Global and Local Indices for Characterizing Biomolecular Flexibility and Rigidity

Christopher Pfleger,[a] Sebastian Radestock,[a] Elena Schmidt,[b] and Holger Gohlke*[a]

Understanding flexibility and rigidity characteristics of biomolecules is a prerequisite for understanding biomolecular structural stability and function. Computational methods have been implemented that directly characterize biomolecular flexibility and rigidity by constraint network analysis. For deriving maximal advantage from these analyses, their results need to be linked to biologically relevant characteristics of a structure. Such links are provided by global and local measures (‘indices’) of biomolecular flexibility and rigidity. To date, more than 14 indices are available with sometimes overlapping or only vague definitions. We present concise definitions of these indices, analyze the relation between, and the scope and limitations of them, and compare their informative value. For this, we probe the structural stability of the calcium binding protein α-lactalbumin as a showcase, both in the ‘ground state’ and after perturbing the system by changing the network topology. In addition, we introduce three indices for the first time that extend the application domain of flexibility and rigidity analyses. The results allow us to provide guidelines for future studies suggesting which of these indices could best be used for analyzing, understanding, and quantifying structural features that are important for biomolecular stability and function. Finally, we make suggestions for proper index notations in future studies to prevent the misinterpretation and to facilitate the comparison of results obtained from flexibility and rigidity analyses. © 2012 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23122

Introduction

Flexibility and its opposite, rigidity, are important characteristics of protein stability and function.[1–3] As such, the mechanical heterogeneity of proteins is a prerequisite for proper enzyme function. Consequently, the distribution of flexible and rigid regions is highly conserved within homologous proteins.[3–5] Likewise, orthologs from meso- and thermophilic organisms are in states of corresponding flexibility at their respective working temperatures.[6] Being able to identify flexible and rigid regions as well as changes in flexibility/rigidity on changes in a protein’s environment, for example, due to binding of a ligand, temperature or solvent change, is essential for understanding protein stability and function. From an application point of view, such information provides a means for optimizing protein stability and function by rational protein engineering[6–9] and is valuable for structure-based ligand design.[10–12]

Flexibility and rigidity are static properties that denote the possibility of motions but do not give any information about directions and magnitudes of actual motions. The flexibility of a biological macromolecule is typically characterized by measuring motions of the structure using techniques such as hydrogen/deuterium (H/D) exchange, neutron scattering, and relaxation measurements by nuclear magnetic resonance (NMR) spectroscopy.[3] However, the experimental characterization and quantification of biomolecular flexibility remains challenging, especially with respect to the diverse types and time-scales of motions.[2,3] Moreover, it remains difficult to determine biomolecular flexibility that is related to function because such flexibility can be limited to small but crucial parts of the structure.[13] Thus, computational methods provide a valuable complement for analyzing flexibility and rigidity of biomolecules.

On the other hand, methods are applied that analyze conformational ensembles, either determined by crystallography or NMR spectroscopy or generated by molecular dynamics (MD) simulations.[14] The outcome of these methods obviously depends on the number and diversity of states in the ensemble. On the other hand, methods have been devised that identify flexible and rigid regions from a single input structure, for example, by determining the spatial variation in the local packing density,[15] or by representing the structure as a connectivity network of interacting residues or atoms.[16–21]

Following another network concept, protein structures are modeled as constraint networks (molecular frameworks), where vertices represent atoms and edges represent covalent bonds and angles.[22] For accurately modeling biomolecular flexibility, noncovalent interactions must also be included in this network.[23–26] Flexible and rigid regions are then determined from the number and spatial distribution of bond-rotational degrees of freedom.[27,28] The Pebble Game algorithm,[27–29] implemented in the Floppy Inclusion and Rigidity

[a] C. Pfleger, S. Radestock, H. Gohlke
Department of Mathematics and Natural Sciences, Institute for Pharmaceutical and Medicinal Chemistry, Heinrich-Heine-University, Düsseldorf, Germany
E-mail: gohlke@uni-duesseldorf.de

[b] E. Schmidt
Department of Biological Sciences, Goethe-University, Frankfurt am Main, Germany

© 2012 Wiley Periodicals, Inc.
Substructure Topology (FIRST) software\textsuperscript{[23]} efficiently assigns each bond as either being part of a flexible or a rigid region. A rigid region results from a collection of interlocked bonds that have no relative motion. If the rigid region has redundant constraints, it is overconstrained. Otherwise, it is isostatically rigid. In a flexible region, the dihedral rotation of one bond is not locked in by other bonds. The theory underlying this approach is rigorous\textsuperscript{[30]} and the parameterization for modeling the constraint network has frequently and successfully been applied in various analyses of biomolecules.\textsuperscript{[31–34]} In addition, comparisons between the constraint network and dynamical approaches have been performed in the past. These include the analysis of changes in the flexibility of proteins on complex formation by constraint network analysis (CNA) and MD simulations\textsuperscript{[26]} as well as a large-scale comparison of protein essential dynamics from MD simulations and coarse-grained normal mode analyses\textsuperscript{[35]} which use information from CNA as input.\textsuperscript{[34]}

On diluting constraints in a network, a phase transition occurs at which the network loses its ability to carry stress.\textsuperscript{[36]} In particular, diluting constraints that model noncovalent interactions in a protein allows simulating the thermal unfolding of the biomolecule.\textsuperscript{[6,7,24,37–39]}

Linking results from CNA to biologically relevant characteristics of a structure is key for deriving maximal advantage from information on biomolecular flexibility. In this context, biologically relevant characteristics are, for example, the (macroscopic) phase transition point where a biomolecule switches from a structurally stable (largely rigid) state to an unfolded (largely flexible) state or the (microscopic) localization and distribution of structurally weak parts. As links, global and local measures ("indices") of biomolecular flexibility and rigidity are applied. To date, more than 14 indices have been introduced, which are summarized in Supporting Information, Table S1. A subset of four global and four local indices that are particularly useful will be described in more detail below and will be related to case studies where the indices have been successfully applied for explaining and interpreting experimental findings. Sometimes, only vague index definitions have been provided from the original authors, leading to similar or identical indices being used under different names. In other cases, the scope of an index is limited to only a subdomain of flexibility and rigidity analysis. Finally, a comparison of indices with respect to their informative value is elusive except for a comparative study of metrics used within the distance constraint model (DCM).\textsuperscript{[39]}

Thus, this study aims at (i) providing concise definitions of the indices, (ii) analyzing the relation between, and the scope and limitations of, indices, and (iii) comparing their informative value. We also introduce three new indices that allow extending the applicability domain of flexibility and rigidity analysis to understanding and improving thermostability, analyzing flexibility changes on complex formation and mutations, and investigating how flexibility information is transmitted between sites in a protein. The majority of these indices can be computed by the CNA package\textsuperscript{[40]} developed in our laboratory, which functions as a front- and backend to the FIRST software. As a test system, we analyze the structural stability of the calcium binding protein \(\alpha\)-lactalbumin both in the ‘ground state’ and after perturbing the system by changing the network topology. Furthermore, we concentrate on monitoring changes in a network along an unfolding trajectory rather than investigating a single static network state. With these showcase analyses, we intend to provide guidelines for future studies suggesting which of these indices could best be used for analyzing, understanding, and quantifying biologically relevant characteristics that are important for protein stability and function.

### Materials and Methods

#### Structure preparation

The structure of \(\alpha\)-lactalbumin determined by X-ray crystallography to 1.7 Å resolution\textsuperscript{[21]} was taken from the Protein Data Bank (PDB code 1HML).\textsuperscript{[42]} The quality of the structure was checked using the PDBREPORT database.\textsuperscript{[43]} Hydrogens were added by the REDUCE program,\textsuperscript{[44]} and, where necessary, Asn, Gln, or His side chains were flipped. The secondary structure information of \(\alpha\)-lactalbumin is summarized in Table 1.

#### Constraint network construction

To construct the constraint network, only the protein molecule was used, whereas water and buffer ions were removed. We decided to not include water molecules in the network based on previous findings\textsuperscript{[26,45]} that showed only a negligible difference in the flexibility characteristics of proteins when structural waters were considered. In fact, when we performed thermal unfolding simulations of \(\alpha\)-lactalbumin with and without crystal water (in the former case, only those water molecules with a distance < 3.5 Å to a protein atom were considered), the computed indices did not change (data not shown). However, we note that in certain cases water can have a pronounced effect on a protein’s flexibility.\textsuperscript{[23]} The molecular network of the covalent and noncovalent bond constraints present within the protein was constructed using the FIRST software (version 6.2).\textsuperscript{[23]} In addition, metal ions were retained when they were part of the structure. Here, a calcium ion is coordinated to the two backbone carbonyl oxygens of K79 and D84, and to the three carboxylate side chains of D82, D87, and D88. Additionally, a zinc ion is coordinated to the carboxyl group of E49. Interactions between ions and protein atoms were treated as covalent bonds and inserted manually.

| Table 1. Domain and secondary structure information for human \(\alpha\)-lactalbumin. |
|--------------------------|-----------------|
| Domain or secondary structure name | Residues |
| \(\alpha\)-domain | K1—T38, D83—L123 |
| \(\beta\)-domain | Q39—D82 |
| Helix A | K5—L11 |
| Helix B | L23—S34 |
| Helix C | T86—K98 |
| Helix D | A106—L110 |
| Strand S1 | I41—E43 |
| Strand S2 | T48—Y50 |
| Strand S3 | I55—S56 |
Hydrogen bonds, salt bridges, and hydrophobic contacts were considered as noncovalent bond constraints as described previously.\textsuperscript{[40]} Hydrogen bond energies \( E_{HB} \) were calculated using an empirical potential.\textsuperscript{[46]} 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 \( \text{\AA} \), S: 1.8 \( \text{\AA} \)) plus 0.25 \( \text{\AA} \).\textsuperscript{[28]} That way, the influence of solvent on the protein stability is considered implicitly.

Then, flexible and rigid regions within the constraint network are identified by FIRST. The algorithm and the underlying theory have been described elsewhere.\textsuperscript{[27–29]}

Constraint network analysis

CNA has been developed in our laboratory with the aim to analyze the structural features that are important for a biomolecule's stability. CNA functions as a front- and backend to FIRST by (i) setting up a variety of constraint network representations for rigidity analysis, (ii) processing the results obtained from FIRST, and (iii) computing indices for characterizing molecular stability both globally and locally. With respect to (i), CNA provides a method to simulate the thermal unfolding of a protein structure by removing noncovalent constraints from the network in the order of increasing strength.\textsuperscript{[6,7]} In this study, only hydrogen bonds (including salt bridges) were removed. For each threshold value \( E_{\text{cut}} \), a new network state \( \sigma = R(T) \) was generated where only hydrogen bonds with an energy \( E_{HB} \leq E_{\text{cut}} \) were considered. This follows the idea that weaker hydrogen bonds will break at lower temperatures than stronger ones. Each network \( \sigma \) was then decomposed into flexible and rigid regions, producing a thermal unfolding trajectory. The number of hydrophobic contacts was kept constant during the simulation. This is motivated by the fact that hydrophobic interactions actually remain constant or even become stronger when the temperature increases.\textsuperscript{[47,48]} We performed two thermal unfolding simulations of \( \alpha \)-lactalbumin, one in which the calcium ion was included and another one in which the ion was removed from the constraint network. The zinc ion was present in both simulations. We note that we applied CNA to only a single structure in the present study but that it can be applied to an ensemble of conformations, too, then yielding averaged index values.\textsuperscript{[39]}

Results

Unfolding simulation of \( \alpha \)-lactalbumin

\( \alpha \)-Lactalbumin is a metallo protein that regulates the lactose biosynthesis by modulating the specificity of galactosyltransferase. The protein contains two distinct ion binding sites: one site is located at the connection of \( \alpha \)- and \( \beta \)-domain (K79–D88) and binds a calcium ion; the second site is located at the active site of \( \alpha \)-lactalbumin and binds a zinc ion. Several studies have demonstrated that the binding of calcium affects the function of \( \alpha \)-lactalbumin,\textsuperscript{[49–52]} whereas the binding of zinc has only a negligible effect to the structure.\textsuperscript{[52]} It has been suggested that binding of calcium induces a conformational transition and enhances the thermal stability of \( \alpha \)-lactalbumin.\textsuperscript{[49]} Accordingly, the removal of calcium causes a transition from a well-ordered, rigid to a less-ordered, floppy binding site.\textsuperscript{[52]}

To relate information about \( \alpha \)-lactalbumin's mechanical flexibility and rigidity to thermal stability, a thermal unfolding of the structure was simulated. Simulating the thermal unfolding of \( \alpha \)-lactalbumin with CNA requires less than 2 min on a standard workstation computer. Snapshots from the unfolding trajectory of \( \alpha \)-lactalbumin are depicted in Figure 1. They show the loss of rigidity in terms of the decay of rigid clusters with increasing temperature. Arrows point to states where the network undergoes a transition as indicated by a sudden drop in rigidity. The first transition relates to the beginning of the collapse of the largest rigid cluster, which is located in the \( \alpha \)-domain of \( \alpha \)-lactalbumin. Still, the cluster continues to dominate the network after this transition. At the second transition point, the cluster stops dominating the network and breaks
into smaller rigid components. Thus, rigidity ceases to percolate through the structure, that is, the structure is not able to transmit stress anymore after this transition. Finally, during the last transition, the rigid cluster capturing the \( \beta \)-domain of \( \alpha \)-lactalbumin collapses, and nearly the whole system becomes flexible. We will use these transition points for the evaluation and comparison of computed global and local indices.

Global flexibility indices

Floppy mode density. Global flexibility indices monitor the degree of flexibility and rigidity within the constraint network at a macroscopic level. One such global index is the number of internal independent degrees of freedom (floppy modes, \( F \)) that are associated with dihedral rotations in a network.\(^{[24,36]} \) By monitoring \( F \) during a thermal unfolding simulation, a phase transition can be detected that relates to the transition from an amorphous, rigid to a polymeric, glassy state in random networks\(^{[36]} \) as well as from a structurally stable (rigid) to an unfolded (flexible) state in proteins.\(^{[24]} \) Usually, \( F \) is normalized by the number of (overall) internal degrees of freedom associated with the \( N \) atoms, resulting in a floppy mode density \( \Phi \), eq. (1))

\[
\Phi = \frac{F}{6N-6} \tag{1}
\]

Figure 2a shows the change in \( \Phi \) during the thermal unfolding simulation of \( \alpha \)-lactalbumin as a function of \( E_{\text{cut}} \). \( E_{\text{cut}} \) relates to...
the mean coordination number \(<r>\) in the constraint network in that \(<r>\) is a monotonic function of \(E_{\text{cut}}\). Previously, \(\Phi = f(<r>)\) has been used in such analyses,\(^{24}\) predominantly in the case of network glasses.\(^{54-56}\) However, for applications to biological systems, \(\Phi = f(E_{\text{cut}})\) seems to be a more intuitive choice because \(E_{\text{cut}}\) can be related to the temperature of the protein system.\(^{61}\) In fact, the shape of \(d^2\Phi/dE_{\text{cut}}^2\) (inset in Fig. 2a) resembles the curve of the specific heat for a phase transition of second order,\(^{24,57,58}\) and its maximum at \(-1.59\) kcal mol\(^{-1}\) corresponds to the second transition in Figure 1. The \(d^2\Phi/dE_{\text{cut}}^2\) curve also exhibits a weak shoulder around \(-0.5\) to \(-1.0\) kcal mol\(^{-1}\), which is related to the earlier transition.

**Mean rigid cluster size.** Originating from percolation theory, moments of the size distribution of rigid clusters (i.e., the microstructure of the network) can be used to analyze macroscopic properties of constraint networks.\(^{59}\) In this context, the change of the mean rigid cluster size \(S\) can be monitored during the thermal unfolding simulation, with the size of the largest rigid cluster always being excluded from the calculation (Supporting Information, Table S1). This leads to \(S\) being zero as long as one rigid cluster dominates the whole network or if all rigid clusters have vanished. Figure 2b shows the changes of \(S\) for \(\alpha\)-lactalbumin as a function of \(E_{\text{cut}}\). Here, \(S\) is not zero at high \(E_{\text{cut}}\) because even then the largest cluster does not cover the whole network. By removing hydrogen bond constraints from the network, the rigid cluster starts to decay, which leads to a steep increase of \(S\) at \(-0.61\) kcal mol\(^{-1}\). At this point, the system is still dominated by one large rigid cluster (see also Fig. 1). In the following, the mean cluster size only monitors the collapse of smaller rigid components that have segregated from the largest cluster. The second transition then highlights the (loss of) rigidity percolation at \(-1.55\) kcal mol\(^{-1}\), which matches the second transition in Figure 1. At this point, the largest rigid cluster collapses as reflected by the steepest increase of the curve in Figure 2b. After a plateau phase, \(S\) starts to decrease due to the fact that the network is less and less dominated by rigid components.

**Rigidity order parameter.** As another index originating from percolation theory and derived from the microstructure of a constraint network, the rigidity order parameter \(P_{\infty}\) has been used to identify transition points during thermal unfolding.\(^{6,7,54,59-61}\) Here, the fraction of the network belonging to the giant percolating cluster (type 1) or the actual largest rigid cluster (type 2) is chosen as an order parameter (Supporting Information, Table S1). That is, compared to \(S\), the extremum of the cluster size distribution is considered. The giant percolating cluster is defined as the largest rigid cluster present at high \(E_{\text{cut}}\) values with all constraints in place (i.e., at low temperatures). As long as the network is in the rigid phase, it is dominated by one rigid cluster and, hence, \(P_{\infty}\) is close to one. In the floppy phase, with a vanishing largest rigid cluster, \(P_{\infty}\) is zero.

Figures 2c and 2d show \(P_{\infty}\) for type 1 and type 2 as a function of \(E_{\text{cut}}\). The giant percolating cluster corresponds to the actual largest rigid cluster at \(-0.30\), \(-1.00\), and \(-2.00\) kcal mol\(^{-1}\) (Fig. 1). At \(-3.00\) kcal mol\(^{-1}\), this identity ceases to exist because now the actual largest cluster is located in the \(\alpha\)-domain, whereas the giant percolating cluster remains in the \(\beta\)-domain of \(\alpha\)-lactalbumin. Accordingly, both \(P_{\infty}\) are identical until \(E_{\text{cut}} = -2.22\) kcal mol\(^{-1}\), where \(P_{\infty}\) type 1 drops to zero. In contrast, \(P_{\infty}\) type 2 continues monitoring the decay of the actual largest rigid cluster until it drops close to zero at \(-3.55\) kcal mol\(^{-1}\). Note the stepwise decrease of the \(P_{\infty}\) curves that reflects a process of multiple smaller transitions during the thermal unfolding. Both rigidity order parameters show distinct transitions at \(-0.61\), \(-1.55\), and \(-2.22\) kcal mol\(^{-1}\), which match the findings from Figure 1.

**Entropy associated with the rigid cluster size distribution.** The cluster configuration entropy \(H\), introduced by Andraud et al.\(^{62}\) as a morphological descriptor for heterogeneous materials, is a particularly useful index for characterizing macroscopic properties of a constraint network in terms of its microstructure. \(H\) has been adapted from Shannon's information theory [eq. (2)] and, thus, is a measure of the degree of disorder in the realization of a given state.

\[
H = - \sum_i w_i \ln w_i \quad (2)
\]

It is defined as a function of the probability that an atom is part of a cluster of size \(s\) (\(s\)-cluster) [eq. (3)]

\[
w_i = \frac{s^n_i}{\sum_i s^n_i} \quad (3)
\]

with \(n_i\) being the cluster number normalized by the total number of atoms \((N)\) [eq. (4)]

\[
n_i = \frac{\text{Number of clusters of sizes}}{N} \quad (4)
\]

Previously, \(w_i\) was calculated for \(k = 1\) in the context of a statistical analysis of the complexity in Monte-Carlo sampled networks,\(^{63}\) which corresponds to the original definition by Andraud et al. \((H_{\text{type } 1})\). A modified version \((H_{\text{type } 2})\) has been introduced by Radestock and Gohlke\(^{67}\) using \(k = 2\) (\(s^2\)-cluster).

As long as the largest rigid cluster dominates the system, \(H\) is zero because there is only one realization of the system possible. For the same reason, \(H\) is zero if all atoms can move independently. In between, \(H\) is nonzero because of multiple possible realizations of the system associated with a heterogeneous cluster size distribution. Figures 2e and 2f show \(H_{\text{type } 1}\) and \(H_{\text{type } 2}\) as a function of \(E_{\text{cut}}\). In the case of \(H_{\text{type } 1}\), two transitions at \(-0.61\) and \(-1.55\) kcal mol\(^{-1}\) can be identified. These transitions reflect changes in the network when the largest rigid cluster starts to decay or stops dominating the network (see also Fig. 1). In addition, a third transition can be identified at \(-2.22\) kcal mol\(^{-1}\) in the case of \(H_{\text{type } 2}\), which corresponds to the collapse of the largest rigid cluster, as can be two later transitions at \(-2.99\) and \(-3.54\) kcal mol\(^{-1}\) indicative of the final loss of the remaining rigid components.

**Local indices**

**Flexibility index.** Local indices characterize the network flexibility and rigidity down to the bond level. The flexibility index \(g_i\),
implemented in the FIRST program is a local analog of the floppy mode density $\Phi$ and quantifies the degree of flexibility or rigidity of a subcomponent of the network.\textsuperscript{[23,64]} The current FIRST implementation models biomolecules as body-bar networks. Here, atoms are treated as bodies with six degrees of freedom, and covalent and noncovalent interactions are modeled by a different number of bars that connect bonded atoms. Each bond $i$ is part of a network subcomponent in one of four states: (i) a dangling end, (ii) a flexible collective mode, (iii) an isostatically rigid region, or (iv) an overconstrained region. For bonds in a dangling end, $g_i = 1$. For bonds that are part of a flexible collective mode $j$, $0 < g_j < 1$ [eq. (5)]:

$$g_i = \frac{F_j}{6E_j - B_j}$$  \hspace{1cm} (5)

$F_j$ is the number of independently rotatable bonds in $j$, $E_j$ is the number of edges that represent rotatable bonds, and $B_j$ is the total number of bars from flexible bonds. Thus, $g_i$ relates the number of independently rotatable bonds to the number of potentially rotatable bonds in this case. In an isostatically rigid region where the number of internal degrees of freedom equals the number of constraints, $g_i = 0$. Finally, in an overconstrained region $k$, $-1.0 < g_k < 0$ [eq. (6)]:

$$g_k = -\frac{C_k - (6V_k - 6)}{(6V_k(6V_k - 1) - (6V_k - 6)}$$  \hspace{1cm} (6)

$C_k$ is the total number of constraints, and $V_k$ indicates the number of atoms in that region. Thus, $g_k$ relates the number of redundant constraints to the maximal number of redundant constraints in this case.

Note that with respect to a previous definition of the flexibility index $f_i$,\textsuperscript{[23]} $g_i = f_i$ for collective modes, but $g_i \neq f_i$ for overconstrained regions. The latter is because $f_i$ relates the number of redundant constraints to the actual number of all constraints (Supporting Information, Table S1).

The index $g_i$ is calculated from a single network state and does not require a thermal unfolding simulation. For visualizing $g_i$ results, it is often useful to derive an atom-based flexibility index as an average over $g_i$ values of bonds an atom is involved in. For example, a flexibility index for $C_a$ atoms can be calculated by averaging over the two backbone bonds (N$\cdots$C$\alpha$ and $C\alpha$--C). Figure 3 shows $g_i$ of $C\alpha$ atoms for two different network states of $\alpha$-lactalbumin before ($E_{cut} = -1.46$ kcal mol$^{-1}$) and after ($E_{cut} = -1.66$ kcal mol$^{-1}$) the phase transition where the largest rigid cluster stops dominating the network (see also Fig. 1). The $g_i$ for the region K79–D84 increases from 0.00 to 0.12, reflecting that this region now forms a flexible hinge region between two newly formed clusters that originated from the largest rigid cluster. Likewise, $g_i$ for the region G100–L110 increases, which indicates that the largest rigid cluster decays resulting in a larger hinge region.

As a downside of the index definition for overconstrained regions, $g_i$ values close to zero are obtained for these regions due to the denominator being dominated by the maximal number of redundant constraints in the region. This hampers a differentiation of regions with varying degrees of rigidity, at least if the same scale is used for $g_i$ values in flexible and overconstrained regions. A more fundamental drawback of the index arises from the fact that $g_i$ values for the majority of all regions of the network after the phase transition, that is, in a structurally less stable state, are lower than $g_i$ values for the network before the transition. Counter intuitively, the index indicates that the network has become less flexible within collective modes and more rigid in overconstrained regions, although constraints have been removed from the network.

**Percolation index.** Here, we introduce for the first time the percolation index $p_\pi$, which is a local analog to the rigidity order parameter $P_\infty$ in that it monitors the percolation behavior of a biomolecule on a microscopic level. As such, it allows identifying the hierarchical organization of the giant percolating cluster during a thermal unfolding simulation. The index value $p_i$ is derived for each covalent bond $i$ between two atoms $A_{i(1,2)}$ by the $E_{cut}$ value during a thermal unfolding.
simulation at which, the bond segregates from the giant percolating cluster \( c_{\text{gpc}} \) [eq. (7)]:

\[
\rho_i = \min\{E_{\text{cut}}|A_1 \land A_2 \in c_{\text{gpc}}(E_{\text{cut}})\}
\]  

(7)

For a \( C_\alpha \) atom-based representation, the lower of the two \( \rho_i \) values of the two backbone bonds is taken. \( \rho_i = 0 \) then indicates that an atom has never been part of the giant percolating cluster, that is, the atom has always been in a flexible region of the biomolecule. In contrast, the lower \( \rho_i \) the longer is a residue part of the giant percolating cluster during the thermal unfolding simulation. The secondary structure profile of \( \alpha \)-lactalbumin is given at the top. On the right, the respective index is mapped in a color-coded fashion on the \( \alpha \)-lactalbumin structure.

Figure 4a shows the \( C_\alpha \) atom-based \( \rho_i \) for \( \alpha \)-lactalbumin. The lower \( \rho_i \) the longer is a residue part of the giant percolating cluster during the thermal unfolding simulation. The horizontal lines correspond to the three transitions as depicted in Figure 1. The secondary structure profile of \( \alpha \)-lactalbumin is given at the top. On the right, the respective index is mapped in a color-coded fashion on the \( \alpha \)-lactalbumin structure.

Figure 4b shows the \( C_\alpha \) atom-based \( \rho_i \) for \( \alpha \)-lactalbumin. The lower \( \rho_i \) the longer is a residue part of a rigid cluster during the thermal unfolding simulation. The secondary structure profile of \( \alpha \)-lactalbumin is given at the top. On the right, the respective index is mapped in a color-coded fashion on the \( \alpha \)-lactalbumin structure.

Figure 4c shows the \( C_\alpha \) atom-based protection factors from H/D exchange experiments in the molten globule state of \( \alpha \)-lactalbumin. The data were taken from Schulman et al.\textsuperscript{[65]} The orange rectangles in b) and c) mark those protein regions that agree in terms of large structural stability and high protections factors.

**Figure 4.** a) Percolation index \( \rho_i \) for \( \alpha \)-lactalbumin. The lower \( \rho_i \) the longer is a residue part of the giant percolating cluster during the thermal unfolding simulation. The horizontal lines correspond to the three transitions as depicted in Figure 1. The secondary structure profile of \( \alpha \)-lactalbumin is given at the top. b) Rigidity index \( r_i \) for \( \alpha \)-lactalbumin. The lower \( r_i \) the longer is a residue part of a rigid cluster during the thermal unfolding simulation. The secondary structure profile of \( \alpha \)-lactalbumin is given at the top. On the right, the respective index is mapped in a color-coded fashion on the \( \alpha \)-lactalbumin structure. c) Protection factors from H/D exchange experiments in the molten globule state of \( \alpha \)-lactalbumin. The data were taken from Schulman et al.\textsuperscript{[65]} The orange rectangles in b) and c) mark those protein regions that agree in terms of large structural stability and high protections factors.
decays. The decay occurs as a multistep process in which residues collectively segregate from the rigid core. Three main steps can be identified, which correspond to the three transitions in Figure 1: first, the N-terminus including helix A and residues D37–A40 segregate at $p_i = -0.62$ kcal mol$^{-1}$; second, helix B and D segregate at $p_i = -1.56$ kcal mol$^{-1}$; finally, the most stable region in the $\beta$-domain collapses at $p_i = -2.22$ kcal mol$^{-1}$.

Additionally, we also analyzed the effect of perturbing the constraint network topology by removing the calcium ion (Fig. 5a). The binding region of the ion (K79–D88) is mostly affected by the removal in that the region now segregates earlier during a thermal unfolding simulation, equivalent to a lower structural stability. In addition, a lower structural stability is found for the region I89–W104. As W104 is about 15 Å away from the ion binding site, this demonstrates the long-range aspect of rigidity percolation in such networks.

Rigidity index. As a generalization of the percolation index $p_i$, we introduce the rigidity index $r_i$ here for the first time. The index is defined for each covalent bond $i$ between two atoms $A_{(1,2)}$ as the $E_{cut}$ value during a thermal unfolding simulation at which the bond changes from rigid to flexible. Phrased differently, this index monitors when a bond segregates from any rigid cluster $c$ of the set of rigid clusters $C^{\infty}$ [eq. (8)]

$$r_i = \min \{ E_{cut} | \exists c \in C^{\infty} : A_1 \land A_2 \in c \}$$

For a $C_a$ atom-based representation, the average of the two $r_i$ values of the two backbone bonds is taken. Accordingly, $r_i = 0$ indicates that an atom has always been in a flexible region of the biomolecule. In contrast, the lower $r_i$ the longer is a residue part of a rigid cluster during the thermal unfolding simulation.

Figure 4b shows the $C_a$ atom-based $r_i$ for $\alpha$-lactalbumin. As expected, secondary structure elements are identified as regions of highest structural stabilities. The most buried helix B ($r_i = -3.00$ to $-3.55$ kcal mol$^{-1}$) is also the most stable component followed by helices C ($r_i = -3.18$ kcal mol$^{-1}$) and A ($r_i = -2.79$ kcal mol$^{-1}$). These findings are in very good agreement with protein regions that have high protection factors according to H/D exchange experiments in the molten globule state of $\alpha$-lactalbumin (Fig. 4c). The two spikes in Figure 4b at residues L23 and V66 reveal that both residues are captured in small clusters that remain rigid until the end of the thermal unfolding simulation.

The effect of perturbing the network topology by removal of the calcium ion is shown in Figure 5b. This affects the first residues K79–D84 of the ion binding site in a similar manner as detected by the percolation index $p_i$ (Fig. 5a). In contrast, $r_i$ reveals a stronger effect on residue I85, whereas no change of $r_i$ is observed for residues T86–W104. The sole exception in the latter region is residue G100, which becomes less stable according to $r_i$. These results demonstrate that the information derived from $r_i$ and the percolation index $p_i$ (see above) is complementary: although the stability within helix C remains unaffected by the ion removal as revealed by $r_i$ (Fig. 5b), the...
percolation index $p_c$ shows that the helix as a whole now segregates earlier from the giant percolating cluster (Fig. 5a). The latter index thus shows at which places in the structure the change of the global percolation behavior due to the ion removal manifests. Furthermore, $r_c$ but not $p_c$ exposes I85 and G100 as potential hinges for the movement of helix C because these residues become locked in only when the helix is fixed at its top by the ion to other parts of the protein. In summary, these results can be interpreted in that ion removal makes helix C movable as a rigid body as it is then encompassed by two hinges. These findings are in very good agreement with those from an experimental study on bovine $\alpha$-lactalbumin\[66\] where helix C and the adjacent helical element (C77–K80) change their relative orientation and where the opposite face of the calcium binding site is perturbed, as implicated from comparing the apo and calcium-bound structures.

**Stability maps.** Stability maps $r_{cij}$ have been introduced by Radestock and Gohlke.\[6] A stability map is a two-dimensional itemization of the rigidity index $r_i$ and is derived by identifying ‘rigid contacts’ between two residues $R_{ij}$ which are represented by their $C\alpha$ atoms. A rigid contact exists if two residues belong to the same rigid cluster $c$ of the set of rigid clusters $C^{\mathrm{rig}}$. 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 [eq. (9)].

$$r_{cij} = \min\{E_{\text{cut}} | 3c \in C^{\text{rig}} : R_i \land R_j \in c\} \quad (9)$$

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 flexibility and rigidity 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 increasing $E_{\text{cut}}$. In Figure 6, the stability map for $\alpha$-lactalbumin is shown. The upper and lower triangles of the map have been derived for the constraint network with and without the calcium ion, respectively. Again, residues belonging to the calcium binding site are mostly affected by the ion removal. Furthermore, the maps intriguingly reveal that losses of rigid contacts do not only occur between isolated pairs of residues but also in a cooperative manner. That is, parts of the protein break away from the rigid cluster as a whole, as can be seen for helices A, B, and C.

**Discussion**

We provided concise definitions of indices for deriving biologically relevant characteristics of a biomolecule from rigidity analysis. The majority of these indices are computed by monitoring the changes in a network along an unfolding trajectory of the showcase example $\alpha$-lactalbumin. The trajectory was computed by consecutively removing hydrogen bond constraints using the CNA package. Here, only a single input structure was used as a starting point for building up the constraint network. Different conformations of a biomolecule can vary in the noncovalent bond network and, hence, can lead to a different outcome of the rigidity analysis as observed by us\[26\] and others.\[45\] In general, this problem can be overcome by analyzing an ensemble of constraint networks, for example, generated from a conformational ensemble obtained by MD simulation.\[26,39,40\] However, we note that using a single structure does not provide a limitation for the current study because here we only compare indices with respect to each other.

A related approach is provided by the DCM.\[31,67,68\] Here, 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 that DCM requires knowledge of experimentally determined heat capacity curves for a protein-specific parameterization of the model.\[8,69\]

We will now analyze the relation between, and the scope and limitations of indices used in connection with CNA and compare their informative value.

**Relations among indices**

The main purpose of applying global indices is to identify (a) phase transition point(s) where a molecule switches from a rigid to a flexible state. This is relevant for analyzing the thermostability of proteins\[16,7,37\] or changes in a protein's global stability on binding of a ligand\[26\] (see Table 2). Monitoring the floppy mode density $\Phi$ provides a measure for the intrinsic
Table 2. Application of global and local indices for explaining and interpreting experimental findings.

<table>
<thead>
<tr>
<th>Application</th>
<th>Experimental data</th>
<th>Index type</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyzing thermostability of and identifying 'weak-spots' in biomolecules</td>
<td>Either optimal growth temperatures of the organism or experimentally determined melting temperatures of the biomolecules. 'Weak-spots' were compared with identified folding cores from H/D exchange NMR studies as well as mutation experiments.</td>
<td>Rigidity order parameter $P_{\infty}$, Cluster configuration entropy $H$</td>
<td>Radstock and Gohlke[62], Rathi et al.[59]</td>
</tr>
<tr>
<td>Identification of folding cores in biomolecules</td>
<td>Folding cores as predicted by H/D exchange NMR experiments.</td>
<td>Rigidity order parameter $P_{\infty}$, Cluster configuration entropy $H$</td>
<td>Rader[37]</td>
</tr>
<tr>
<td>Analyzing the loss of structural stability in biomolecules</td>
<td>Compared to the unfolding behavior of network glasses upon melting.</td>
<td>Floppy mode density $\Phi$</td>
<td>Rader et al.[24]</td>
</tr>
<tr>
<td>Analyzing the flexibility of substrate-binding regions in enzymes</td>
<td>Comparison of different structural information as well as thermal mobility (B-factor) determined by X-ray crystallography.</td>
<td>Stability maps $r_{cij}$</td>
<td>Radstock and Gohlke[6]</td>
</tr>
<tr>
<td>Analyzing the flexibility in proteins as well as RNA structures</td>
<td>Thermal mobility (B-factor) determined by X-ray crystallography.</td>
<td>FIRST flexibility index $f_i$, FIRST flexibility index $g_i$</td>
<td>Jacobs et al.[25], Fulle and Gohlke[26,71,72], Tan and Rader[24]</td>
</tr>
<tr>
<td>Analyzing changes in protein flexibility upon protein-protein complex formation</td>
<td>Thermal mobility (B-factor) determined by X-ray crystallography and atomic fluctuations by MD simulations.</td>
<td>FIRST flexibility index $f_i$, Gohlke et al.[26]</td>
<td></td>
</tr>
</tbody>
</table>

[a] The author used the notations $X_C$ and $X_m$ that match the definition of the type 2 rigidity order parameter $P_{\infty}$; (b) FIRST dilution plots are graphical representations of the rigid cluster decomposition along the thermal unfolding simulation. The percolation index $p$, and the rigidity index $r$, are measures that allow a numerical interpretation of the dilution plots.

flexibility along the unfolding trajectory. Alternatively, indices have been adapted from percolation theory by analyzing the microstructure of the network. These indices are derived from properties of the set of rigid clusters generated along the unfolding trajectory.[60] The rigidity order parameter $P_{\infty}$ considers the extremum, that is, the largest rigid cluster, at a certain network state. In contrast, the largest rigid cluster is excluded in the case of the mean rigid cluster size $S$; therefore, $S$ monitors the size distribution of smaller rigid clusters. The cluster configuration entropy $H$ is a morphological descriptor of heterogeneity in networks.[62] Compared to the original implementation $H_{\text{type } 1}$, the alternative $H_{\text{type } 2}$ is more sensitive with respect to transition points that occur later in the unfolding trajectory, that is, when the network is already largely flexible. Both indices $S$ and $H$ monitor the complexity of the network; phrased differently, they measure the degree of disorder in the realization of a given network state.

The main purpose of applying local indices is to monitor the location and distribution of structurally weak or strong parts in biomolecules. This is relevant for guiding protein engineering efforts aimed at identifying unfolding nuclei (structurally weak parts) that, when mutated, may lead to an increase in the thermostability (Table 2).[6,7] These indices can either reflect structural stability on a per-residue basis or characterize correlations of stability between pairs of residues. The FIRST flexibility index $g_i$ belongs to the first class and is a local analog of the global index $\Phi$ in regions of collective modes. The index monitors how degrees of freedom and redundant constraints, respectively, are distributed throughout the network. It is the only index presented here that is computed from a static network state. The percolation index $p$, is a local analog of $P_{\infty}$ and monitors the percolation behavior of the giant percolating cluster on a microscopic level. The rigidity index $r$, is a generalization of $p$, and monitors the transition when a residue segregates from any rigid cluster. Stability maps belong to the second class and characterize the correlation of stability in biomolecules; hence, they itemize the information provided by $r$, that they reflect the microscopic stabilities between all pairs of residues in the network. The relationship between global and local indices is shown in Figure 7.

Ambiguity in index definitions

To date, at least 14 indices have been introduced in the literature with sometimes overlapping or identical definitions. Rader et al.[61] have used a global order parameter $X_C$ with the goal to describe the percolation behavior in biomolecules. This index definition matches type 2 of the rigidity order parameter $P_{\infty}$; $P_{\infty}$ had been introduced before by Stauffer.[60] Likewise, a parameter that considers the fraction of the network belonging to the percolating rigid cluster was used in the work of Chubynsky and Thorpe.[55] The definition of this parameter matches type 1 of $P_{\infty}$. We thus recommend using the original $P_{\infty}$-based notation in future studies for clarity. The local indices $p$, and $r$, introduced in this study can be seen as ‘envelopes’ of spikes in FIRST dilution plots.[70] FIRST dilution plots
have been used as graphical representations of the rigid cluster decomposition along a thermal unfolding simulation so far. The local indices \( p_i \) and \( r_i \) allow exploiting the information from a rigid cluster decomposition in a quantitative manner (Fig. 5), for example, to calculate the stability change of a biomolecule on ligand binding or mutation. With a similar aim, Rader introduced a local index \( P_{\text{rc}} \) for the projection of the percolation behavior on a microscopic level.\[37\] This index monitors the propensity of a residue for being part of the largest rigid cluster along an unfolding trajectory. Rader’s definition is related to the percolation index \( p_l \) defined here for the first time. In our opinion, analyzing the giant percolating cluster as in \( p_l \) instead of the actual largest rigid cluster as in \( P_{\text{rc}} \) is more appropriate for identifying that part of a protein from where rigidity starts to propagate through the network.

Supporting Information, Table S1 also contains four index definitions used in the DCM.\[9\] The global DCM flexibility index \( \theta \) is the average of the number of independent degrees of freedom \( F \) over the DCM ensemble. The local DCM flexibility index \( \vartheta \) is the ensemble average of the local density of floppy modes and redundant constraints\[8\] and, therefore, related to the FIRST flexibility index \( f \).\[18\] A second local index \( P_{\text{R}} \) has been introduced for quantifying backbone flexibility by monitoring the probability whether backbone dihedral angles \((\phi, \psi)\) are rotatable over the ensemble.\[68\] This definition is related to the rigidity index \( r_l \) that monitors when a bond segregates from a rigid cluster during the thermal unfolding. Finally, cooperativity correlation plots are provided by DCM\[68,69\] that quantify the correlated stability for pairs of residues in terms of rotatable dihedral backbone angles. Related to this, stability maps \( r_{ci} \) monitor the correlated stability between pairs of residues according to how long a rigid contact exists along the unfolding trajectory. Note that, although index definitions used within the frameworks of CNA or DCM are related, the way they are derived substantially differs: in that CNA monitors changes in the constraint network along a thermal unfolding trajectory, whereas DCM performs Monte-Carlo sampling of network topologies at a fixed temperature. We thus recommend using the respective CNA- or DCM-based index notations in future studies to make these differences clear.

### Informative value, scope, and limitations of global indices

We applied four global indices for analyzing the unfolding trajectory of \( \alpha \)-lactalbumin, the floppy mode density \( \Phi \), the mean rigid cluster size \( S \), the rigidity order parameter \( P_{\text{so}} \), and the cluster configuration entropy \( H \). In general, all global indices were able to identify transition points in agreement with results from visual inspection of the unfolding trajectory. However, the indices dramatically differ in terms of the sensitivity, which leads to different numbers of transitions being identified. Monitoring \( \Phi \) allows identifying those transitions points where the structural features in the network change dramatically. As such, \( \Phi \) identifies the second transition when the largest rigid cluster stops dominating the network but fails at identifying the other two transitions. The index \( S \) identifies the first and second transitions and, thus, detects the beginning of the network collapse; however, it fails at identifying the last transition. The indices \( P_{\text{so}} \), type 1 and 2 detect all three transition points identified visually; the type 2 index additionally identifies two more transitions highlighting the collapse of last rigid fragments. As for index \( H \), type 1 and 2 are complementary: \( H_{\text{type 1}} \) identifies the earlier transition points when the network is dominated by the largest rigid cluster, whereas \( H_{\text{type 2}} \) identifies the later transition points when the network consists of multiple smaller rigid clusters.

The loss of stability in biomolecules during thermal unfolding is a multistep process.\[7\] This is because biomolecules have a hierarchical organization in terms of structural stability as opposed to network glasses. This has implications for index sensitivity (early vs. late transitions) and the application domain of indices (e.g., folding core detection vs. thermostability analysis). The index \( \Phi \) is particularly insensitive for detecting any transition points other than the one associated with the rigidity percolation threshold because the change in the number of independent degrees of freedom markedly changes only at this point.\[36\] Accordingly, when applied for analyzing the (un-)folding of mainly globular biomolecules, \( \Phi \) only provides a description of a two state nature of this transition, that is, of the biomolecule switching from a globally stable to a flexible state. \( \Phi \) has been used for analyzing the loss of structural stability on protein unfolding and for comparing a protein’s unfolding behavior to that of network glasses on melting by Rader et al.\[24\] (Table 2). In contrast, indices \( S \) and \( H \), based on the cluster size distribution of the network, show an increased sensitivity with respect to the detection of additional transition points. This is because removing a single constraint can lead to the collapse of a rigid region, which can strongly affect the cluster size distribution even if the network is already in the floppy state. Both indices \( S \) and \( H_{\text{type 1}} \) are most sensitive for earlier transitions in unfolding and, hence, best

![Figure 7. Relationship between global and local indices. Fine dashed arrows indicate ‘local analog of global count’ relationship. Coarsely dashed arrows indicate a ‘generalization’ relationship.](image)
applied for analyzing the loss of the rigidity percolation in biomolecules. Phrased differently, \(H_{\text{type } 1}\) is preferably applied for detecting when the first fragments segregate from the largest rigid cluster in a network. In contrast, \(H_{\text{type } 2}\) is most sensitive for later transitions in unfolding and, hence, best applied for analyzing the decay of the largest rigid fragments.\(^{[27]}\) Consequently, \(H_{\text{type } 2}\) was used to analyze the shift in the melting temperature of 20 pairs of orthologous proteins from mesophilic and thermophilic organisms\(^{[6,7]}\) and five citrate synthase structures that cover a wide range of thermostability\(^{[39]}\) (Table 2).

Finally, \(P_{\infty}\) shows relatively pronounced steps also in between the transition points due to monitoring the decay of the giant or largest rigid cluster and, thus, is even more sensitive to network changes than \(S\) or \(H\). This makes \(P_{\infty}\) better suited for a detailed analysis of the loss of stability in a biomolecule, for example, when monitoring the segregation of secondary structure units from the core\(^{[38]}\) rather than for identifying transition points in an automated fashion, for example, as needed when computing melting points of a protein over an ensemble of structures.\(^{[6,7]}\) Along these lines, \(P_{\infty}\) and the related index \(X_{\text{c}}\) were used as absolute measures to identify key amino acids that are important for the stability of rhodopsin\(^{[61]}\) and as relative measures to analyze the global shift in stability on ligand binding to HIV-1 gp120\(^{[64]}\) and the thermostability of rubredoxin structures\(^{[37]}\) (Table 2). Notably, homologous proteins have \(P_{\infty}\) curves of a very similar shape,\(^{[6,7]}\) reflecting the evolutionary conservation of flexibility and rigidity in proteins and providing direct evidence for the hypothesis of corresponding states for orthologs from meso- and thermophilic organisms.\(^{[6]}\)

Besides providing qualitative information that allow for the successful prediction of folding cores\(^{[61]}\) and weak spots,\(^{[6,7,37,39]}\) some of the indices have also been used to connect with quantitative experimentally measurable data. In particular, this holds for (changes in) the stability of proteins measured in terms of melting temperatures (\(T_m\)) or, more indirectly, in terms of optimal growth temperatures (\(T_{\text{opt}}\)) of the respective organisms. DCM index definitions describe the Landau free energy \(G(T,\theta)\) as a function of the temperature \(T\) and a global flexibility order parameter \(\theta\) and thus directly relate to a protein’s thermostability.\(^{[8,68]}\) Likewise, Radestock and Gohlke found a linear relationship between computed phase transition temperatures and experimental \(T_m\) or \(T_{\text{opt}}\)\(^{[6,7]}\) which was also applied in subsequent studies.\(^{[37,39]}\)

**Informative value, scope, and limitations of local indices**

We applied the four local indices FIRST flexibility index \(g_i\), percolation index \(p_i\), rigidity index \(r_i\), and stability maps \(r_{ci}\) for analyzing the unfolding trajectory of \(\alpha\)-lactalbumin. The indices \(p_i\) and \(r_i\) are introduced here for the first time; \(p_i\) reflects the stability of the giant percolating cluster along the unfolding trajectory, whereas \(r_i\) monitors the rigid-to-flexible transition of bonds in the whole network. As such, \(p_i\) reflects the hierarchical organization of the giant percolating cluster and, hence, is useful for identifying how rigidity starts propagating through the network. Furthermore, because \(p_i\) provides a microscopic view of the percolation behavior, the index allows tracing how structural changes (e.g., a mutation or the removal of a ligand or an ion) affect this (macroscopic) behavior. In contrast, \(r_i\) maps rigidity and flexibility at a more local scale, as demonstrated for the comparison of \(r_i\) with protection factors from H/D exchange experiments in the molten globule state of the protein (Fig. 4c). This comparison reveals that the structurally most stable regions are also those with the highest protection factors.\(^{[65]}\) Such regions have been interpreted as folding cores in this context.\(^{[70]}\) Thus, \(r_i\) may be useful for detecting residues that are part of folding cores in biomolecules. The complementarity of information provided by \(p_i\) and \(r_i\) is markedly demonstrated also when perturbing the constraint network as in the case of removing the calcium ion from \(\alpha\)-lactalbumin (Figs. 5a and 5b): \(p_i\) monitors the stabilizing effect of the calcium ion on the giant percolating cluster (which includes helix C in the presence of the ion), whereas \(r_i\) detects no influence of the ion on the (local) stability of the helix. Notably, in both cases long-range changes in flexibility and rigidity are detected despite perturbing the network only by removing short-range constraints between the ion and its surroundings. Therefore, both indices should be particularly useful for detecting changes in a biomolecule’s flexibility on ligand binding that may exert an entropic effect and for identifying adequate sites for mutations to increase the thermostability of a biomolecule. As an alternative to \(r_i\), the related largest rigid cluster propensity \(P_{\text{rc}}\) introduced by Rader et al.\(^{[37]}\) (Supporting Information, Table S1) has been used to identify the mechanically most stable residues in rubredoxin; these results agree well with results from H/D experiments (Table 2). The authors concluded that using this index is hence appropriate to analyze unfolding pathways in proteins. Finally, stability maps are applied to monitor correlations in structural stability for each pair of residues in a biomolecule; these maps have been successfully used to demonstrate that the distribution of functionally important flexible regions is conserved between meso- and thermophilic orthologs when considering the appropriate working temperatures of the enzymes (Table 2).\(^{[6]}\)

Indices to monitor the distribution of structural stability in biomolecules should be well behaved. In particular, one expects that, when constraints are removed from a network, the network’s stability either remains unchanged or decreases. This relation is reflected by both \(p_i\) and \(r_i\) on removal of the calcium ion, as was also to be expected from the definitions of these indices. The FIRST flexibility index \(g_i\) did not behave as expected; however (Fig. 3), on removal of constraints, \(g_i\) decreased in some regions of collective modes and overconstrained structural parts. An explanation is given in Figure 8. \(g_i\) relates the number of independently rotatable bonds to the number of potentially rotatable bonds in collective modes (number of redundant constraints to the maximal number of redundant constraints in overconstrained regions), with both the enumerator and denominator evaluated for that particular region. Breaking up this region into two parts by removing some constraints will in general lead to a disproportionate change in the enumerator and denominator values. As a
Figure 8. Example illustrating the unexpected behavior of the FIRST flexibility index $g$, on removal of constraints. a) Schematic view of the collapse of a rigid cluster into a smaller one and a dangling end. b) A worked example of a) with two network states before and after the removal of constraints. Each edge represents five bars that connect atoms in a body-bar network. Solid edges represent regular bonds. Dashed edges represent additional bonds for a fully connected network. Applying eq. (6) to both network states leads to $g_{\text{before}} = -0.024$ and $g_{\text{after}} = -0.028$. (Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.)

determined stability changes in the case of wildtype eglin C and 11 of its mutants (Pfleger and Gohlke, unpublished results).

In conclusion, in this study, we presented concise definitions for four global and four local indices for describing stability characteristics in biomolecules. Three index definitions were introduced and applied to analyze a biomolecule’s stability for the first time. Showcase analyses of the thermal unfolding of $\alpha$-lactalbumin demonstrated the scope and limitations, and the informative value of each index. This allowed us to provide guidelines for future studies suggesting which of these indices could best be used for analyzing, understanding, and quantifying structural features that are important for protein stability and function. Finally, we made suggestions for proper index notations in future studies to prevent the misinterpretation and to facilitate the comparison of results obtained from flexibility and rigidity analyses.

Acknowledgments

We are grateful to Doris L. Klein and Prakash C. Rathi (Heinrich-Heine-University, Düsseldorf) for fruitful discussions. The CNA program for calculating local and global flexibility indices is available from the authors upon request. CNA functions as a front- and back-end to the FIRST program; the latter is available from http://flexwe-b.asu.edu/software/.

Keywords: percolation theory · protein stability · phase transition · folding core · protein engineering · drug design


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