Protein rigidity and thermophilic adaptation

Sebastian Radestock and Holger Gohlke*

Mathematisch-Naturwissenschaftliche Fakultät, Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, Germany

ABSTRACT

We probe the hypothesis of corresponding states, according to which homologues from mesophilic and thermophilic organisms are in corresponding states of similar rigidity and flexibility at their respective optimal temperatures. For this, the local distribution of flexible and rigid regions in 19 pairs of homologous proteins from meso- and thermophilic organisms is analyzed and related to activity characteristics of the enzymes by constraint network analysis (CNA). Two pairs of enzymes are considered in more detail: 3-isopropyldehydrogenase and thermolysin-like protease. By comparing microscopic stability features of homologues with the help of stability maps, introduced for the first time, we show that adaptive mutations in enzymes from thermophilic organisms maintain the balance between overall rigidity, important for thermostability, and local flexibility, important for activity, at the appropriate working temperature. Thermophilic adaptation in general leads to an increase of structural rigidity but conserves the distribution of functionally important flexible regions between homologues. This finding provides direct evidence for the hypothesis of corresponding states. CNA thereby implicitly captures and unifies many different mechanisms that contribute to increased thermostability and to activity at high temperatures. This allows to qualitatively relate changes in the flexibility of active site regions, induced either by a temperature change or by the introduction of mutations, to experimentally observed losses of the enzyme function. As for applications, the results demonstrate that exploiting the principle of corresponding states not only allows for successful thermostability optimization but also for guiding experiments in order to improve enzyme activity in protein engineering.

INTRODUCTION

In the recent years, there has been considerable effort to understand the structural determinants of stability and activity of proteins from thermophilic and hyperthermophilic organisms (henceforth referred to together as thermophilic proteins) with optimal growth temperatures of more than 60°C.1,2 These proteins are required to retain their native structures even under extreme environmental conditions, when homologues from mesophilic organisms already denature.1,2 Consequently, thermophilic proteins often have a substantially higher intrinsic thermostability than their mesophilic counterparts, while retaining the basic fold characteristics, and in the case of enzymes, the catalytic site properties of the particular protein family.1,3

Because of their adaptation to high temperature, thermophilic enzymes are interesting for the biotechnological industry as biocatalysts.4–6 However, the identification of suitable thermophilic enzymes with a specific temperature-dependent activity profile by screening is tedious.6,7 As a valuable alternative, mesophilic enzymes can be re-engineered to become more thermostable.8,9 Thus, there is great interest in the biotechnology industry for having tools in hand that aid in developing thermostable enzymes by protein engineering. In addition to stability, optimizing enzyme activity for elevated target temperatures is mandatory. This requires both a thorough understanding of the structural determinants of thermostability and temperature-dependent activity.10

Numerous studies have been performed to identify principles of thermostability and, subsequently, to apply those in rational or data-driven protein engineering.8,11 These include a large and growing number of comparative studies using mesophilic and thermophilic protein homologues.12 The influence of various sequence and structure determinants of thermophilic adaptation has been studied.13–21 Often, thermophilic adaptation comes along with a better packing of hydrophobic interactions and an increased density of salt-bridges or charge-assisted hydrogen bonds.19,20,22 In many cases, a combination of different sequence or structure features was found to be responsible for an increased thermostability.10,21,23

Additional Supporting Information may be found in the online version of this article.

Sebastian Radestock’s current address is Computational Structural Biology Group, Max Planck Institute of Biophysics, Max-von-Laue-Str. 3, 60438 Frankfurt am Main, Germany.

*Correspondence to: Holger Gohlke, Universitätstr. 1, 40225 Düsseldorf, Germany.

E-mail: gohlke@uni-duesseldorf.de.

Received 5 August 2010; Revised 28 September 2010; Accepted 7 November 2010

Published online 19 November 2010 in Wiley Online Library (wileyonlinelibrary.com).

DOI: 10.1002/prot.22946
These structural changes contribute to the improvement of the underlying network of noncovalent interactions within the structure, presumably leading to an increase in mechanical rigidity of the structure and a decrease in flexibility.1

Motional properties of proteins with different thermal stabilities have been compared using a variety of different experimental techniques, including hydrogen/deuterium (H/D) exchange, neutron scattering, and nuclear magnetic resonance (NMR) spectroscopy.24 As an intriguing conclusion, the local distribution of flexible or rigid regions in a thermophilic protein is similar to that of a mesophilic homologue, provided that the proteins are compared at their respective optimal temperatures. This has led to the hypothesis that homologous mesophilic and thermophilic proteins are in corresponding states of similar rigidity and flexibility at their respective optimal temperature.1,25 This should be especially true for enzymes, where stability characteristics of catalytic and binding site residues are an important factor in determining the activity.3 In fact, many thermophilic enzymes have a maximal activity at elevated temperature, but are far less active at ambient temperatures, where the structures are more rigid.26–29 Thus, rigidity is an important factor for the integrity of the native folded structure, whereas a certain degree of flexibility is required for activity, although the general validity of this relation is being discussed.30

The experimental characterization and quantification of protein flexibility and rigidity are challenging, especially in regard to the diverse types and timescales of motions that are of interest.24,31 Moreover, changes in flexibility due to sequence variations can be limited to a small part of the structure only, which makes them difficult to detect.32 Computational approaches are interesting alternatives. As an example, molecular dynamics (MD) simulations provide information on the motional properties of atoms in a protein structure. By performing simulations at different temperatures or by running nonequilibrium simulations, motional properties of the structure during thermal unfolding can be analyzed.33–36 By comparing equilibrium and nonequilibrium MD simulations, Creveld et al.37 studied the unfolding of cutinase and distinguished two types of flexible regions: those that promote unfolding and those that are important for enzyme activity. Thereby, unfolding regions were predicted where mutations modulate thermostability and do not interfere with enzyme activity. Despite these successes, such simulations are computationally expensive, and it is not clear whether force fields and solvent models are adequate for use at high temperatures. Without doing expensive MD simulations, moving protein parts can also be predicted from a single protein conformation using methods related to normal mode analysis (NMA).38 Using such an approach, Su et al.39 studied changes in the fluctuation of residues upon protein unfolding.

Both MD and NMA studies report on the mobility of atoms by analyzing atomic fluctuations or the correlation of movements, rather than flexibility and rigidity characteristics of structural regions. Here, flexibility denotes the possibility of motion within the region, that is, the region can be deformed. In contrast, no relative motion is allowed within rigid (structurally stable) regions, and such regions can only move as a rigid body with six degrees of freedom. Thus, flexibility and its opposite, rigidity, are static properties that describe the possibility of motion.40,41 The distinction between mobility and flexibility/rigidity becomes obvious in the case of a mobile rigid body, such as a moving helix or domain.

In a previous study, we showed for a dataset of 20 pairs of mesophilic and thermophilic protein homologues that protein thermostability can be predicted by directly characterizing the mechanical rigidity of a protein structure during a thermal-unfolding simulation.42 The approach is referred to as constraint network analysis (CNA) and based on a graph-theoretical method that determines rigidity and flexibility within a protein structure in atomic resolution.43 We showed that the approach is helpful for understanding and exploiting the relationship between microscopic structure and macroscopic stability and, thus, 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.

Here, we go beyond the previous study, which related protein thermal stability and rigidity by analyzing the local distribution of flexible or rigid regions and relating it to activity characteristics of enzyme structures. We show that CNA implicitly captures and unifies many different mechanisms that contribute to increased thermostability and to activity at high temperatures. Two pairs of enzymes from our data set are considered in more detail: 3-isopropylmalate dehydrogenase (IPMDH) and thermolysin-like protease (TLP). By comparing microscopic stability features of the meso- and thermophilic protein homologues, we show that adaptive mutations of thermophilic enzymes maintain the balance between overall rigidity, important for thermostability, and local flexibility, important for activity, at the respective temperature at which the protein functions. Thus, direct evidence is offered that supports the hypothesis of corresponding states. To the best of our knowledge, the present study is the first one probing this hypothesis for a large data set by computational means. Moreover, unlike other computational studies that probed this hypothesis,35,36 protein rigidity and flexibility are characterized directly. Our results demonstrate that exploiting the principle of corresponding states not only allows for successfully optimizing thermostability but also for guiding experiments in order to improve enzyme activity in protein engineering.
THEORY

In the following, we will briefly introduce the theoretical background of CNA. With this method, the mechanical rigidity of a protein structure is characterized by simulating the thermal unfolding of a protein.

Unfolding simulations on constraint network representations of proteins

Protein stability is determined by the large number of interactions that contribute to the native-folded state. Thus, to theoretically study protein stability features, proteins can be described as molecular networks. In the present study, we go beyond describing proteins as topological networks. Instead, protein structures are represented as constraint networks, in particular molecular frameworks. These networks are a subtype of 3D-networks in which vertices represent atoms and edges represent covalent and noncovalent bond and angle constraints. As noncovalent bond constraints, hydrogen bonds, including salt bridges, and hydrophobic interactions are used. The rigidity and flexibility within a constraint network can be fully characterized using constraint counting. As a result, the protein-structure network is decomposed into rigid clusters and flexible regions. A rigid cluster is a set of atoms that move together as a rigid body, that is, no internal movement is allowed. Otherwise, an atom is part of a flexible region. The rigid cluster decomposition can be calculated using the FIRST (Floppy Inclusion and Rigid Substructure Topography) software.

Thermal unfolding of the protein structure can now be simulated by gradually removing noncovalent bond constraints from the constraint network. Thus, starting from a rigid network with a large number of constraints, a flexible network is obtained with only a few constraints left. Here, hydrogen bonds (including salt bridges) are removed from the network in the order of increasing strength, that is, only hydrogen bonds with an energy \( E_{hh} \leq E_{cu} \) are retained for a certain network state. This follows the idea that stronger hydrogen bonds will break at higher temperatures than weaker ones. The number of hydrophilic contacts is kept constant during the thermal unfolding, because the strength of hydrophobic interactions remains constant or even increases with increasing temperature. A new network is generated for each hydrogen bond that is removed, and the rigid cluster decomposition is performed on this network.

Identifying the folded-unfolded transition

During the thermal-unfolding simulation, a phase transition from a largely rigid to a largely flexible network is observed. At the transition, the percolating rigid cluster (the giant cluster) suddenly breaks and stops dominating the system. The transition defines the rigidity percolation threshold. Both proteins and glasses are similar in that such a transition can be observed. However, the percolation behavior of protein networks is usually more complex, and several transitions can be observed. This is related to the fact that protein structures are modular, as they are assembled from secondary structure elements, subdomains, and domains, which often break away from the giant cluster as a whole. As every protein fold has its unique number and pattern of spatially arranged modules, a percolation behavior characteristic for each fold is observed.

To describe the general percolation behavior of a network, the microstructure of the network, that is, properties of the set of clusters generated by the bond-dilution process are analyzed. Here, we choose the fraction of the network belonging to the percolating rigid cluster as an order parameter for characterizing the rigidity transition (the rigidity-order parameter, \( P_c \)). The rigidity-order parameter denotes the probability that an atom belongs to the giant cluster and is zero in the floppy phase. Monitoring the decay of the giant cluster by the rigidity-order parameter provides a global and intuitive description of the rigidity within the protein structure during thermal unfolding. Although the percolation behavior of protein networks is complex, the curves of the rigidity-order parameter are similar to curves observed for network models of glasses and amorphous solids.

The first transition observed describes the break down of the completely rigid network into a number of rigid clusters. The last transition describes the loss of the remaining piece of rigidity and the onset of a network that is completely flexible, that is, rigid clusters cease to exist. The first transition is referred to as rigidity percolation transition, because the network is unable to transmit stress afterward. However, the last transition is biologically most relevant, because it corresponds to the folded-unfolded transition in experimental protein unfolding.

To identify the last steps of the folded–unfolded transition that is relevant from a biological point of view, a parameter for analyzing macroscopic properties of the network associated with the rigid cluster-size distribution is used. In this context, Andraud et al. introduced the cluster configuration entropy (\( H \)) as a morphological descriptor for heterogeneous materials. \( H \) 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. Originally, \( H \) was defined as a function of the probability that an atom is part of a cluster of size \( s \) (\( s \)-cluster). This \( s \)-cluster configuration entropy definition has been shown to be particularly useful for identifying early steps of the transition of protein networks between largely rigid states and partially flexible states.

However, we are interested in identifying the late steps of the transition between partially rigid states, where a
rigid core in the structure network is still present, and largely flexible states, where the core is broken. This is in accordance with the idea that the core then represents the folding nucleus that dominates the transition state in experimental protein (un)folding. To identify these steps of the transition, we used a configuration entropy

\[ H = - \sum_i w_i \ln w_i, \]  

where the probability \( w_i \) that an arbitrarily occupied site belongs to a \( s^2 \)-cluster is given by

\[ w_i = \frac{s^2 n_s}{\sum_s s^2 n_s} \]  

and the normalized cluster number \( n_s \) is

\[ n_s = \frac{\text{number of clusters of size } s}{N}. \]

\( N \) equals the total number of atoms.

Our definition of the configuration entropy is useful for identifying the temperature at which the giant cluster finally stops to dominate the system, that is, where a transition takes place that corresponds to the folded–unfolded transition in protein unfolding. The temperature at which this transition occurs is referred to as the phase-transition temperature (\( T_p \)), which can be related to the experimental melting temperature (\( T_m \)). See Methods section for a description as to how the phase-transition temperature is determined.

**RESULTS AND DISCUSSION**

**Data set of mesophilic and thermophilic homologues**

The hypotheses underlying the present study are that thermostability comes along with an increased mechanical rigidity of the structure and that mesophilic and thermophilic proteins are in corresponding states of similar rigidity and flexibility at their respective optimal temperature. To probe these hypotheses, a data set containing homologous protein structures from mesophilic and thermophilic organisms was constructed (Table S1).\(^{42}\) The structures in the data set fulfill stringent quality criteria based on previous results that indicate that a good structural quality is necessary for successful rigidity analysis.\(^{43}\) The data set consists of 19 pairs of protein structures from 19 different protein families (Table S1). The corresponding mesophilic and thermophilic proteins within each family are highly similar, with sequence identities varying from 37 to 88% and backbone root mean square deviation (rmsd) values between 0.6 and 2.5 Å. Higher rmsd values were only obtained for proteins of families 5 (elongation factor TU) and 11 (phosphoglycerate kinase) due to differences in the configuration of domains.

The high sequential and structural similarities point to close relationships with respect to the three-dimensional topology of the molecules. The presence of conserved catalytic residues suggests that the homologues are closely related, if not identical, from a mechanistic point of view, too. Sequence alignments of 3-IPMDH and TLP are shown in the Supporting Information (Figs. S1 and S2), where catalytic and other functionally important residues are highlighted.

Ideally, the melting temperature (\( T_m \)) of a protein should be used as a descriptor of thermostability. Unfortunately, only a few proteins in our data set have melting temperatures currently available, because the \( T_m \) of many proteins cannot be determined experimentally. The limited number of proteins with known \( T_m \), thus, forced us to use the optimal growth temperature of the corresponding organism (\( T_{ \text{opt} } \)) as a descriptor of thermostability. Optimal growth temperatures have previously been used in studies comparing meso- and thermophilic proteins.\(^{16}\)

**Macroscopic stability is characterized by the final decay of a rigid core**

All protein structures were subjected to thermal-unfolding simulations using CNA. CNA is based on a rigidity analysis of protein constraint networks, as implemented in the FIRST software.\(^{43}\) This method is remarkably fast, and unfolding trajectories can be produced within a few minutes. Using CNA, global (macroscopic) stability is characterized by monitoring the size of the largest rigid cluster (the giant cluster) and the rigid cluster-size distribution within the network as the thermal-unfolding simulation proceeds. These properties are described by the rigidity-order parameter (\( P_{ns} \)) and the \( s^2 \)-cluster configuration entropy (\( H \)), respectively. From these quantities, the temperature of a biologically relevant phase transition (\( T_p \)) is obtained that relates to the protein’s melting temperature (\( T_m \)).\(^{42}\) Here, only differences in the phase-transition temperatures of homologous proteins are considered.\(^{42}\)

The rigidity-order parameter plots of two representative pairs of proteins from our data set, IPMDH and TLP, are shown in Figure 1(a,b). As described in the Theory section, the curves characterize the general percolation behavior of the constraint networks during unfolding. In respect to the pattern of transitions, the rigidity-order parameter curves of thermophilic proteins are almost identical to their mesophilic counterparts. Nevertheless, the rigidity-order parameter curves of thermophilic proteins are shifted to higher temperatures (except in low-temperature regions). From a global point of view, these observations provide initial support to 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.

The $s^*$-cluster configuration entropy plots of the two representative pairs of proteins from our data set are shown in Figure 1(c,d). The cluster configuration entropy is used to detect the rigid-to-flexible phase transition that is biologically relevant, that is, where the last rigid region integrating multiple secondary structure elements breaks, and the structure is no longer dominated by a rigid cluster. This melting is associated with small changes in the topology of the constraint network, which have a large effect on the rigid cluster-size distribution. Accordingly, the curve shows a steep increase at the phase-transition temperature, where the relevant transition takes place. As can be seen from Figure 1(c,d), the thermophilic protein has a higher phase-transition temperature compared to its mesophilic homologue.

Results for all pairs of protein structures from the data set are given in Table I. Notably, for approximately two thirds of the 19 different protein families, the $T_p$ of the thermophilic protein is predicted to be higher than that of the mesophilic counterpart, as exemplarily demonstrated for IPMDH and TLP. This shows that the rigid

**Table I**

Differences in experimental $T_{og}$ values versus differences in computed phase-transition temperatures ($T_p$) between a mesophilic protein and its thermophilic homologue.

<table>
<thead>
<tr>
<th>Protein family</th>
<th>$\Delta T_{og}$</th>
<th>$\Delta T_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Pyrrolidone carboxyl peptidase</td>
<td>53.0</td>
<td>18.0</td>
</tr>
<tr>
<td>2 Aspartate aminotransferase-like</td>
<td>35.5</td>
<td>-7.2</td>
</tr>
<tr>
<td>3 $\beta$-Glucanase</td>
<td>20.0</td>
<td>-11.3</td>
</tr>
<tr>
<td>4 Xylose isomerase</td>
<td>34.5</td>
<td>11.5</td>
</tr>
<tr>
<td>5 Elongation factor TU</td>
<td>35.5</td>
<td>9.7</td>
</tr>
<tr>
<td>6 Signal recognition particle (SRP)</td>
<td>34.0</td>
<td>15.3</td>
</tr>
<tr>
<td>7 Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)</td>
<td>15.5</td>
<td>-0.4</td>
</tr>
<tr>
<td>8 Methylguanin-DNA methyltransferase</td>
<td>58.0</td>
<td>10.9</td>
</tr>
<tr>
<td>9 Carboxypeptidase</td>
<td>18.0</td>
<td>7.2</td>
</tr>
<tr>
<td>10 3-Isopropylmalate dehydrogenase (IPMDH)</td>
<td>38.0</td>
<td>15.3</td>
</tr>
<tr>
<td>11 Phosphoglycerate kinase</td>
<td>43.0</td>
<td>1.1</td>
</tr>
<tr>
<td>12 Cyclodextrin glycosyltransferase</td>
<td>25.0</td>
<td>29.8</td>
</tr>
<tr>
<td>13 Carbonic anhydrase</td>
<td>13.0</td>
<td>4.9</td>
</tr>
<tr>
<td>14 Cellulase</td>
<td>20.0</td>
<td>-16.8</td>
</tr>
<tr>
<td>15 Thermolysin-like protease (TLP)</td>
<td>25.0</td>
<td>23.6</td>
</tr>
<tr>
<td>16 Subtilase</td>
<td>34.0</td>
<td>2.8</td>
</tr>
<tr>
<td>17 Endocellulase</td>
<td>20.0</td>
<td>-25.9</td>
</tr>
<tr>
<td>18 Histidine carrier protein</td>
<td>18.0</td>
<td>16.2</td>
</tr>
<tr>
<td>19 Adenylate kinase, zinc finger</td>
<td>18.0</td>
<td>-4.7</td>
</tr>
</tbody>
</table>

*Optimal growth temperature difference of the corresponding species, in K.

In K, $\Delta T_p$ is marked in bold if the phase-transition temperature of the thermophilic protein was predicted to be higher than that of the mesophilic counterpart.
protein core of a thermophilic protein is more resistant to high temperatures than that of a mesophilic protein, which lends immediate support to the hypothesis that thermostability is linked to an enhanced structural rigidity of the folded native state. Values in Table I differ slightly from previously published ones,42 because, in the present study, a modified, fully automated method for identifying the phase-transition temperature was used (see Methods section). This method yields phase-transition temperatures that are determined in a more consistent manner across the whole data set.

In approximately one third of the investigated pairs of homologues, a higher phase-transition temperature for the thermophilic protein could not be found. The following reasons may account for this. First, in a few cases, mesophilic proteins have been reported to be as thermostable (if not more thermostable) than their thermophilic counterpart.58 Second, extrinsic factors such as glycosylation, salt, pressure effects, or solvent viscosity may further influence thermostability.59 As only for a few proteins in our data set melting temperatures are available, let alone data on the dependency of \( T_m \) on the solvent viscosity, we are currently unable to investigate this solvent influence any further. Finally, rigidity analysis is sensitive with respect to deficiencies in the structures43; in particular, as only one static representation per structure is analyzed. Analyzing multiple representations instead, for example, generated by MD simulations, and averaging over the results is expected to alleviate deficiencies in the constraint model and, hence, led to more robust results.60

**Identifying sites that are important for macroscopic stability**

To understand protein structure adaptations that lead to higher thermostability, we next consider microscopic properties of the protein structures in the context of the macroscopic percolation behavior of the constraint networks. Here, the unfolding simulations of IPMDH and TLP are considered in detail, and local structural regions are identified at which the unfolding starts. The identified unfolding regions are compared to experimental data.

**3-Isopropylmalate dehydrogenase**

3-Isopropylmalate dehydrogenase (IPMDH) is an enzyme in the leucine biosynthesis pathway that catalyzes the dehydrogenation and concomitant decarboxylation of 3-isopropylmalate (IPM) substrate, using nicotinamide adenine dinucleotide (NAD\(^+\)) as a cofactor. In the present study, IPMDH from *E. coli* and *T. thermophilus* is compared. IPMDH is a functional dimer, composed of two identical subunits (SU), each with \( \sim 350 \) amino acid residues.61 The polypeptide chain of one subunit is folded into two domains with similar topologies based on parallel \( \alpha/\beta \) motifs: domain 1 contains the N- and the C-terminus, and domain 2 contains the subunit interface. The active site is located in a cleft between the two domains and is made up of residues from both subunits.62

In Figure 1(c), cluster configuration entropy plots are shown for mesophilic and thermophilic IPMDH. Mesophilic and thermophilic IPMDH show a corresponding percolation behavior. The biologically relevant transitions, corresponding to the folded–unfolded transitions in protein unfolding, can be identified at 340 and 355 K for the *E. coli* and the *T. thermophilus* IPMDH, respectively. The difference in experimental optimal growth temperatures of the corresponding species is 38 K (Table I). The rigid-cluster decomposition of the networks before and after the phase transition is shown in Figure 2. The events that take place at the unfolding transitions in *E. coli* and *T. thermophilus* IPMDH are quite similar: before the transition, the giant cluster extends over the subunit-interface and the domain-interfaces, although only a few residues of domain 1 are incorporated by the giant cluster [Fig. 2(a,c)].

After the transition, the giant cluster does not cover the subunit-interface anymore [Fig. 2(b,d)]. Likewise, the rigid connection between domains 1 and 2 is lost. The close correspondence of the rigid cluster distribution in the constraint networks of the homologues before and after the phase transition is a striking result of our analysis. Particularly, the location of the giant cluster is similar within the pair of homologues. As for the size of the giant cluster, the one in the thermophilic structure just before the phase transition is larger than the one in the mesophilic protein. This is in agreement with the experiment, where it has been shown that the folding intermediate of the *E. coli* structure only retains 50% of the original secondary structure, whereas that of the *T. thermophilus* IPMDH retains 90%.64,65

By comparing the structure networks before and after the phase transition, we can identify microscopic structural features, where the unfolding starts, so-called unfolding regions or weak spots. The decay of the giant cluster in IPMDH initially causes residues at the domain interface, in domain 1, and at the subunit interface (domain 2) to become flexible. In Figure 2(b), these regions are indicated by arrows. The unfolding regions identified by the thermal-unfolding simulation of *E. coli* IPMDH have been compared to experiment (Table II). Encouragingly, sites within the predicted regions have also been successfully targeted by experiment to increase thermostability of various mesophilic66–71 and chimeric IPMDH.72–77 The experimental examination of thermophilic IPMDH has further revealed that most of the sites that are important for thermostability are at the subunit-interface,74,71,78 the domain-interface,73,79–81 and in domain 1.82,83 These regions are also detected as unfolding regions in our simulation for both the
mesophilic and the thermophilic protein (Table II). Overall, the agreement between the sites predicted by CNA and the experimentally identified unfolding regions is good. However, weak agreement is found for the domain 1 region, of which only a few residues are part of the giant cluster, although mutations in this domain have been shown to influence protein stability.82,83

Table II

<table>
<thead>
<tr>
<th>Structure region</th>
<th>Residues in predicted unfolding regions&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Experimentally verified thermostabilizing mutations&lt;sup&gt;k,h&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>III Domain 1</td>
<td>—</td>
<td>20, 24, 42, 57, 60–62, 84, 86, 92, 350, 356–360</td>
</tr>
</tbody>
</table>

<sup>a</sup>Numbering in the alignment, as shown in Figure S1.
<sup>b</sup>See text for details. According to ref. 66–74,77,79,81–83.
<sup>c</sup>Previously,42 these residues have been classified as belonging to the domain 1 region.

**Thermolysin-like protease**

Thermolysin-like proteases (TLP) are members of the peptidase family M4 of which thermolysin is the prototype.84 Thermolysin originates from *B. thermoproteolyticus*. TLP are proteases and require a Zn<sup>2+</sup> ion for their activity. They contain multiple Ca<sup>2+</sup> ions, which

---

**Figure 2**

Snapshots from the thermal unfolding simulation of meso- (**a, b**) and thermophilic (**c, d**) 3-isopropylmalate-dehydrogenase (IPMDH) just before (**a, c**) and after the phase transitions (**b, d**) at 340 and 355 K, respectively. The rigid cluster decomposition of the network is shown. The giant cluster is shown in blue. Other clusters are shaded black. Arrows in (**b**) indicate potential unfolding nuclei. Roman numbers refer to the numbering of the unfolding nuclei in Table II. The subunits and domains are marked. The locations of the active sites are indicated by asterisks. Figures were generated using PyMOL.63 [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
have been reported to be important for stability.\textsuperscript{85} TLP consists of an N-terminal (mainly $\beta$-sheet) and a C-terminal (mainly $\alpha$-helical) domain. Both domains are connected by a central $\alpha$-helix (residues 139–154), which is located at the bottom of the active site cleft and contains several of the catalytically important residues. In the present study, mesophilic TLP from \textit{B. cereus} and thermoophilic thermolysin is compared. The structures share a high sequence and structure similarity (Table S1). The two enzymes differ in their thermostability and in the level of enzyme activity, when measured at the same temperature.

The simulated thermal-unfolding behavior of \textit{B. cereus} TLP and thermolysin is similar, as expressed by a corresponding percolation behavior. The analysis of the cluster configuration entropy plot revealed that the biologically relevant transitions that correspond to the folded–unfolded transitions can be identified at 350 and 373 K, respectively [Fig. 1(d)]. The difference in experimental optimal growth temperatures of the corresponding species is 25 K (Table I). In Figure 3, the rigid cluster decomposition of the networks of \textit{B. cereus} TLP and thermolysin, respectively, is shown directly before and after the phase transition. Before the transition, the giant cluster dominates the system in both cases. Moreover, the giant cluster is located in the same region of the proteins: it extends over the N-terminal domain and comprises the $\beta$-sheet region and an $\alpha$-helix in the N-terminal domain (residues 67–89). The size of the giant cluster in thermolysin differs from that reported previously.\textsuperscript{42} This is because, in the present study, a modified method to determine the phase-transition temperature was applied (see Methods section).

In addition to these common features, some differences also exist. The networks also contain several smaller rigid clusters, mainly located in the C-terminal domain, but also in the N-terminal one. The size and number of these clusters vary, with more and larger rigid clusters existing in \textit{B. cereus} TLP compared to thermolysin.

After the phase transition, the giant cluster decays into smaller rigid clusters and regions that are flexible. Interestingly, the decay of the giant cluster is in the N-terminal domain, where the main site of proteolytic degradation is assumed. Thus, the thermal-unfolding simulation predicts that the global unfolding starts at similar sites as observed for local-unfolding events that lead to kinetic degradation.

Comparing the constraint networks before and after the phase transition allows identifying the microscopic structural origins of the thermostability of thermolysin. For both mesophilic TLP and thermolysin, similar unfolding regions can be identified. The first region comprises almost all residues of the $\beta$-sheet in the N-terminal domain, the second one a Ca$^{2+}$-binding site (residues 56, 58, 60, and 68), the hydrophobic region around...
by the emergence of flexible regions

Table III
Comparison of predicted with experimentally verified unfolding sites for mesophilic thermolysin-like protease (TLP)

<table>
<thead>
<tr>
<th>Structure region</th>
<th>Residues in predicted unfolding regions&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Experimentally verified thermostabilizing mutations&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>II Ca&lt;sup&gt;2+&lt;/sup&gt;-binding site, hydrophobic region around Phe 63, and N-terminus of the α-helix in the N-terminal domain</td>
<td>54, 56–62, 68–70</td>
<td>4, 8, 57, 59, 61, 64, 66–70</td>
</tr>
</tbody>
</table>

<sup>a</sup>Numbering in the alignment, as shown in Figure S2.

<sup>b</sup>See text for details. According to ref. 86–94.

Phe 63, and the N-terminal residues of the α-helix in the N-terminal domain (residues 67–89). The predicted unfolding regions are in good agreement with sites where stabilizing mutations have successfully been introduced experimentally into TLP to increase thermostability (Table III).<sup>86–94</sup> This holds true in particular for the unfolding region comprised residues 56–70 in the N-terminal domain (region II). Little correspondence, however, was found for the residues of the β-sheet region in the N-terminal domain (region I). No unfolding regions were found in the C-terminal domain. This is in agreement with the experimental studies showing that mutations at various sites in the C-terminal domain only have a marginal effect on thermal stability.<sup>93,95–99</sup>

For both IPMDH and TLP, we found a strong agreement between the events that take place at the unfolding transition in the mesophilic enzyme and its thermophilic homologue. In particular, the location and the size of the giant cluster were comparable. In both the mesophilic and the thermophilic enzyme, corresponding unfolding regions were identified. The only difference in the global-unfolding behavior between the mesophilic and the thermophilic protein was the higher thermostability of the unfolding regions. Apparently, thermostability is determined by the temperature dependence of local changes in flexibility and rigidity of the systems.

Microscopic stability is determined by the emergence of flexible regions

By analyzing the decay of the giant cluster, we identified the above unfolding regions that are important for macroscopic stability, because their breaking determined the global folding–unfolding transition. This demonstrates that, within a rigid cluster, a hierarchy of regions of varying stability can exist.<sup>50,100</sup> Revealing such hierarchies in detail provides information about microscopic stability features, which can be associated with local unfolding events. To characterize microscopic stability features from the thermal-unfolding simulation, “stability maps” are introduced here. For this, “rigid contacts” between two residues, represented by their Ca 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, $E_{\text{cut}}$ is identified at which a “rigid contact” between two residues is lost. [Note that, by Eq. (4), $E_{\text{cut}}$ also relates to a temperature at which the “rigid contact” is lost.] 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. Stability maps are comparable to cooperativity correlation plots used in related studies.<sup>101,102</sup> However, stability maps provide direct information on the distribution of rigidity and flexibility within the system, and they also identify regions that are flexibly or rigidly correlated across the structure. In turn, cooperativity correlation plots are calculated from an ensemble of realizations of a given state of the protein, and they primarily identify regions that are correlated across the entire ensemble of these realizations.<sup>101,102</sup>

In Figures 4 and 5, the stability maps for IPMDH and TLP are shown, respectively. The upper and lower triangles of the matrices display the stability maps for the mesophilic and the thermophilic homologues, with corresponding rows and columns determined by a structural alignment. Gaps in the structural alignment are colored in gray. Relative microscopic stability values are color coded in that contacts that are stable at temperatures above the phase-transition temperature are shown in red, whereas contacts that have already been broken when this temperature is reached are shown in white. For both IPMDH and TLP, the size and the localization of the giant cluster at temperatures slightly below the respective phase-transition temperature are similar for the mesophilic and the thermophilic proteins. Moreover, the size and the localization of many other rigid clusters are corresponding.

Histograms of the differences between microscopic stability values of rigid contacts in thermophilic IPMDH and TLP and microscopic stability values of corresponding rigid contacts in the mesophilic proteins are shown in Figure S3(a,b), respectively. A negative stability difference indicates that a rigid contact in the thermophilic protein is stronger than that in the mesophilic protein. For both IPMDH and TLP, stability differences mostly range from 0 to −2.0 kcal mol<sup>−1</sup>, with ~55 and 75%,
respectively, of the contacts in the range of 0 and 
−1.0 kcal mol⁻¹. Thus, it becomes apparent that, for the
thermophilic protein, the majority of rigid contacts are
stronger when compared with the mesophilic homologue,
but that the increase in strength per contact is moderate.

To finally quantify the relation of microstability features
of those pairs of meso- and thermophilic homologous
proteins in our data set for which the phase-transition
temperature of the thermophilic protein was predicted to
be higher than that of the mesophilic counterpart, corre-
lation coefficients were calculated for the respective stabi-
lity maps (Table IV). These values range from 0.25 to
0.90, with an average value of 0.53 ± 0.19, showing weak
to strong correlations in all cases. These correlations are
significant with P-value ≤ 0.05. For example, scatter plots
of the microscopic stability values of rigid contacts in
mesophilic versus thermophilic IPMDH and TLP, respec-
tively, are shown in Figure S4(a,b).

Overall, these results demonstrate that microscopic
stability features are conserved between mesophilic and
thermophilic homologues. In agreement with what has
been found when analyzing the general percolation beha-
vor based on global parameters, the results demonstrate
that mesophilic and thermophilic proteins are in corre-
sponding states of similar rigidity and flexibility, even at
the microscopic level. Such a conservation of microscopic
stability features strongly suggests that the features are
important for enzyme function and activity.¹,²⁴,¹⁰³

Relating microscopic stability and
enzyme activity

Experimentally, the interplay between flexibility, stabi-
lity, and activity has been studied by measuring kinetic
parameters that describe enzyme function and catalysis
(K_m and k_cat).³ However, these parameters do not
provide insights at an atomic level into the importance
of flexibility and rigidity for enzyme activity. A
more direct view on protein flexibility and microstability
is obtained by biophysical techniques that analyze

Figure 4
Combined stability map for E. coli and T. thermophilus IPMDH subunits (SU) 1 and 2. In the upper (lower) half of the matrix, stability
information for the mesophilic (thermophilic) protein is shown. Relative microscopic stability values are color coded: contacts that are stable at
temperatures above the phase-transition temperature are shown in red, and instable contacts are shown in white. Gaps in the alignment of both
structures or residues that do not form any rigid contacts are shown in gray. Catalytic residues are highlighted by black arrows, and other
functionally important residues are highlighted by gray arrows. Contacts between the residues in the giant cluster (at a temperature just below T_p)
are surrounded by a solid line. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
dynamical properties of the structure, such as H/D exchange, neutron scattering, and NMR spectroscopy. However, even then, it remains difficult to quantify the influence of flexibility and rigidity due to the different types of motions and timescales detected. Such difficulties may explain why, in some cases, thermophilic enzymes were unexpectedly found by H/D exchange to display a similar, or even increased, flexibility when compared with their mesophilic homologues.\textsuperscript{58,106–108}

Using CNA, no assumption needs to be made with respect to timescales and types of motions. This method allows studying rigidity and flexibility in atomic resolution and thus characterizing microscopic stability features. Here, we analyze how adaptive mutations of thermophilic enzymes maintain the balance between rigidity, important for macroscopic stability, and local flexibility, important for activity, regardless of the temperature at which the protein functions. The role of local flexibility is most apparent for regions involved in substrate binding and product release.

**Figure 5**
Combined stability map for \textit{B. cereus} thermolysin-like protease (TLP) and thermolysin. In the upper (lower) half of the matrix, stability information for the mesophilic (thermophilic) protein is shown. Relative microscopic stability values are color coded: contacts that are stable at temperatures above the phase-transition temperature are shown in red, and instable contacts are shown in white. Gaps in the alignment of both structures or residues that do not form any rigid contacts are shown in gray. Catalytic residues are highlighted by black arrows, and other functionally important residues are highlighted by gray arrows. Contacts between the residues in the giant cluster (at a temperature just below \(T_p\)) are surrounded by a solid line. Contacts between the residues belonging to the \(\alpha\)-helix in the N-terminus are surrounded by a dashed line. The latter region is shown enlarged. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

**Table IV**
Correlation between the stability maps from mesophilic and thermophilic homologues

<table>
<thead>
<tr>
<th>Protein family(a)</th>
<th>(r^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.46</td>
</tr>
<tr>
<td>4</td>
<td>0.30</td>
</tr>
<tr>
<td>5</td>
<td>0.48</td>
</tr>
<tr>
<td>6</td>
<td>0.55</td>
</tr>
<tr>
<td>8</td>
<td>0.69</td>
</tr>
<tr>
<td>9</td>
<td>0.35</td>
</tr>
<tr>
<td>10</td>
<td>0.65</td>
</tr>
<tr>
<td>11</td>
<td>0.66</td>
</tr>
<tr>
<td>12</td>
<td>0.33</td>
</tr>
<tr>
<td>13</td>
<td>0.53</td>
</tr>
<tr>
<td>15</td>
<td>0.74</td>
</tr>
<tr>
<td>16</td>
<td>0.25</td>
</tr>
<tr>
<td>18</td>
<td>0.90</td>
</tr>
<tr>
<td>Mean</td>
<td>0.53 ± 0.19</td>
</tr>
</tbody>
</table>

\(a\)Numbers refer to family numbers in Table I. Only those families were considered for which the phase-transition temperature of the thermophilic protein was predicted to be higher than that of the mesophilic counterpart.

\(b\)Pearson’s correlation coefficient.
3-Isopropylmalate dehydrogenase

The active site of 3-isopropylmalate dehydrogenase (IPMDH) is located in a cleft between the two subunits and is made up of residues from both subunits. The catalytic residues Lys 195 and Asp 227 belong to one subunit, whereas Tyr 145 belongs to the other (the numbering refers to the alignment shown in the Supporting Information). The catalytic residues are indicated by black arrows in Figure 4. The catalytic residues are located in domain 2 of the enzyme, which has an important role for macroscopic stability (see above). As illustrated in Figure 6(a,d) for meso- and thermophilic IPMDH, respectively, the active site is in the immediate vicinity of the giant cluster at the working temperature (see Methods section) of the corresponding enzyme (337 and 352 K, respectively), indicating that the catalytic residues are in a relatively stable region of the protein. Examining the stability map in Figure 4 reveals that contacts between the catalytic residues and their neighboring residues are relatively unstable when compared with contacts between residues in the giant cluster. (In Fig. 4, contacts between residues of the giant cluster are surrounded by a solid line.) Thus, the catalytic residues are not part of the giant cluster, but located in a locally flexible region, allowing a certain degree of flexibility for them, which is anticipated to be necessary for catalysis. Still, because of the vicinity to the giant cluster, a certain degree of global stability is retained.

As shown for the thermophilic enzyme, this view is corroborated in that the microstability values of the active site show that these residues become part of the giant cluster as early as at temperatures 7–10 K below the respective phase-transition temperature [Figs. 4 and 6(c)]. Thus, at the working temperature of the mesophilic enzyme (337 K), the active site of the thermophilic homologue is rigidified. It is highly likely that this is the

Figure 6
Active site of meso- (a, b) and thermophilic (c, d) 3-isopropylmalate dehydrogenase (IPMDH) at the working temperature of the mesophilic enzyme (337 K) (a, c) and the working temperature of the thermophilic enzyme (352 K) (b, d). The rigid cluster decomposition of the network is shown. The giant cluster is shown in blue. Other clusters are shaded black. Catalytic residues are shown in red, and other functionally important residues are shown in green.
reason why the *T. thermophilus* IPMDH has been found to be only marginally active then. In turn, at the working temperature of the thermophilic enzyme (352 K), the active site of the mesophilic homologue is identified to be completely flexible and, hence, has lost all structural stability [Fig. 6(b)].

Aside from the microscopic stability of the catalytic residues, the rigidity and flexibility characteristics of residues involved in substrate and cofactor binding are expected to have a prominent role for the activity of IPMDH, even more so, because, upon binding of IPM and NAD\(^+\), IPMDH shows an induced fit of the binding site. Especially, the cofactor binding triggers conformational changes in the immediate vicinity of both the adenosine- and the nicotinamide-binding sites: the C\(_\alpha\) atoms of residues 14–18, 284–299, and 334–347 in domain 1 and 261–266 at the domain interface move by up to 2.5 Å to bind the adenine moiety (the numbering refers to the alignment shown in Fig. S1). In turn, residues 77–91 in domain 1 move at the nicotinamide-binding site. These residues are indicated by gray arrows in Figure 4. All these residues are located in loops, with the exception of residues 16–18 and 90–91, which are located at the amino terminal ends of \(\alpha\)-helices. The longest of the moving segments comprises residues 284–299. This loop closes over an analog of NAD\(^+\), ADP-ribose, when bound to IPMDH, allowing Asp 298 and Leu 290 to bind the adenine moiety.

As a prerequisite for these movements, contacts between residues in the moving regions and neighboring regions need to be unstable. This is shown in the stability map (see Fig. 4), where it can be seen that the microscopic stabilities of these residues are in fact low: residues 14–18 and 77–91 become rigidified (i.e., they are integrated into the giant cluster) only at temperatures 10–20 K below the phase-transition temperature [Fig. 6(c)]. At this stage, the regions cannot move toward the binding site anymore, and the closed conformation of IPMDH cannot be adopted, which may be another reason why the thermophilic enzyme is only marginally active at the temperature of maximal activity of the mesophilic enzyme [Fig. 6(a)]. Notably, cold-active forms of *T. thermophilus* IPMDH have been engineered by introducing mutations into exactly these fragments in order to improve binding by favoring the closed conformation.

### Thermolysin-like protease

The active site of TLP is located in a cleft between the N- and the C-terminal domain. TLP binds a single-catalytic Zn\(^{2+}\) ion. Two mechanisms of action have been proposed for TLP: one in which Glu 144 acts as a proton acceptor in catalysis, and a second in which His 232 acts as a general base instead of the glutamate (the numbering refers to the alignment shown in Fig. S2). In either mechanism, residues Glu 144, His 232, Tyr 158, and a Zn\(^{2+}\) bound water play important roles during catalysis. The Zn\(^{2+}\) is coordinated by two histidine residues (His 143 and His 147), a glutamate (Glu 167), and a water molecule. Most of the catalytic and Zn\(^{2+}\)-coordinating residues belong to the central \(\alpha\)-helix (residues 139–154) that represents the bottom of the binding site. The catalytic and Zn\(^{2+}\) coordinating residues are indicated by black arrows in Figure 5.

For TLPs, a hinge-bending motion has been postulated to be important for substrate binding and catalysis. In this context, the central \(\alpha\)-helix (residues 139–154) has been identified as a flexible joint. Particularly, residues Gly 136 and Gly 137 at the N-terminus of this \(\alpha\)-helix are important. In addition, Gly 79 in the center of an \(\alpha\)-helix in the N-terminal domain (residues 67–89) has been found to be important for hinge bending. These residues are indicated by gray arrows in Figure 5. Mutations of residue 79 influenced substrate binding by altering mobility in the binding cleft. Mutations of residues 136 and 137 affected motions directly involved in catalysis and, hence, influenced \(k_{\text{cat}}\).

The microscopic stabilities of the functionally important residues 79, 136, and 137 can be determined from the unfolding simulation. To this aim, the stability map of meso- and thermophilic TLP in Figure 5 is examined. It is expected that contacts between the functionally important residues and their neighboring residues are unstable, thus allowing for a local flexibility. For Gly 79, this can be observed primarily in the case of mesophilic TLP: contacts between residue 79 and other residues of the long \(\alpha\)-helix in the N-terminal domain (residues 67–89) are relatively unstable compared to mutual contacts of other residues within the \(\alpha\)-helix. (In Fig. 5, contacts between these residues are surrounded by a black-dashed line, and this region is shown enlarged.) The result shows that the influence of residue 79 is limited to a local region, most likely because Gly 79 acts as a \(\alpha\)-helix breaker.

The importance of residues 136 and 137 as hinge residues can also be understood from examining the stability map. Contacts between the residues of the giant cluster and contacts between residues of a large cluster in the C-terminal domain 138–208 are relatively stable, whereas contacts between residues 125–137 are relatively unstable. At temperatures just below the respective phase-transition temperature, residues 125–137 form a flexible region that assures that the giant cluster does not extend into the C-terminal domain and does not fuse with residues 138–208 forming the other cluster. That way, residues 125–137 can act as a hinge. As shown in Figure 7(a,d) for meso- and thermophilic TLP, respectively, some of the catalytic and Zn\(^{2+}\)-coordinating residues belong to the large cluster in the C-terminal domain at the working temperature (see Methods section) of the corresponding enzyme (342 and 364 K, respectively). Thus, these residues are in a locally rigid region, which can nevertheless move globally due to the hinge. The global mobility and
local rigidity characteristics have been confirmed by recent H/D exchange experiments that were performed to study the conformational dynamics of free thermolysin.\textsuperscript{122}

As shown for thermolysin, at temperatures 20–30 K below the phase-transition temperature, the global mobility is lost, because the hinge region rigidifies [Fig. 7(c)]. This happens because contacts are then established between residues of the hinge (125-136) and residues of the giant cluster in the C-terminal domain (137-208). This result explains experimental findings that show that thermolysin is only marginally active at a temperature that corresponds to the $T_{og}$ of the mesophilic enzyme, 25 K below its own $T_{og}$.\textsuperscript{123} In turn, at a temperature where the thermophilic enzyme is predicted to be fully active (364 K), the active site of the mesophilic homologue is completely flexible [Fig. 7(b)].

The importance of the flexibility of the hinge region has also been confirmed by mutational studies.\textsuperscript{90,124} Mutating Leu 145 by Ser increased $k_{cat}$ and, thus, improved the enzyme's activity but did not influence the thermostability of the protein.\textsuperscript{124} On the other hand, mutating Gly 120 to Glu improved the activity, too, but had a negative effect on macroscopic stability.\textsuperscript{90} These observations can be understood with the help of the CNA results: whereas the mutation of residue 145 was performed in a region that is not important for macroscopic stability, the mutation of residue 120 targets an identified unfolding region. Apparently, activity could be modulated independently from stability when a mutation was introduced at a site that does not correspond to an unfolding region. It is encouraging to note that CNA allows understanding this complex relationship between activity and thermostability.

![Figure 7](image-url)

Active site of meso- (a, b) and thermophilic (c, d) thermolysin-like protease (TLP) at the working temperature of the mesophilic enzyme (342 K) (a, c) and the working temperature of the thermophilic enzyme (364 K) (b, d). The rigid cluster decomposition of the network is shown. The giant cluster is shown in blue. Other clusters are shaded black. Catalytic residues are shown in red, and other functionally important residues are shown in green.
CONCLUSIONS

On the basis of the constraint network representations of proteins, and by applying rigidity theory, we analyzed the local distribution of flexible and rigid regions in 19 pairs of homologous proteins from meso- and thermophilic organisms and related it to activity characteristics of the enzyme structures. The present study thus extends a previous one by us, which related protein structural rigidity and thermostability.42

From a biological point of view, the most intriguing result of our study is that adaptive mutations of thermophilic enzymes maintain the balance between overall rigidity, important for thermostability, and local flexibility, important for activity, at the respective temperature at which the protein functions: our results demonstrate that thermophilic adaptation leads to an increase of structural rigidity in general, but that the distribution of functionally important flexible regions is conserved between meso- and thermophilic homologues. This finding provides direct evidence for the hypothesis of corresponding states. To the best of our knowledge, this study is the first one probing this hypothesis for a large data set by computational means and by directly characterizing protein rigidity and flexibility. Our results demonstrate that exploiting the principle of corresponding states not only allows for successful thermostability optimization but also for guiding experiments in order to improve enzyme activity in protein engineering.

In more detail, our results show that specific regions are crucial for macroscopic stability of protein structures. When these unfolding nuclei become flexible, the structural integrity of the native folded state is lost, and the protein unfolds. Notably, unfolding nuclei identified for IPMDH and TLP are in good agreement with sites that are known to be important for thermostability and where thermostabilizing mutations have been introduced. Thus, we suggest that the unfolding regions represent sites where mutations should be introduced that might modulate a protein’s thermostability. Detailed analysis of the flexibility in the IPMDH and TLP substrate-binding regions with the help of “stability maps,” introduced here for the first time, showed that enzyme function requires certain regions of the protein to be flexible at the working temperature of the respective enzyme. We were able to qualitatively relate changes in the flexibility of those regions, induced either by a temperature change or by the introduction of mutations, to experimentally observed losses of the enzymes’ functions.

From a methodological point of view, it is important to note that protein thermostability can be predicted by characterizing the mechanical rigidity of a protein structure during a thermal-unfolding simulation. CNA is based on a graph-theoretical method that determines rigidity and flexibility within a protein structure in atomic resolution. The approach is extremely fast and computationally inexpensive. Moreover, no assumption on the type and the timescale of motion is made. Remarkably, CNA implicitly captures and unifies many different mechanisms that contribute to increased thermostability and to activity at high temperatures. Thus, CNA provides an appealing alternative to MD simulations or experiments. Finally, the approach is entirely based on the native topology of a protein. This can be rationalized in that we are identifying when the giant rigid cluster breaks, which is controlled by native contacts.

Network representations of proteins have already been used to predict flexibility and stability characteristics of protein structures. Vendruscolo et al.125,126 showed that protein structure networks are “small world” networks. In such networks, atoms or residues can be identified that determine the topology of the network.45 Heringa and Argos127 used protein structure networks to identify clusters of tightly packed side chains. Later on, they showed that these clusters were important for protein thermostability.128 Brinda et al.129,130 used a similar approach to identify differences between meso- and thermophilic proteins. Robinson-Rechavi et al.19 used such an approach to identify structural properties of thermophilic proteins, too. In the present study, we go beyond analyzing topological properties of protein structure networks. Instead, protein structures are represented as constraint networks (molecular frameworks), where rigid and flexible regions can be identified.42

A shortcoming of our method is the fact that a structure of high quality is needed for analysis. Moreover, rigidity analysis is sensitive with respect to deficiencies in the structures, and extrinsic factors such as glycosylation, salt, pressure effects, or solvent viscosity may further influence thermostability. Thus, future work is needed to make the approach viable also in those cases where only low-resolution structures or homology models are available and/or to include such extrinsic factors. An approach to reduce sensitivity toward a single structure has been developed by Jacobs et al.101,102 Using the distance constraint model (DCM), a large number of combinations of constraints is generated as an input to rigidity analysis, and the results are then averaged. However, the DCM requires a fit to experimental data, which limits the general applicability of the method.

Compared to other computational methods for protein engineering,131–133 CNA provides an interesting alternative. Most importantly, we expect CNA to be a valuable tool for predicting “weak spots” in a protein structure, where mutations could improve microscopic stability and consequently macroscopic stability, ultimately leading to an increased thermostability. In addition, CNA provides detailed information about regions that should not be stabilized in order to not interfere with enzyme activity. Such information will be very valuable, because even in an elaborate study describing the successful design of a thermostable enzyme,131 only coarse assumptions have
been made so as to not influence the active site’s stability during the design process.

**METHODS**

**Data set**

A detailed description of the data set selection is given in Ref. 42. Here, we summarize the procedure. Crystallographic models of homologous pairs of mesophilic and thermophilic protein structures were collected from the Protein Data Bank (PDB)\(^{134}\) using data sets from the recent literature as a starting point.\(^{15,130}\) Only homologous pairs were considered that fulfilled the following criteria: (i) the resolution is \(\leq 2.6\) Å; (ii) structures were excluded in cases where the crystallographic model did not correspond to the biological unit; (iii) the quality of the structures was checked using the PDBREPORT database,\(^{135}\) and structures with the qualification “bad” were excluded. Excluded structures were replaced by other available structures of the same protein if possible.

For enzymes, information about catalytic residues was taken from the Catalytic Site Atlas.\(^{136}\) To ensure structural homology, only sequences sharing \(\geq 35\%\) residue identity were considered, as were only pairs of structures that shared the same oligomeric and ligand-binding state and had a similar resolution. In some cases, it was not possible to find structures that met all of the above criteria. In the case of the thermophilic protein of family 10 (IPMDH), a mutant was taken, because the wild-type enzyme was only available as a monomer from the PDB. The biological unit, however, is a homodimer. Nevertheless, the mutant shows only slight differences in thermostability and no difference in enzyme activity when compared with the wild-type protein.\(^{80}\) For further details on other protein families, see ref. 42.

The final data set contained 19 homologous pairs of mesophilic and thermophilic protein structures (Table S1). The data set differs from a previously published one,\(^{42}\) in that every protein family is now being represented by only one pair of structures. Structural similarity was determined using the DaliLite server.\(^{137}\) Where necessary, the crystallographic model was modified by flipping Asn, Gln, or His side chains using REDUCE.\(^{138}\) The optimal growth temperature \((T_{\text{opt}})\) was assigned to each protein by considering the living environment temperature or the average of a range of habitat temperatures of the corresponding species.\(^{21,139}\) Melting temperatures were taken from the literature and the ProTherm database.\(^{140}\)

**Constraint network analysis**

**Network construction**

To construct the constraint network, only the protein structure was used. Water, solvent molecules, buffer ions, substrates, and cofactor molecules were removed from the crystallographic models. Metal ions were retained when they were part of the structure. Bonds between ions and protein atoms were treated as covalent bonds and inserted manually. The constraint network of the covalent and noncovalent bonds present within the protein was constructed using the FIRST software (version 6.2).\(^{43}\) Hydrogen bonds (including salt bridges) were modeled by a bond between the hydrogen and the acceptor atom and two additional angular constraints associated with these atoms.\(^{43}\) Hydrogen-bond energies \(E_{\text{hb}}\) were calculated from an empirical potential based on donor-hydrogen-acceptor atom geometries as given in the crystallographic model.\(^{51}\) The position of the hydrogen atoms was used as predicted by the REDUCE software.\(^{138}\) Hydrophobic interactions between carbon or sulfur atoms were taken into account if the distance between these atoms was less than the sum of their van der Waals radii (C: 1.7 Å, S: 1.8 Å) plus 0.25 Å.\(^{50}\)

**Rigid cluster decomposition**

Rigidity within the constraint network can be fully characterized using constraint counting.\(^{48}\) To this aim, in FIRST, the “pebble game” algorithm is implemented.\(^{43}\) This algorithm determines for every bond whether it is part of a rigid cluster or a flexible region.\(^{48,141,142}\) A rigid cluster is defined as a set of atoms that move together as a rigid body in any floppy motion. Otherwise, an atom is part of a flexible region. The size of a rigid cluster was determined as the number of atoms that form the cluster. Cluster size was further normalized by the total number of atoms within the network. By definition, the giant cluster is the largest cluster present in a given network.

**Thermal-unfolding simulation**

By gradually removing noncovalent bond constraints from the network, one can go from a rigid network with a large number of bonds to a flexible network with only a few bonds. Knowing the energies \(E_{\text{hb}}\) of all potential hydrogen bonds (including salt bridges) in the network, the energy values were used for excluding hydrogen bonds from the network with \(E_{\text{hb}} > E_{\text{cut}}\). A new network was generated for every single hydrogen bond that was removed. The networks were then decomposed into rigid clusters, as described earlier. As the energy of a hydrogen bond relates to a temperature \(T\) at which the bond breaks\(^{43,51}\) (see also below), thermal unfolding can be simulated that way.\(^{50}\) The number of hydrophobic contacts was kept fixed during the thermal-unfolding simulation.

**Network parameter evaluation**

Global changes in the rigidity characteristics of the network during the thermal-unfolding simulation were characterized using the rigidity-order parameter \((P_{\infty})\)
and the \( s^2 \)-cluster configuration entropy (\( H \)) as defined in the Theory section. A phase transition was identified as the most prominent change in a \( H \) versus \( E_{\text{cut}} \) plot. To do so, a spline was fitted to the entropy values. \( E_{\text{cut,p}} \) is the hydrogen bond energy cutoff where the first-order derivative of the spline reaches its maximum.

To determine a phase-transition temperature \( T_p \) from \( E_{\text{cut,p}} \) a relation \( T = f(E_{\text{cut}}) \) needs to be established. For this, we calibrated \( E_{\text{cut}} = E_{\text{cut,p}} \) values with respect to experimental melting temperatures \( T_m \), considering only those pairs of meso- and thermophilic proteins for which \( E_{\text{cut,p}} \) of the mesophilic protein was correctly predicted to be larger than \( E_{\text{cut,p}} \) of the thermophilic protein. For those proteins, for which only optimal growth temperatures of the corresponding species (\( T_{\text{opt}} \)) were available, \( T_m \) values were estimated by assuming that they are 25 K above the \( T_{\text{opt}} \) value.\(^\text{143}\) Optimal growth temperatures have previously been used in studies comparing meso- and thermophilic proteins.\(^\text{16}\) This resulted in the following relation:

\[
T = -20 \frac{K}{(\text{kcal mol}^{-1})} E_{\text{cut}} + 300 \text{ K} \quad (4)
\]

Considering that, at \( E_{\text{cut}} \), only hydrogen bonds with \( E_{\text{hh}} \leq E_{\text{cut}} \) are retained in the network, Eq. (4) also provides a mapping of the procedure for removing hydrogen bonds onto a scale of temperature. In contrast to previous studies\(^\text{43,60}\) where the hydrogen bond energy was related to thermal fluctuations in terms of \( kT \), Eq. (4) provides a less steep relation of hydrogen bond energy with temperature. We note that Eq. (4) works well for determining differences in \( T_p \) for pairs of homologous proteins and to relate these to differences in experimentally determined melting temperatures (\( T_m \)) or optimal growth temperatures of the corresponding species (\( T_{\text{opt}} \)). However, the relation does not allow to compare \( T_p \) values between nonhomologous proteins. This is because the size of a constraint network and the topology of the underlying protein influence the absolute \( T_p \) value.

Local stability characteristics were identified based on a stability map. At a given temperature, rigidity analysis labels portions of the protein as rigid or flexible. A stability map shows how pairs of residues at a given temperature are related to each other in terms of their rigidity characteristics. On the map, a color is assigned to a pair of residues according to the stability of their “contact.” Here, two residues form a “contact” if both are part of the same rigid cluster. The contact stability is then defined as the \( E_{\text{cut}} \) value at which the contact breaks during the thermal-unfolding simulation. Stability maps of the same size can be compared by measuring Pearson’s correlation of the two matrices. For pairs of homologous proteins, corresponding residue positions in the sequence were identified based on structural alignments obtained from the DaliLite server.\(^\text{137}\)

For comparing a pair of homologous proteins in their respective functional states, working temperatures for the enzymes were determined, such that (i) these temperatures are lower by the same amount than the respective \( T_p \) values and (ii) structural elements that need to move for the enzymes to be functional come out as flexible in the rigidity analysis. This leads to working temperatures 3 K below the respective \( T_p \) values in the case of IPMDH and 8 K in the case of TLP.

**ACKNOWLEDGMENTS**

We are grateful to Daniel Cashman, Doris Klein, Prakash Rathi (Heinrich Heine-University, Düsseldorf), and Benjamin Radestock (Johannes Gutenberg-Universität, Mainz) for critically reading the manuscript.

**REFERENCES**


103. Somero GN. Temperature adaptation of enzymes—biological optimi-
104. Hollen J, Marqusee S. Structural distribution of stability in a ther-
106. Fitter J, Heberle J. Structural equilibrium fluctuations in meso-
1636.
107. Hernandez G, Jenney FE, Adams MWW, LeMaster DM. Milli-
108. Hernandez G, LeMaster DM. Reduced temperature dependence of collective conformational opening in a hyperthermophile rub-
109. Hurley JH, Dean AM. Structure of 3-isopropylmalate dehydro-
111. Suzuki T, Yasugi M, Arisaka F, Oshima T, Yamagishi A. Cold-adapt-
tion mechanism of mutant enzymes of 3-isopropylmalate dehy-
112. Yasugi M, Amino M, Suzuki T, Oshima T, Yamagishi A. Cold ad-
aptation of the thermophilic enzyme 3-isopropylmalate dehydro-
116. Mock WL, Aksamawati M. Binding to thermolysin of phenolate-
118. Fitter J, Heberle J. Structural distribution of stability in a ther-
121. Holland DR, Tronrud DE, Pley HW, Flaherty KM, Stark W, Jans-
122. Liu YH, Konermann L. Conformational dynamics of free and cata-
lytically active thermolysin are indistinguishable by hydrogen/deu-
terium exchange mass spectrometry. Biochemistry 2008;47:6342–
6351.
123. de Kreij A, van den Burg B, Venema G, Vriend G, Eijsink VG, Nielsen JE. The effects of modifying the surface charge on the cata-
124. Yasukawa K, Inouye K. Improving the activity and stability of ther-
126. Mock WL, Aksamawati M. Binding to thermolysin of phenolate-