application of Rigidity Analysis to Biomacromolecules

18.3.1
Coarse-Graining for Simulating Conformational Transitions in Proteins

Specific functions of biomacromolecules often require conformational transitions. Such conformational changes range from the movements of single side chains and loops to large-scale domain motions. The ability to describe and predict conformational changes of biomacromolecules is not only important for understanding their impact on biological function, but will also have implications for modeling (biomacro)molecular complex formation [62] and in structure-based drug design [63]. As modeling (large-scale) conformational transitions of biomacromolecules is computationally challenging, coarse-grained simulation methods have emerged as efficient alternatives [64]. Decomposing the biomacromolecule into rigid clusters and flexible links in between by rigidity analysis provides a natural coarse-graining [22], and has already been used in several simulation methods [23, 25, 26, 65, 66].

Notably, a three-step approach has been developed for the multiscale modeling of biomacromolecular conformational changes that also relies on such a coarse-graining in the first step [23, 66]. In the second step, the dynamic properties of the biomacromolecule are revealed by the rotations–translations of blocks (RTB) approach [67], using an elastic network model representation of the coarse-grained protein (termed *R*igid *C*luster *N*ormal *M*ode *A*nalysis; RCNMA) [23]. Thus, in this step, only rigid body motions are allowed for rigid clusters, while links between them are treated as fully flexible. In the final step, the recently introduced idea of constrained geometric simulations of diffusive motions in proteins [25] is extended. New macromolecule conformers are generated by deforming the structure along low-energy normal mode directions predicted by RCNMA plus random direction components. The generated structures are then iteratively corrected regarding steric clashes or constraint violations; this module is termed NMSim [66]. Constraints to be satisfied include torsions of the main-chain and side-chains, distances and angles due to noncovalent interactions such as hydrogen bonds or hydrophobic interactions, and bond, angle, and planarity constraints. In total, when applied repetitively over all three steps, the procedure efficiently generates series of conformations that lie preferentially in the low energy subspace of normal modes.

The RCNMA approach was initially tested on a data set of ten proteins that show conformational changes upon ligand binding [23]. In terms of efficiency, coarse-graining the protein results in a remarkable reduction of memory requirements and computational times by factors of 9 and 27 on average and up to 25 and 125, respectively. In terms of accuracy, directions and magnitudes of motions predicted by this approach agree well with experimentally determined values, despite embracing in extreme cases more than 50% of the protein into one rigid cluster. In fact, the results of the present method are in general comparable to if no or a uniform coarse-graining is applied, and become superior if the movement is dominated by loop or fragment motions. This indicates that explicitly distinguishing between flexible and

rigid regions is advantageous when using a simplified protein representation in the second step. Finally, it should be noted that motions of atoms in rigid clusters are also well predicted by this approach.

The NMSim approach was applied to a dataset of eight proteins with experimentally observed conformational changes (A. Ahmed, H. Gohlke, unpublished results). For proteins that show domain motions, conformational variabilities are reproduced very well for four out of five proteins, with correlation coefficients $r > 0.70$, and as high as $r = 0.92$ in the case of adenylate kinase. In seven out of eight cases, NMSim simulations starting from unbound structures are able to sample conformations that are similar (RMSD $= 1.0 - 3.1$ Å) to ligand-bound conformations. Thus, the generated conformations can serve as an input to ensemble-based docking approaches, as has been demonstrated successfully for peptide–protein docking [68] and docking multiple small-molecule ligands to HIV-1 TAR RNA [24], using a simulation method related to NMSim [25]. Remarkably, an NMSim-generated pathway of conformational change of adenylate kinase correctly describes the sequence of domain closing, very similar to what was found in an all-atom MD simulation [69]. The NMSim approach thus is a computationally efficient alternative to MD simulations for conformational sampling of proteins. Pathways of conformational transitions generated by this method can serve as starting points for more sophisticated sampling techniques, such as umbrella sampling.

18.3.2 Themostabilization of Proteins

Organisms can be classified according to their optimal growth temperatures (T_{og}) into psychrophilic, mesophilic, thermophilic, and hyperthermophilic, with $T_{og} \approx 5$–$25\,^{\circ}$C, 25–50 °C, 50–85 °C, and >85 °C, respectively [70, 71]. Usually, proteins from thermophilic and hyperthermophilic organisms (hereafter referred to as "thermophilic proteins") are thermostable, in that they retain their native fold even at high temperatures. Enzymes with high thermostability are valuable for industrial [72, 73] and biotechnological applications [74]. Therefore, increasing the thermostability of proteins is an important task in protein engineering.

By comparing homologs from mesophilic and thermophilic organisms, different mechanisms have been revealed that lead to increased thermostability. Among these, a better packing of hydrophobic regions and an increased density of salt-bridges or charge-assisted hydrogen bonds are most frequently described [75–77]. In many cases, a complex interplay of these mechanisms was found to be responsible for an increased thermostability [78–80]. As a unifying concept, it was suggested that these changes lead to an improved network of noncovalent interactions within the structure and, thus, to an overall increase in mechanical stability/rigidity of the structure [81].

In order to investigate and improve the thermal stability of proteins, CNA was applied to identify structural features from which a destabilization of a protein structure originates upon thermal unfolding [27]. These unfolding nuclei have been investigated before by experiment and computational studies [82–84]. Unfolding

nuclei are detected by considering the microscopic properties of a constraint network during a thermal unfolding simulation: unfolding nuclei are formed by residues that are part of the giant cluster before the phase transition but are in a flexible region afterwards. In a validation study on pairs of homologous proteins from mesophilic and thermophilic organisms [27], unfolding nuclei identified in thermolysin-like protease (TLP) from *Bacillus cereus* and thermolysin from *Bacillus thermoproteolyticus* are in good agreement with sites where thermostabilizing mutations have been successfully introduced by experiment. Likewise, a good agreement between computed and experimentally verified unfolding nuclei was found for the homologs 3-isopropylmalate dehydrogenase (IPMDH) from *Escherichia coli* and *Thermus thermophilus*. These results demonstrated that unfolding nuclei identified by CNA can be used to guide data-driven protein engineering: unfolding nuclei are prominent candidates for introducing mutations in order to increase thermostability.

In a subsequent study on 19 pairs of homologous proteins from mesophilic and thermophilic organisms [52], the local distribution of flexible and rigid regions in these proteins was analyzed with the help of stability maps, and the findings were related to activity characteristics of the enzyme structures. Again, TLP/thermolysin and IPMDH were considered in more detail. The study results revealed that adaptive mutations in enzymes from thermophilic organisms maintain the balance between overall rigidity (which is important for thermostability), and local flexibility (important for activity) at the appropriate temperature at which the protein functions. Thus, thermophilic adaptation in general leads to an increase of structural rigidity, but conserves the distribution of functionally important flexible regions between homologs from mesophilic and thermophilic organisms. This finding provides direct evidence for the hypothesis of corresponding states [81, 85]. Notably, changes in the flexibility of active-site regions, induced either by a temperature change or by mutations, were related to experimentally observed losses of the enzyme function. From an application point of view, this suggests that exploiting the principle of corresponding states by means of CNA not only allows for successful thermostability optimization but also for guiding experiments in order to improve enzyme activity in protein engineering.

18.3.3
Flexibility of Antibiotics Binding Sites and Allosteric Signal Transmission in Ribosomal Structures

18.3.3.1 Deriving a New Constraint Network Parameterization for RNA Structures

RNA structures are highly flexible biomolecules that show a remarkable ability to undergo large, but controlled, conformational changes to achieve their diverse functional roles [86, 87]. In contrast to globular proteins, RNAs are mostly elongated and more loosely packed [88]. Moreover, both systems have different structural features: the core of proteins is predominantly determined by interactions of hydrophobic side chains [89], while the stability of RNA (and DNA) structures is predominantly governed by hydrogen bonds, base-stacking interactions, and solvation effects [88–90]. Thus, a constraint network representation that has been

developed for proteins may not be appropriate for RNA systems. Indeed, it could be shown that a protein-based parameterization does not capture the flexibility characteristics of RNA structures satisfyingly, but rather leads to too-rigid RNA structures [91]. This led to the development of a new topological network representation of RNA structures, which allows for the reliable determination of flexible and rigid regions within these biomacromolecules [33, 91].

Starting out by analyzing the network rigidity of a canonical A-form RNA, it became obvious that it is the inclusion of hydrophobic contacts into the RNA topological network that is crucial for an accurate flexibility prediction, and that the number of contacts between adjacent bases needs to be limited. Different criteria were then thoroughly tested to include hydrophobic interactions and hydrogen bonds in a constraint network representation of RNA structures [91]. These criteria were adjusted based on comparing results from rigidity analysis with crystallographic B-values of a $tRNA^{ASP}$ structure and NMR order parameters of RNA hairpins. In addition, conformational variabilities of NMR-derived ensembles of 12 RNA structures were compared with atomic fluctuations determined from structural ensembles. The latter ensembles were generated by constraint geometric simulations (similar to the NMSim approach described in Section 18.3.1). Notably, one parameterization was found to be optimal for both predicting infinitesimal motions, as obtained by rigidity analysis, and finite amplitude motions, as obtained by constraint geometric simulations. With this parameterization, it was possible to identify those nucleotides (U8 and U48, G26 and G45) in a $tRNA^{ASP}$ structure as flexible that have been known to function as hinge regions by experiment [91].

18.3.3.2 **Analyzing the Ribosomal Exit Tunnel**

The derived parameterization was then applied to analyze the ribosomal exit tunnel within the large ribosomal subunit [35]. The ribosome is the protein synthesis machinery of the cell. After peptide bond formation at the peptidyl transferase center (PTC) [92], the nascent polypeptide chain leaves the ribosome via the ribosomal exit tunnel, which spans the entire large subunit of the ribosome and has an active role in cotranslational processes [93–97]. Two striking results stand out from this study:

1) By determining a hierarchy of regions of varying stabilities of the large subunit, it was possible to propose a pathway of allosteric signal transmission from the ribosomal tunnel region to the PTC (Figure 18.4a). This finding was later supported by cryo-electron microscopy data of a stalled ribosome structure [98] and mutation studies [99]. The results indicate that the signal transmission is based on mechanical coupling between specific structurally stable regions, which is reminiscent of a tensegrity architecture, which consists of a tensed network of structural members that resist shape distortion (Figure 18.4b). This type of architecture particularly suits mechanical signal transmission due to a local force, as generated by the interactions of nascent polypeptides with the tunnel wall.
2) By investigating ribosomal structures from different organisms, characteristic flexibility patterns were identified in the highly conserved antibiotics binding pocket at the PTC for the different kingdoms that could be linked to antibiotics

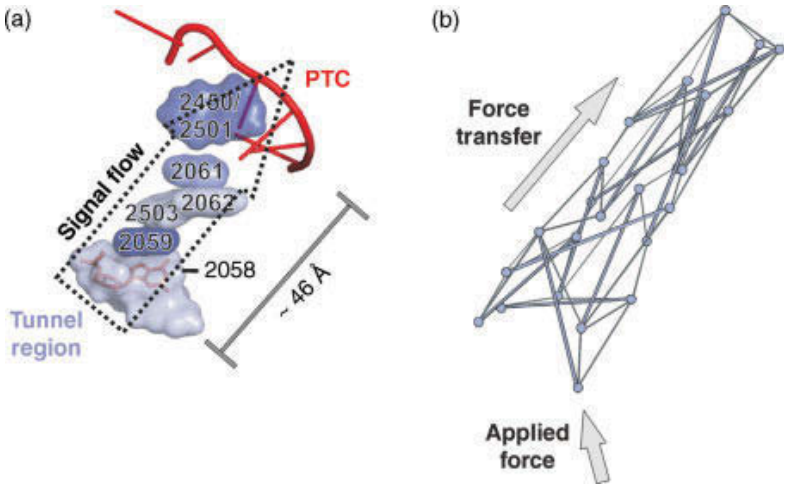


Figure 18.4 (a) Sequence of coupled rigid clusters that allows signal transmission from the ribosomal exit tunnel to the peptidyl transferase center (PTC) by mechanical coupling between specific structurally stable regions (depicted in surface representation with bluish hues; the numbering refers to *E. coli* nucleotides) [35]. The signal transmission occurs over a distance of ~46 Å; (b) A tensegrity structure ("obelisk") depicting how local forces generated from interactions of nascent polypeptides with the tunnel wall can produce structural rearrangements at the PTC. The steel bars (blue) represent the structurally stable regions in the ribosomal structure; the tension cables (black lines) correspond to flexible regions that support/carry the force transfer.

selectivity (Table 18.1). Whereas, the glycosidic bonds of the crevice-forming nucleotides show a dual flexibility character in the case of the archaeal structure (which possesses typical eukaryotic elements at the principal antibiotic target sites), the two glycosidic bonds are structurally stable across all three analyzed bacterial structures. These differences in flexibility characteristics are related to differences in the crevice sizes. As such, a wider active site crevice is found for bacterial structures than for the archaeal structure [100]. As an already open conformation would not require any deformation energy to accommodate to the

Table 18.1 Flexibility characteristics of the antibiotics binding crevice at the PTC.

Group	Organism	Active-site crevice	
		Crevice size[a)]	Flexibility[b)]
Archaea	*Haloarcula marismortui*	Too small	Dual
Bacteria	*Deinococcus radiodurans*	Fits	Stable
	Escherichia coli	Fits	Stable
	Thermus thermophilus	Fits	Stable

a) As reported in Ref. [100].
b) Flexibility characteristics of glycosidic bonds of nucleotides A2451 and C2452 (*E. coli* numbering) obtained by rigidity analysis [35].

bound conformation, this could be the reason why bacteria are more sensitive to some of the active-site crevice antibiotics than archaea [100]. The constraint counting results further support this hypothesis. Bacteria are vulnerable to antibiotics because of an open conformation of the active site crevice that is structurally stable. In contrast, archaea (eukaryotes) can only bind antibiotics if the narrow conformation of the crevice can widen, as given by the dual flexibility characteristics. Overall, the study results show that, in order to explain the binding selectivity of antibiotics, it is necessary to take flexibility characteristics of the binding sites into account.

18.4 Conclusions

During recent years, encouraging progress has been made in applying graph-theoretical approaches for characterizing the flexibility and rigidity of biomacromolecules down to the bond level, and linking this information to biological function. The underlying theory, computational approaches, and sample applications have been reviewed in this chapter. Rigidity analysis usually takes a few seconds on proteins of hundreds or thousands of residues, and so can also be efficiently applied to large biomacromolecules, such as the ribosome. Promising applications of rigidity analysis include supporting data-driven protein engineering by identifying structural parts that impact protein thermostability, probing signal transmission in order to identify new putative allosteric binding sites, and assisting in the assessment of flexibility characteristics of binding sites. These are areas of active research by the present authors and others.

References

1 Schellman, J.A. (1975) Macromolecular binding. *Biopolymers*, **14** (5), 999–1018.

2 Daniel, R.M., Dunn, R.V., Finney, J.L., and Smith, J.C. (2003) The role of dynamics in enzyme activity. *Annu. Rev. Biophys. Biomol. Struct.*, **32**, 69–92.

3 Cozzini, P., Kellogg, G.E., Spyrakis, F., Abraham, D.J., Costantino, G., Emerson, A., Fanelli, F., Gohlke, H., Kuhn, L.A., Morris, G.M., Orozco, M., Pertinhez, T.A., Rizzi, M., and Sotriffer, C.A. (2008) Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.*, **51** (20), 6237–6255.

4 Fischer, E. (1894) Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch Chem. Ges.*, **27** (3), 2985–2993.

5 Koshland, D.E. Jr (1958) Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl Acad. Sci. USA*, **44** (2), 98–104.

6 Ma, B., Kumar, S., Tsai, C.J., and Nussinov, R. (1999) Folding funnels and binding mechanisms. *Protein Eng.*, **12** (9), 713–720.

7 Tsai, C.J., Kumar, S., Ma, B., and Nussinov, R. (1999) Folding funnels, binding funnels, and protein function. *Protein Sci.*, **8** (6), 1181–1190.

8 Vihinen, M. (1987) Relationship of protein flexibility to thermostability. *Protein Eng.*, **1** (6), 477–480.

9 Závodszky, P., Kardos, J., Svingor, Á., and Petsko, G.A. (1998) Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins.

Proc. Natl Acad. Sci. USA, **95** (13), 7406–7411.

10 Ishima, R. and Torchia, D.A. (2000) Protein dynamics from NMR. *Nat. Struct. Mol. Biol.*, **7** (9), 740–743.

11 Weiss, S. (1999) Fluorescence spectroscopy of single biomolecules. *Science*, **283** (5408), 1676–1683.

12 Zhang, X.J., Wozniak, J.A., and Matthews, B.W. (1995) Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozymes. *J. Mol. Biol.*, **250** (4), 527–552.

13 Frank, J. and Agrawal, R.K. (2000) A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature*, **406** (6793), 318–322.

14 Smith, D.K., Radivojac, P., Obradovic, Z., Dunker, A.K., and Zhu, G. (2003) Improved amino acid flexibility parameters. *Protein Sci.*, **12** (5), 1060–1072.

15 Palmer, A.G. 3rd, Kroenke, C.D., and Loria, J.P. (2001) Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Methods Enzymol.*, **339**, 204–238.

16 Dodson, G.G., Lane, D.P., and Verma, C.S. (2008) Molecular simulations of protein dynamics: new windows on mechanisms in biology. *EMBO Reports*, **9** (2), 144–150.

17 Karplus, M. and McCammon, J.A. (2002) Molecular dynamics simulations of biomolecules. *Nat. Struct. Mol. Biol.*, **9** (9), 646–652.

18 Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., and Bahar, I. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J.*, **80** (1), 505–515.

19 Bahar, I., Atilgan, A.R., and Erman, B. (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, **2** (3), 173–181.

20 Case, D.A. (1994) Normal mode analysis of protein dynamics. *Curr. Opin. Struct. Biol.*, **4** (2), 285–290.

21 Jacobs, D.J., Rader, A.J., Kuhn, L.A., and Thorpe, M.F. (2001) Protein flexibility predictions using graph theory. *Proteins: Struct. Funct. Bioinf.*, **44** (2), 150–165.

22 Gohlke, H. and Thorpe, M.F. (2006) A natural coarse graining for simulating large biomolecular motion. *Biophys. J.*, **91** (6), 2115–2120.

23 Ahmed, A. and Gohlke, H. (2006) Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. *Proteins: Struct. Funct. Bioinf.*, **63** (4), 1038–1051.

24 Fulle, S., Christ, N.A., Kestner, E., and Gohlke, H. (2010) HIV-1 TAR RNA spontaneously undergoes relevant apo-to-holo conformational transitions in molecular dynamics and constrained geometrical simulations. *J. Chem. Inf. Model.*, **50** (8), 1489–1501.

25 Wells, S., Menor, S., Hespenheide, B.M., and Thorpe, M.F. (2005) Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.*, **2** (4), S127–S136.

26 Farrell, D.W., Speranskiy, K., and Thorpe, M.F. (2010) Generating stereochemically acceptable protein pathways. *Proteins: Struct. Funct. Bioinf.*, **78** (14), 2908–2921.

27 Radestock, S. and Gohlke, H. (2008) Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.*, **8** (5), 507–522.

28 Livesay, D.R. and Jacobs, D.J. (2006) Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins: Struct. Funct. Bioinf.*, **62** (1), 130–143.

29 Hespenheide, B.M., Rader, A.J., Thorpe, M.F., and Kuhn, L.A. (2002) Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graphics Model.*, **21** (3), 195–207.

30 Rader, A.J. and Bahar, I. (2004) Folding core predictions from network models of proteins. *Polymer*, **45** (2), 659–668.

31 Del Carpio, C.A., Iulian Florea, M., Suzuki, A., Tsuboi, H., Hatakeyama, N., Endou, A., Takaba, H., Ichiishi, E., and Miyamoto, A. (2009) A graph theoretical approach for assessing bio-macromolecular complex structural

stability. *J. Mol. Model.*, **15** (11), 1349–1370.

32 Gohlke, H., Kuhn, L.A., and Case, D.A. (2004) Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins: Struct. Funct. Bioinf.*, **56** (2), 322–337.

33 Fulle, S. and Gohlke, H. (2009) Constraint counting on RNA structures: linking flexibility and function. *Methods*, **49** (2), 181–188.

34 Tan, H.P. and Rader, A.J. (2009) Identification of putative, stable binding regions through flexibility analysis of HIV-1 gp120. *Proteins: Struct. Funct. Bioinf.*, **74** (4), 881–894.

35 Fulle, S. and Gohlke, H. (2009) Statics of the ribosomal exit tunnel: implications for cotranslational peptide folding, elongation regulation, and antibiotics binding. *J. Mol. Biol.*, **387** (2), 502–517.

36 Mottonen, J.M., Jacobs, D.J., and Livesay, D.R. (2010) Allosteric response is both conserved and variable across three CheY orthologs. *Biophys. J.*, **99** (7), 2245–2254.

37 Jacobs, D.J. and Dallakyan, S. (2005) Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys. J.*, **88** (2), 903–915.

38 Maxwell, J.C. (1864) On the calculation of the equilibrium and stiffness of frames. *Philos. Mag. Series 4*, **27** (182), 294–299.

39 Laman, G. (1970) On graphs and rigidity of plane skeletal structures. *J. Eng. Math.*, **4** (4), 331–340.

40 Tay, T.S. and Whiteley, W. (1984) Recent advances in the generic rigidity of structures. *Struct. Topol.*, **9**, 31–38.

41 Katoh, N. and Tanigawa, S. (2009) A proof of the molecular conjecture, in Proceedings of the 25th Annual Symposium on Computational Geometry, 8–10 June 2009, Aarhus, Denmark. ACM, pp. 296–305.

42 Jacobs, D.J. and Hendrickson, B. (1997) An algorithm for two-dimensional rigidity percolation: the pebble game. *J. Comput. Phys.*, **137** (2), 346–365.

43 Jacobs, D.J. and Thorpe, M.F. (1995) Generic rigidity percolation: the pebble game. *Phys. Rev. Lett.*, **75** (22), 4051–4054.

44 Jacobs, D.J. (1998) Generic rigidity in three-dimensional bond-bending networks. *J. Phys. A: Math. Gen.*, **31**, 6653–6668.

45 ProFlex, a program for analyzing flexibility of networks. Available from: http://www.bch.msu.edu/~kuhn/software/proflex/ (accessed 12 March 2011).

46 FIRST, a program for analyzing flexibility of networks. Available from: http://flexweb.asu.edu/ (accessed 12 March 2011).

47 Hespenheide, B.M., Jacobs, D.J., and Thorpe, M.F. (2004) Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. *J. Phys. Condens. Matter*, **16** (44), S5055–S5064.

48 Stevens, M., Boolchand, P., and Hernandez, J.G. (1985) Universal structural phase transition in network glasses. *Phys. Rev. B Condens. Matter Mater. Phys.*, **31** (2), 981–991.

49 Thorpe, M.F., Jacobs, D.J., Chubynsky, M.V., and Phillips, J.C. (2000) Self-organization in network glasses. *J. Non-Cryst. Solids*, **266–269** (Part 2), 859–866.

50 Wang, Y., Wells, J., Georgiev, D.G., Boolchand, P., Jackson, K., and Micoulaut, M. (2001) Sharp rigid to floppy phase transition induced by dangling ends in a network glass. *Phys. Rev. Lett.*, **87** (18), 185503.

51 Rader, A.J. (2010) Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys. Biol.*, **7** (1), 16002.

52 Radestock, S. and Gohlke, H. (2011) Protein rigidity and thermophilic adaptation. *Proteins: Struct. Funct. Bioinf.*, **79** (4), 1089–1108.

53 Rader, A.J., Hespenheide, B.M., Kuhn, L.A., and Thorpe, M.F. (2002) Protein unfolding: rigidity lost. *Proc. Natl Acad. Sci. USA*, **99** (6), 3540–3545.

54 Dahiyat, B.I., Gordon, D.B., and Mayo, S.L. (1997) Automated design of

the surface positions of protein helices. *Protein Sci.*, **6** (6), 1333–1337.

55 Makhatadze, G.I. and Privalov, P.L. (1995) Energetics of protein structure. *Adv. Protein Chem.*, **47**, 307–425.

56 Mamonova, T., Hespenheide, B., Straub, R., Thorpe, M.F., and Kurnikova, M. (2005) Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.*, **2** (4), S137–S147.

57 Taverna, D.M. and Goldstein, R.A. (2002) Why are proteins marginally stable? *Proteins: Struct. Funct. Bioinf.*, **46** (1), 105–109.

58 Livesay, D.R., Huynh, D.H., Dallakyan, S., and Jacobs, D.J. (2008) Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family. *Chem. Cent. J.*, **2** (17), 1–20.

59 Albert, R. and Barabási, A.L. (2002) Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74** (1), 47–97.

60 Stauffer, D. and Aharony, A. (1992) *Introduction to Percolation Theory*, Taylor & Francis, London.

61 Andraud, C., Beghdadi, A., and Lafait, J. (1994) Entropic analysis of random morphologies. *Physica A*, **207** (1–3), 208–212.

62 Carlson, H.A. (2002) Protein flexibility and drug design: How to hit a moving target. *Curr. Opin. Chem. Biol.*, **6** (4), 447–452.

63 Ahmed, A., Kazemi, S., and Gohlke, H. (2007) Protein flexibility and mobility in structure-based drug design. *Front. Drug Des. Discovery*, **3** (1), 455–476.

64 Tozzini, V. (2005) Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.*, **15** (2), 144–150.

65 Lei, M., Zavodszky, M.I., Kuhn, L.A., and Thorpe, M.F. (2004) Sampling protein conformations and pathways. *J. Comput. Chem.*, **25** (9), 1133–1148.

66 Ahmed, A. and Gohlke, H. (2009) Multiscale modeling of macromolecular conformational changes, in 1st International Conference on Computational & Mathematical Biomedical Engineering – CMBE09, 29 June–1 July 2009, Swansea, UK, pp. 219–222.

67 Durand, P., Trinquier, G., and Sanejouand, Y.-H. (1994) A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers*, **34** (6), 759–771.

68 Zavodszky, M.I., Ming, L., Thorpe, M.F., Day, A.R., and Kuhn, L.A. (2004) Modeling correlated main-chain motions in proteins for flexible molecular recognition. *Proteins: Struct. Funct. Bioinf.*, **57** (2), 243–261.

69 Maragakis, P. and Karplus, M. (2005) Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J. Mol. Biol.*, **352** (4), 807–822.

70 Demirjian, D.C., Moris-Varas, F., and Cassidy, C.S. (2001) Enzymes from extremophiles. *Curr. Opin. Chem. Biol.*, **5** (2), 144–151.

71 Vieille, C. and Zeikus, G.J. (2001) Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.*, **65** (1), 1–43.

72 Polizzi, K.M., Bommarius, A.S., Broering, J.M., and Chaparro-Riggers, J.F. (2007) Stability of biocatalysts. *Curr. Opin. Chem. Biol.*, **11** (2), 220–225.

73 Ferrer, M., Golyshina, O., Beloqui, A., and Golyshin, P.N. (2007) Mining enzymes from extreme environments. *Curr. Opin. Microbiol.*, **10** (3), 207–214.

74 Podar, M. and Reysenbach, A.L. (2006) New opportunities revealed by biotechnological explorations of extremophiles. *Curr. Opin. Biotechnol.*, **17** (3), 250–255.

75 Robinson-Rechavi, M., Alibes, A., and Godzik, A. (2006) Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of *Thermotoga maritima*. *J. Mol. Biol.*, **356** (2), 547–557.

76 Szilagyi, A. and Zavodszky, P. (2000) Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein

subunits: results of a comprehensive survey. *Structure*, **8** (5), 493–504.
77 Vogt, G., Woell, S., and Argos, P. (1997) Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.*, **269** (4), 631–643.
78 Russell, R.J. and Taylor, G.L. (1995) Engineering thermostability: lessons from thermophilic proteins. *Curr. Opin. Biotechnol.*, **6** (4), 370–374.
79 Querol, E., Perez-Pons, J.A., and Mozo-Villarias, A. (1996) Analysis of protein conformational characteristics related to thermostability. *Protein Eng.*, **9** (3), 265–271.
80 Vieille, C. and Zeikus, J.G. (1996) Thermozymes: Identifying molecular determinants of protein structural and functional stability. *Trends Biotechnol.*, **14** (6), 183–190.
81 Jaenicke, R. and Böhm, G. (1998) The stability of proteins in extreme environments. *Curr. Opin. Struct. Biol.*, **8** (6), 738–748.
82 Eijsink, V.G.H., Veltman, O.R., Aukema, W., Vriend, G., and Venema, G. (1995) Structural determinants of the stability of thermolysin-like proteinases. *Nat. Struct. Biol.*, **2** (5), 374–379.
83 Creveld, L.D., Amadei, A., van Schaik, R.C., Pepermans, H.A., de Vlieg, J., and Berendsen, H.J. (1998) Identification of functional and unfolding motions of cutinase as obtained from molecular dynamics computer simulations. *Proteins: Struct. Funct. Bioinf.*, **33** (2), 253–264.
84 Gianese, G., Bossa, F., and Pascarella, S. (2002) Comparative structural analysis of psychrophilic and meso- and thermophilic enzymes. *Proteins: Struct. Funct. Bioinf.*, **47** (2), 236–249.
85 Somero, G.N. (1978) Temperature adaptation of enzymes: biological optimization through structure-function compromises. *Annu. Rev. Ecol. Syst.*, **9**, 1–29.
86 Al-Hashimi, H.M. and Walter, N. (2008) RNA dynamics: it is about time. *Curr. Opin. Struct. Biol.*, **18** (3), 321–329.
87 Fulle, S. and Gohlke, H. (2010) Molecular recognition of RNA: challenges for modelling interactions and plasticity. *J. Mol. Recognit.*, **23** (2), 220–231.
88 Van Wynsberghe, A.W. and Cui, Q. (2005) Comparison of mode analyses at different resolutions applied to nucleic acid systems. *Biophys. J.*, **89** (5), 2939–2949.
89 Hyeon, C., Dima, R.I., and Thirumalai, D. (2006) Size, shape, and flexibility of RNA structures. *J. Chem. Phys.*, **125** (19), 194905.
90 Gohlke, H., Bozilovic, J., and Engels, J.W. (2011) Synthesis and properties of fluorinated nucleobases in DNA and RNA, in *Fluorine in Pharmaceutical and Medicinal Chemistry: From Biophysical Aspects to Clinical Applications*, 1st edn (eds V. Gouverneur and K. Mueller) World Scientific Publishing Co., New Jersey, USA.
91 Fulle, S. and Gohlke, H. (2008) Analyzing the flexibility of RNA structures by constraint counting. *Biophys. J.*, **94** (11), 4202–4219.
92 Nissen, P., Hansen, J., Ban, N., Moore, P.B., and Steitz, T.A. (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science*, **289** (5481), 920–930.
93 Berisio, R., Schluenzen, F., Harms, J., Bashan, A., Auerbach, T., Baram, D., and Yonath, A. (2003) Structural insight into the role of the ribosomal tunnel in cellular regulation. *Nat. Struct. Biol.*, **10** (5), 366–370.
94 Etchells, S.A. and Hartl, F.U. (2004) The dynamic tunnel. *Nat. Struct. Mol. Biol.*, **11** (5), 391–392.
95 Gilbert, R.J., Fucini, P., Connell, S., Fuller, S.D., Nierhaus, K.H., Robinson, C.V., Dobson, C.M., and Stuart, D.I. (2004) Three-dimensional structures of translating ribosomes by cryo-EM. *Mol. Cell*, **14** (1), 57–66.
96 Nakatogawa, H. and Ito, K. (2002) The ribosomal exit tunnel functions as a discriminating gate. *Cell*, **108** (5), 629–636.
97 Woolhead, C.A., McCormick, P.J., and Johnson, A.E. (2004) Nascent membrane and secretory proteins differ in FRET-detected folding far inside the ribosome and in their

exposure to ribosomal proteins. *Cell*, **116** (5), 725–736.

98 Seidelt, B., Innis, C.A., Wilson, D.N., Gartmann, M., Armache, J.P., Villa, E., Trabuco, L.G., Becker, T., Mielke, T., Schulten, K., Steitz, T.A., and Beckmann, R. (2009) Structural insight into nascent polypeptide chain-mediated translational stalling. *Science*, **326** (5958), 1412–1415.

99 Vázquez-Laslop, N., Ramu, H., Klepacki, D., Kannan, K., and Mankin, A. (2010) The key function of a conserved and modified rRNA residue in the ribosomal response to the nascent peptide. *EMBO J.*, **29** (18), 3108–3117.

100 Blaha, G., Gürel, G., Schroeder, S.J., Moore, P.B., and Steitz, T.A. (2008) Mutations outside the anisomycin-binding site can make ribosomes drug-resistant. *J. Mol. Biol.*, **379** (3), 505–519.

101 Whiteley, W. (2005) Counting out to the flexibility of molecules. *Phys. Biol.*, **2** (4), S116–S126.