Published in final edited form as:
J Am Chem Soc. 2009 May 20; 131(19): 6785–6793. doi:10.1021/ja809133u.

Characterizing the Role of Ensemble Modulation in Mutation-induced Changes in Binding Affinity

Anthony Manson[1], **Steven T Whitten**[1,2], **Josephine C. Ferreon**[1], **Robert O Fox**[1], and **Vincent J Hilser**[1,*]

[1]Department of Biochemistry and Molecular Biology, and Sealy Center for Structural Biology and Biophysics, University of Texas Medical Branch, Galveston, TX 77555, USA

[2]RedStorm Scientific, Inc., Galveston, TX 77550, USA

Abstract

Protein conformational fluctuations are key contributors to biological function, mediating important processes such as enzyme catalysis, molecular recognition and allosteric signaling. To better understand the role of conformational fluctuations in substrate:ligand recognition, we analyzed, experimentally and computationally, the binding reaction between an SH3 domain and the recognition peptide of its partner protein. The fluctuations in this SH3 domain were enumerated by using an algorithm based on the hard sphere collision model, and the binding energetics resulting from these fluctuations were calculated using a structure-based energy function parameterized to solvent accessible surface areas. Surprisingly, this simple model reproduced the effects of mutations on the experimentally determined SH3 binding energetics, within the uncertainties of the measurements, indicating that conformational fluctuations in SH3, and in particular the RT loop region, are structurally diverse and are well-approximated by the randomly configured states. The mutated positions in SH3 were distant to the binding site, and involved Ala and Gly substitutions of solvent exposed positions in the RT loop. To characterize these fluctuations, we applied principal coordinate analysis to the computed ensembles, uncovering the principal modes of conformational variation. It is shown that the observed differences in binding affinity between each mutant, and thus the apparent coupling between the mutated sites, can be described in terms of the changes in these principal modes. These results indicate that dynamic loops in proteins can populate a broad conformational ensemble, and that a quantitative understanding of molecular recognition requires consideration of the entire distribution of states.

Protein conformational fluctuations are essential for biological function [1], and perturbations to fluctuations brought about by mutation or environmental changes are known to affect proteins functionally [2]. Despite clear experimental evidence for a dependence of function on fluctuations [3,4,5], the impact of conformational heterogeneity on important biological processes, such as substrate:ligand recognition, remains poorly understood [3,6]. Clearly, methods that can relate the functional and structural character of proteins are needed to understand mechanistically, and quantitatively, the role that fluctuations have in biological function.

To investigate the impact of structural fluctuations on molecular recognition, we selected the C-terminal src homology 3 (SH3) domain of the *C. Elegans* protein SEM5 (SEM5 C-SH3). SH3 domains are a conserved structural motif found in many regulatory proteins [7]. The typical function of these domains is to mediate protein:protein interactions in signaling pathways by

[*]Corresponding author: E-mail: vjhilser@utmb.edu.

recognizing proline-rich sequences that adopt a left-handed polyproline II helical structure [8,9,10,11]. Previous studies have demonstrated that conformational fluctuations in SH3 domains contribute significantly to the binding energetics observed for this important regulatory motif [6,10], particularly fluctuations in the RT loop [6], a region of the SH3 structure that is adjacent to its binding site. Specifically, it was shown that the energetics of binding cannot be reconciled solely in the context of those changes that are observed in the high resolution structures of the bound and unbound forms [6]. In that study, our group applied a series of Ala to Gly substitutions at surface exposed positions in the RT loop of SEM5 C-SH3 [12]. The residue positions in which the mutations were applied do not contact ligand in the high-resolution structure of bound SH3 [13], suggesting that these mutations can affect binding only through perturbations to the conformational fluctuations in the protein. Calorimetric titration experiments demonstrated a clear position and context-specific dependence in the effect of these Ala to Gly substitutions on the binding energetics of SEM5 C-SH3 [12].

In order to better understand the results of those titration experiments in terms of the conformational fluctuations around the average structure, we have performed a series of structure-based, all-atom simulations of the binding reaction between SEM5 C-SH3 and one of its recognition peptides within the Son of Sevenless (Sos) protein. Conformational heterogeneity in the SEM5 C-SH3 domain was simulated using an algorithm [14] based on the hard sphere collision (HSC) model [15]. For each SEM5 C-SH3 state generated by this algorithm, unbound and bound states of SH3 were generated by docking the Sos peptide in the same position and orientation as observed in the high-resolution structure [13]. Through the use of a solvent-accessible surface area-based energy function [16], the computed SH3 ensembles were energetically weighted. Energetic weighting of the ensembles in this manner allowed for a calculation of the difference in the free energy of binding ($\Delta\Delta G_{bind}$) between the SEM5 C-SH3 mutants, which reproduced the values measured experimentally with a correlation coefficient of 0.95. The agreement between the simulated and measured $\Delta\Delta G_{bind}$ values suggests that the changes in the conformational manifold for each mutant were accurately represented by the computed redistributions within the ensembles of randomly generated states. To aid in the structural interpretation of the fluctuations in the individual computed ensembles, a principal coordinate analysis (PCA) [17] of the conformational space was performed on each SEM5 C-SH3 variant. The results of the PCA indicate that the overall variability in the RT loop conformation can be well-approximated by three general motions: 1) a libration (i.e., up and down motion) relative to the core of the protein, 2) a radial motion (i.e., in and out) relative to the protein core, and 3) a twisting motion of the loop. The mutations applied to the RT loop appeared to affect these three motions with only subtle differences, however, these differences were sufficient to change the binding energetics in a predictable and measurable way.

Results and Discussion

The ability of proteins to undergo conformational transformations that can be modulated by environmental factors, or by interactions with other large or small molecules, is central to their biological function [1]. NMR techniques can be used to observe the very fast motions of protein structures [3,4]. Slow backbone relaxation events likely to be relevant to biological processes [5], however, are not sampled sufficiently by these NMR methods, or by most molecular dynamics simulations [18], and rapid, coarse-grain or meso-scale methods to investigate such phenomena could be of significant utility. To investigate the role of protein conformational fluctuations in substrate:ligand recognition, our group has conducted a mutational analysis of the binding energetics in SEM5 C-SH3 [12]. Structurally, SH3 domains are characterized by a β-sandwich core and a highly flexible RT loop (Fig. 1A), named for the arginine and threonine residues contained within it [13,19]. The binding site in these domains is situated between its β-sandwich core and the RT loop, and is specific for proline-rich sequences, such as that of the Sos peptide [8]. Mutations to SEM5 C-SH3 were applied at residue positions in the RT loop that were observed to be both fully solvent-exposed in high resolution structures [6,13], and to not play a direct role in contacting ligand in the bound complex [13]. As a result of these constraints, the effects of the mutations on the binding energetics will be to redistribute the ensemble of SEM5 C-SH3 conformations.

Simulations of structural fluctuations in an SH3 domain

Previous experimental studies from our lab, which measured NMR order parameters in SEM5 C-SH3, have demonstrated that the structural fluctuations within this domain vary significantly among the residue positions [6]. Data from these studies are reproduced in Figure 1B, and show that the most dynamic regions of the SEM5 C-SH3 structure under native conditions are found among residues ~ 162–172, as well as the N- and C-termini. Residues 162–172 correspond to the RT loop and contain the sites that were mutated in our earlier analysis of the SEM5 C-SH3 binding energetics [12]. The mutations represent a double mutant cycle, whereby Gly and Ala substitutions were made at residue positions 170 and 171 (i.e., G171A170, G171G170, A171A170, and A171G170). In our present study, we have simulated the fluctuations of residues 162–172 using an all-atom approach. Briefly, an algorithm based on hard-sphere collision (HSC) was used to generate conformational ensembles of the RT loop in the unbound state [14]. The algorithm employs a look up table for bond angles and bond lengths [20], and randomly samples backbone dihedral angles as well as a rotamer library [21] to define the backbone and side chain conformation, respectively. Those computed conformations that violated steric contact limits (22) between the atoms of the protein were discarded. To apply the HSC algorithm to the RT loop of SEM5 C-SH3, the positions of the backbone heavy atoms of the anchoring residues (i.e., 162 and 172) were constrained to the positions that were observed in the high-resolution structure [13], a constraint that is justified by the NMR relaxation data (See Fig. 1B). A conformational ensemble generated in this manner for the RT loop is shown in Figure 1C. As is evident in this figure, the ensemble of states generated for the RT loop represents a spatially diverse conformational manifold.

The number of conformations generated in construction of the RT loop ensemble was determined by convergence in the simulated values of $\Delta\Delta G_{bind}$ calculated for each SH3 variant, as well as convergence in the principal coordinates (both discussed below). The onset of convergence occurred with surprisingly few states generated by the algorithm, approximately 100. To ensure that the results of this study were not dependent on the number of states generated in the construction of the SH3 ensembles, 500 states were generated randomly for each of the SH3 variants. Random removal of 10% of the states in each ensemble showed a variation of less than 2% on both the calculated binding affinity differences between the SH3 variants and the principal coordinates (see online supporting material, Fig. S1).

The explicit contribution of electrostatics in modeling the structural fluctuations of the RT loop was not included in our simulations because they were determined by computation to not play a significant role. Briefly, the pKa values of each of the ten residues within 10Å of the binding interface were determined for a random sampling of states in the ensembles of each SH3 variant. The pKa values were calculated by numerical solution of the linearized Possion-Boltzmann equation with the method of finite-differences, as implemented in the H++ web-program [23,24]. These ten residues included ionizable groups within the flexible RT loop region (D164, E169, E172) as well as others located in the more rigid regions of the protein (K176, R177, D188, E193, R199, R200, Y207). The pKa values calculated for residues within the RT loop were observed to vary by no more than ± 0.4 pKa units among the states of each SH3 variant, whereas the pKa values calculated for the other groups were observed to vary by no more than ± 0.1 pKa units among the ensemble states (see online supplementary material, Fig. S2). These results suggest that in our simulations, electrostatics would not significantly bias the structural character of the RT loop ensembles. Consistent with these findings, direct determination of pKa values, obtained by fitting the pH-dependence of the NMR chemical shift data, revealed no significant differences between the mutants (Ferreon and Hilser, unpublished data).

To model the binding interactions between the Sos peptide and SEM5 C-SH3, the peptide ligand was docked to each SEM5 C-SH3 conformer in the same position and orientation as was observed in the high-resolution structure of the wild-type complex [13]. Surprisingly, docking the ligand in this manner resulted in steric collisions between atoms of the Sos peptide and the SEM5 C-SH3 conformers in only ~ 10% of the states generated for each ensemble. The computed ensembles showed a remarkable absence of conformations in which the RT loop overlapped with the volume of space occupied by the ligand in the structure of the bound complex [13]. This result is important because it suggests that while the active site may have a higher degree of conformational heterogeneity (i.e., there is a higher degree of folding frustration [25]), the additional conformational states do not dramatically reduce the number of states that can bind ligand. Instead, most SEM5 C-SH3 conformations tend to enlarge the binding pocket, producing a greater number of states with suboptimal substrate:ligand surface complementarity. The amount of surface area buried in the high-resolution structure of the wild-type versus the distribution of buried surface areas for the simulated ensembles of the SEM5 C-SH3 variants demonstrates this point (Fig. 2A). As can be seen, most ensemble states have less surface area buried upon docking the ligand, relative to the high-resolution complex. Among these mutants, the G171A170 variant buries the most surface area upon docking the ligand, followed by G171G170, A171A170, and A171G170 in order of decreasing amounts.

From the surface area calculations, a free energy (ΔG) of binding for each SEM5 C-SH3 variant was determined as follows. First, those states in each ensemble that had steric collisions with the docked Sos peptide were classified as binding incompetent. Thus, the bound ensembles consisted of all states generated for SEM5 C-SH3 in which docking of the Sos peptide was successful. A ΔG was determined for each of those states by use of a solvent accessible surface area-based energy function, which has been calibrated and tested extensively [16]. The ΔG of each state i of the bound ensemble was calculated with docked Sos peptide ($\Delta G_{bound,i}$). For the unbound ensemble, the ΔG of all ensemble states were calculated in the absence of the Sos peptide ($\Delta G_{unbound,i}$). An energy-weighted ΔG value was then calculated for the bound and unbound ensembles, which allowed for an approximate ΔG_{bind} value determined for each SEM5 C-SH3 variant,

$$\langle \Delta G_{bind} \rangle = \langle G_{bound} \rangle - \langle G_{unbound} \rangle, \quad (1)$$

, where $\langle \Delta G_{bound} \rangle = -RT \ln Q_{bound}$, $\langle G_{unbound} \rangle = -RT \ln Q_{unbound}$, and Q_{bound} and Q_{bound} are the sums of the statistical weights of the bound and unbound ensembles,

$Q_X = \sum_{i=1}^{N_x} \exp(-\Delta G_i / RT)$. In these expressions, R is the gas constant and T the absolute temperature. The summed probability of the binding incompetent states, relative to the bound states, was minimal (<0.1%) and, as such, excluding those states in the calculations of $\langle \Delta G_{bound} \rangle$ did not affect the simulated values of $\langle \Delta G_{bind} \rangle$. Structural refinement of the binding incompetent states by energy minimization techniques [26,27], to increase the number of binding competent conformers in an ensemble, similarly did not alter the relative binding energies significantly, although the absolute values were affected (see online supporting material, Fig. S3). This is an important point, of which we take advantage when comparing the computed mutational effects with experiment (shown below). Because of the lack of sensitivity of the differences in binding free energy between each mutant, relaxation methods were not employed. The binding simulation was repeated for all SEM5 C-SH3 variants. A calculation of $\Delta\Delta G_{bind}$ between each mutant was then determined as,

$$\Delta\Delta G_{bind} = \langle \Delta G_{bind,mut1} \rangle - \langle \Delta G_{bind,mut2} \rangle . \tag{2}$$

.

Comparison of the simulated $\Delta\Delta G_{bind}$ values to those observed experimentally [12] is shown in Figure 2B. The correlation between the experimental and simulated binding energetics is striking, producing a slope of 1.09 and a correlation coefficient of 0.95. Surprisingly, the simulated values of $\Delta\Delta G_{bind}$ were within the error range of the values measured experimentally. We note that the error bars in the predicted binding affinity represents the standard deviation of the values calculated by randomly excluding 10% of states from the calculation for each mutant. The ability of the HSC-based simulations to accurately reproduce the effect of mutation on the binding energetics for the SH3:Sos interaction suggests that the conformational fluctuations in the RT loop is well-approximated by the ensemble of randomly configured, coil-like states.

Detailed comparison of the $\Delta\Delta G_{bind}$ values between the SEM5 C-SH3 variants revealed additional trends. For instance, the impact of the Ala to Gly substitution at position 170 is greater in the context of Ala at position 171 than the Gly at that position. Conversely, the impact of the Gly to Ala substitution at position 171 was greater in the context of Gly at position 170 than the Ala. The ability to accurately simulate these trends of the experimental $\Delta\Delta G_{bind}$ data suggests that the effects of mutation in this region of the RT loop are manifested both in changes in the allowable conformational space as well as changes in the solvation energetics, a point that is discussed below.

Structural character of the RT loop ensemble

To structurally characterize the conformational ensembles of each SEM5 C-SH3 variant, we applied a principal coordinate analysis (PCA) [17] to the cartesian positions of the α and β carbons in residues 162–172 of the RT loop, a procedure similar to that employed previously by Sims and colleagues in the analysis of the distribution of dihedral angles in polypeptide chains [28]. Three key observations were made when applying this PCA technique, demonstrating its suitability for our structural analysis. First, the first three principal modes (i.e., "coordinates") expressed >90% of the total variation among the dataset. This condition allowed for an acute reduction in the dimensionality of the conformational space of each SH3 variant. Second, the conformer density along each principal axis was found to be predominantly Gaussian, allowing the unit distance along the corresponding principal axis to be interpreted as the standard deviation. Third, for this set of SEM5 C-SH3 mutants, the principal modes of variation in the Cα and Cβ positions for residues 162–172 of the loop represented concerted displacement in the backbone chain (discussed below) that were similar in character among the SEM5 C-SH3 variants, which allowed for direct comparison of the conformational spaces among the mutants as represented in principal space. This similar structural character in the mutant ensembles likely resulted from each ensemble describing structural variations in the same flexible region subject to similar sequence and contextual constraints. To highlight the utility of the principal space for interpreting the effects of the mutations, the envelopes of the allowed conformational space for mutants G171A170 and A171G170 are plotted against the first three principal coordinates in Figure 3A. Comparison of the conformational envelopes for these two mutants reveals discernable differences in their relative size, position and shape. The G171A170 and A171G170 variants have the lowest and highest affinity, respectively, for the Sos peptide (see Fig. 2B).

Each principal axis can be viewed as a concerted variation in the conformation of the RT loop. To illustrate this, the individual ensemble states at the extremes of each principal coordinate are shown in Figures 3B–D. The structures at the extremes of the principal coordinates reflect the breadth of the conformational envelope, and varying the position along each axis can be interpreted as sampling the structural intermediates of those extremes. For instance, sampling structures along the axis of the first principal coordinate (pc1) reveals displacements in the backbone that are orthogonal to the path of the native chain, and which represents an up and down libration motion in the RT loop. Variations in structure that tracked with changes in the second principal coordinate (pc2) exhibited an “in and out” shift of the backbone in the RT loop, toward and away from the centroid of the protein. Structural changes in the ensemble that followed changes in the third principal coordinate (pc3) represent a twisting displacement in the backbone. Importantly, the structural motif of each of the three principal coordinates was found to be similar in character among the SEM5 C-SH3 variants: only the breadth and positioning of the conformation envelope along each principal axis varied, albeit subtly, among the mutants.

To understand the structural effects of Ala and Gly substitutions, we note that the absence of a β-carbon in Gly, relative to Ala, dramatically increases the accessible φ and Ψ space for a residue in a protein [29]. This mutation strategy as applied to positions 170 and 171 in SEM5 C-SH3 is expected to affect the binding energetics in distinct position-specific ways. First, according to the high-resolution X-ray and NMR structures of the apo and holo protein [6,13], the φ and Ψ angles for position 170 are in a region that is accessible to both Ala and Gly (i.e., $\varphi = -140$, $\Psi = 150$) (Fig. 4A). Thus, substitution of Ala for Gly at 170 expands the conformational space of the RT loop, whereas the reverse substitution (Gly for Ala) constricts the ensemble. For position 171, the φ and Ψ angles are in a region that is accessible only to Gly. Consequently, substitution for Ala at position 171 would introduce conformational strain in the ensemble and should shift the allowable space. The effects of introducing conformational strain, and/or increasing the allowable conformation space were clearly observed in the principal space. As shown in Figure 4B, the Ala to Gly substitution at position 170 increased the allowable conformational space of the SEM5 C-SH3 ensemble dramatically. Interestingly, the increase in conformational space was not evenly distributed among all three principal coordinates, suggesting that this substitution increased the up and down displacement of the RT loop (i.e., pc1) much more so than the other concerted changes. With regards to the Gly to Ala substitution at position 171, which is shown in Figure 4C, the conformational space was shifted along the pc1 and pc2 axes with minimal changes in pc3 (i.e., the twisting displacement).

Energetic minima in the RT loop ensemble

Additional information about the structural character of the SEM5 C-SH3 ensemble can be ascertained from the energetic hierarchy of states in each ensemble. The weighted conformers of each ensemble, obtained by applying the same surface area-based energy function discussed above, formed low energy (i.e., high probability) clusters within the principal coordinate space of each ensemble. The clusters correspond to groups of states that occupy similar regions of principal space, and as such, the states within a cluster have similar structural character, shown schematically in Figure 5 (Note: the simulated data that produced Fig. 5 is provided in the online supporting material, Fig. S4). In this figure, conformational clusters having a significant statistical weight (an aggregate probability of the states of the cluster >10%) were plotted against pc1 and pc2. These two principal coordinates retained ~ 70% of the total variation in the original dataset. As the data in the figure show, the distribution of high occupancy clusters varied from mutant to mutant. The least constrained mutant, G171G170, exhibited a greater number of clusters, which was expected since the lack of Cβ's at positions 170 and 171 would allow the RT loop to sample a larger conformational space relative to an Ala residue at one or both of these two positions. Accordingly, the G171G170 mutant can be considered as having the set of “basis” clusters for the mutant cycle. The G171A170 mutant showed two fewer clusters, relative to G171G170, reflecting a significant contraction of the conformational space, which was visible also in Figure 4B. The A171G170 mutant provided access to one more cluster than G171A170, but did not have access to two others present in G171A170, reflecting a shift in the conformational space in response to the transposition in the location of the Gly residue. The most constrained mutant, A171A170, showed exclusion of all but one cluster and exhibited the smallest allowed space. In total, the effect of energetically weighting the ensembles was to differentially induce basins in the free energy landscapes of each mutant. The phenomena of conformational strain and degeneracy as introduced by the individual substitutions, however, were clearly apparent in these data.

The introduction of ligand to the ensembles decreased the averaged free energy level for each of the SEM5 C-SH3 variants, relative to their unbound ensembles, favoring complexed states; a result that was not unexpected given the known burial of apolar surface involved in the SEM5 C-SH3:Sos interaction [13]. Interestingly, when viewing entire ensembles in principal space, both the number and the positions of the high occupancy clusters for the unbound and bound ensembles were different, suggesting a change in the overall character of the SEM5 C-SH3 ensemble upon binding ligand. This is shown in Figure 6A for the G171A170 and A171A170 mutants. In this figure, the energetic landscapes for both the unbound and bound ensembles were plotted against pc1 and pc2. As can be seen, the allowable conformational space for A171A170 was a subset of the conformational space for G171A170. Also, the regions of principal space common to both of these mutants had similar, although not identical, terrain features. This observation suggests that the major differences in the binding energetics between these two mutants were caused by favorable regions in the conformational space of the bound G171A170 ensemble that were not allowed in A171A170. These regions of G171A170 conformational space expressed prominent low energy clusters that provided for, in aggregate, more favorable interactions between SEM5 C-SH3 and Sos, relative to that of A171A170 (i.e., G171A170 had a higher affinity for Sos, see Fig. 2B). In addition, the restriction of conformational space due to the presence of Ala at positions 170 and 171 had a significant entropic cost to the binding reaction, which is shown in Figure 6B. The conformational entropy, ΔS_{conf}, was calculated from the Boltzmann relation,

$$\Delta S_{conf} = -R \cdot \sum_i P_i \cdot \ln P_i. \tag{3}$$

.

The effect of Ala at position 171 on the allowed conformational space of SEM5 C-SH3 appeared to be amplified by the presence of Ala at position 170. The entropic cost of the Gly to Ala substitution at position 171 when Gly occupied position 170 was calculated to be ~ 200 cal/mol ($-T\Delta S_{conf}$ at 25 °C), whereas it was ~ 400 cal/mol when Ala also occupied position 170. Similarly, the presence of Ala at position 171 amplified the entropic cost to binding for the Gly to Ala substitution at position 170 (300 cal/mol versus 500 cal/mol; which can be calculated directly from Fig. 6B (not shown)). Thus, the conformational strain introduced by adjacent Cβ's at positions 170 and 171 of the RT loop had a significant cooperative affect on the binding energetics of SH3 (by ~ 200 cal/mol), and was reproduced accurately by these HSC-based simulations (see Fig. 2B). Interestingly, this result demonstrates that the phenomenon of conformational strain can be reconciled in the context of redistributing the conformational ensemble, as opposed to distortions in bond lengths or bond angles, which are not accounted for in the current analysis.

To characterize the structural features of those states in the bound ensemble of G171A170 that were not represented in the A171A170 bound ensemble, the most probable states (states having a probability > 20%) from both were selected and visualized. In the low-energy states of the A171A170 ensemble, the presence of the side chain methyl group of Ala at position 171 caused a concerted displacement in the backbone throughout the RT loop region. This is shown in Figure 7A, where instead of extending the backbone of the RT loop outward from the center of the protein, as was observed in the most probable states of the G171A170 ensemble, the RT loop was directed tangentially away from the binding site. As such, this low-energy state of the A171A170 ensemble possessed a binding site that was more open relative to that observed in many of the low-energy states of the other mutants (Fig. 7B). This would also prevent optimal surface and residuepair complementarity with the ligand in the double Ala variant (see Fig. 2A).

Conclusions

The results presented here indicate that conformational fluctuations in the RT loop of the C-terminal SH3 domain of SEM5 are well-represented by a broad conformational repertoire of unfolded-like states with regard to those residue positions spanning 162–172, residues known from NMR relaxation experiments to be conformationally heterogeneous both in the apo and holo forms. By applying Ala and Gly mutations to two solvent-exposed positions in the loop, and analyzing the mutation-induced changes in the ensemble using PCA, we were able to decipher the nature of the mutational effects. Our results indicate that the change from Ala to Gly at position 170 had the general effect of increasing the size of the ensemble of low affinity conformations, thus decreasing the binding affinity. On the other hand, because the wild-type Gly at position 171 is in a region of φ/Ψ space that is disallowed for Ala, mutation from Gly to Ala shifts the ensemble away from high affinity states, decreasing the affinity. The binding affinity of each of the mutants is a nonlinear combination of these two effects. Importantly, our analysis was able to reproduce the experimental trends in the data, quantitatively capturing the cooperativity (i.e., nonadditivity) between sites.

Further, by performing a principal coordinate analysis on the backbone positions of the RT loop in each ensemble, structural characterization of the fluctuations showed three primary modes of structural variation in the ensemble, with the nature of each mode varying only subtly between the mutants. These structural variations in the RT loop involved an up and down libration, an in and out displacement relative to the protein core, and a twisting of the loop. In short, the analysis of the RT loop fluctuations using principal space provide a unique and detailed picture of how structural perturbations remote from binding sites may assert their influence mechanistically. Indeed, previous studies have demonstrated that high-resolution structures are often insufficient to account for the observed binding energetics in many protein:ligand systems, and in particular, are challenged to account for mutational effects [6, 30]. The results presented here demonstrate that conformational fluctuations in the protein, or in the ligand as shown elsewhere [14], can contribute significantly to the binding energetics, and that an adequate representation of the conformational manifold is a prerequisite to a mechanistic understanding of molecular recognition.

Finally, we note that it seems somewhat paradoxical that a coarse-grain (or meso-scale) energy function, which is based on solvent-accessible surface area, should prove sufficient for the current studies. The reason for the success is rooted in the details of the energy function as well as the specific nature of the problem being addressed. Because the energy function was empirically determined from experimental data and averaged over all types of polar and apolar surface area, it neglects specific details such as hydrogen bonds. Although such an approach would be inappropriate if one or a few specific structural states were the determinants of the observed behavior, the large number and diversity of states that contribute to the SEM5 C-

SH3:Sos binding (as observed directly in the NMR relaxation data for both the bound and unbound state – See Fig 2B) apparently provides a reasonable average energy for the different ensembles. Indeed, the robustness of the results to jackknife analysis suggests that the contours in the energy landscape (Fig. 5 & Fig. 6), which are derived from the combination of sampling technique and energy calculations, are sufficient to capture the observed mutational effects. The success of this approach suggests that in cases where large-scale backbone relaxation of a region of a protein may play an important role in function, such as with intrinsically disordered proteins, a simplified (albeit all-atom) strategy that focuses on enumerating conformationally diverse states represents a useful alternative to molecular dynamics approaches that may not yet be capable of adequately sampling the entire manifold of relevant states.

Methods

Modeling structural fluctuations in the RT loop

An algorithm based on the HSC model [14] was used to generate conformational ensembles of residues 162–172 in the RT loop, and is described in detail elsewhere [14]. The positions of the backbone heavy atoms of residues 162 and 172 were constrained to the positions that were observed in the high-resolution structure [13]. The HSC model, based upon van der Waals atomic radii [22], was used as the only scoring function to eliminate grossly improbable conformations. The Gibbs free energy of each state i, $\Delta G_{unbound,i}$, was determined from an energy function based upon solvent accessible surface areas, as calculated using the method of Lee and Richards [31], and has been calibrated previously and tested extensively [16]. The fluctuations in the RT loop were also modeled using the X-PLOR [26] and Gromacs [32] simulation software packages, using a slow cooling protocol, for comparison to the ensembles obtained using the HSC model, and demonstrated that the overall conclusions presented in this communication were independent of the software used to model the RT loop ensemble.

Calculation of pKa values for ionizable groups within and near the binding interface

The pKa values of ten residues within 10 Å of the binding interface were calculated using the H++ web-program (http://biophysics.cs.vt.edu/H++; refs [23,24]), as applied to the SEM5 C-SH3 structure [13]. The pKa calculations were repeated on a random sampling of 10% of the ensemble states for each SH3 variant, parameterized to a solvent ionic strength of 150 mM, an internal protein dielectric of 6, and a temperature of 25°C. Variations in the calculated pKa values for the ten residues among the random sampling of states is provided in the online supporting material.

Simulating the SH3:Sos binding reaction

The binding interaction between SH3 and Sos was simulated by docking the Sos peptide to each state in the RT loop ensemble, using the Sos peptide as represented in the high-resolution structure of the native complex [13] as the template for the peptide's position and orientation with respect to the SH3 molecule. The composition of the Sos peptide was also modeled as identical to its representation in the high-resolution complex. Upon docking the ligand, those states with steric collisions between SH3 and Sos (complexes that violated radii contact limits [22]) were considered binding incompetent in our simulations. The Gibbs free energy of each bound state in the ensemble ($\Delta G_{bound,i}$) was determined using the same structure-based energy function used for the unbound ensemble, with the exception of bound Sos being present in each state. The difference in free energies between the bound and unbound ensembles was calculated as given by Eq. 1 to yield a simulated binding free energy. The cratic entropy [33], which is the entropic cost related to the reduction in the number of system components, was not included in these calculations as this term would cancel when determining the $\Delta\Delta G_{bind}$ values as given by Eq. 2.

Structural refinement of binding incompetent states

The structures of the binding incompetent states of an ensemble were refined by the X-PLOR software package [26], using Powell's conjugate gradient minimization technique [27].

Principal Coordinate Analysis of the RT loop ensembles

PCA applies a coordinate rotation on a data set such that the transformed axes become aligned with the directions of maximum variance, which often allows for the reduction of a multidimensional data set to a smaller set of characteristic dimensions [17,34] . In our simulations, PCA allowed the majority of the structural variations in the RT loop ensembles to be reduced to a manageable number of coordinates that served to order the random set of conformations into a meaningful landscape. A feature vector representing backbone structural variations in the RT loop ensemble was defined as,

$$\boldsymbol{c}_i=[\boldsymbol{r}_1,\cdots,\boldsymbol{r}_k]\{j{:}1,\boldsymbol{m}\}, \tag{4}$$

, where c_i is a vector of each state i in an ensemble, r_k are the components of the vectors describing the Cα and Cβ atomic positions, and $j{:}1,m$ is the index spanning the contiguous residues (162–172) of the RT loop. This metric has the virtue of capturing the position of each residue and its orientation along the backbone. PCA was then applied to a metric distance function between the different feature vectors, which was defined as

$$d_{i,j}=\left(\sum_k (r_k^i-r_k^j)^2\right)^{1/2}, \tag{5}$$

, where the distance between conformers i and j, $d_{i,j}$, was determined over all components k of the position vectors. Full details for applying PCA to the RT loop ensembles via the metric distance function given by Eq. 5 are provided in the online supplementary material.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by NIH grants to VJH (GM63747) and ROF (AI056326) and Welch Foundation grants to VJH (H-1461) and ROF (H-1345).

References

1. Henzler-Wildman K, Kern D. Nature 2007;450:964–972. [PubMed: 18075575]
2. (a) Shoichet BK, Baase WA, Kuroki R, Matthews BW. Proc. Nat. Acad. Sci. USA 1995;92:452–456. [PubMed: 7831309] (b) Gozavi S, Whitford PC, Jennings PA, Onuchic JN. Proc. Nat. Acad. Sci. USA 2008;105:10384–10389. [PubMed: 18650393] (c) Bai Y, Sosnick TR, Mayne L, Englander SW. Science 1995;269:192–197. [PubMed: 7618079] (d) Gillespie B, Dahlquist FW, Marqusee S. Nature 1999;6:1072–1078.
3. Frederick KK, Marlow MS, Valentine KG, Wand AJ. Nature 2007;448:325–329. [PubMed: 17637663]
4. (a) Lee AL, Kinnear SA, Wand AJ. Nature Struct. Biol. Science J. Mol. Biol 2000;7:72–77. (b) Volkman BF, Lipson D, Wemmer DE, Kern D. Science 2001;291:2429–2433. [PubMed: 11264542] (c) Yang D, Kay LE. J. Mol. Biol 1996;263:369–382. [PubMed: 8913313]

5. (a) Lu HP, Xun L, Xie XS. Science 1998;282:1877–1882. [PubMed: 9836635] (b) Yang H, Luo G, Karnchanaphanurach P, Louie TM, Rech I, Cova S, Xun L, Xie XS. Science 2003;302:262–266. [PubMed: 14551431]

6. Ferreon JC, Volk DE, Luxon BA, Gorenstein DG, Hilser VJ. Biochemistry 2003;42:5582–5591. [PubMed: 12741814]

7. (a) Koch CA, Anderson D, Moran MF, Ellis C, Pawson T. Science 1991;252:668–674. [PubMed: 1708916] (b) Rozakis-Adcock M, Fernley R, Wade J, Pawson T, Bowtell D. Nature 1993;363:83–85. [PubMed: 8479540]

8. Egan SE, Giddings BW, Brooks MW, Buday L, Sizeland AM, Weinberg RA. Nature 1993;363:45–51. [PubMed: 8479536]

9. Li SSC. Biochemical J 2005;390:641–653.

10. Yu HT, Chen JK, Feng SB, Dalgarno DC, Brauer AW, Schreiber SL. Cell 1994;76:933–945. [PubMed: 7510218]

11. (a) Mayer BJ. Journal of Cell Science 2001;114:1253–1263. [PubMed: 11256992] (b) Mayer BJ, Gupta R. Protein Modules in Signal Transduction 1998;228:1–22.

12. (a) Ferreon JC, Hilser VJ. Protein Science 2003;12:447–457. [PubMed: 12592015] (b) Ferreon JC, Hilser VJ. Biochemistry 2004;43:7787–7797. [PubMed: 15196021] (c) Hamburger JB, Ferreon JC, Whitten ST, Hilser VJ. Biochemistry 2004;43:9790–9799. [PubMed: 15274633]

13. Lim WA, Richards FM, Fox RO. Nature 1994;372:375–379. [PubMed: 7802869]

14. Whitten ST, Yang HW, Fox RO, Hilser VJ. Protein Science 2008;17:1200–1211. [PubMed: 18577755]

15. Richards FM. Annu. Rev. Biophys. Bioeng 1977;6:151–176. [PubMed: 326146]

16. (a) Murphy KP, Freire E. Adv. Protein Chem 1992;43:313–361. [PubMed: 1442323] (b) Murphy KP, Bhakuni V, Xie D, Freire E. J. Mol. Biol 1992;227:293–306. [PubMed: 1522594] (c) D'Aquino JA, Gómez J, Hilser VJ, Lee KH, Amzel LM, Freire E. Proteins: Struct. Funct.Genet 1996;25:143–156. [PubMed: 8811731] (d) Gómez J, Hilser VJ, Xie D, Freire E. 1995;22:404–412. (e) Xie D, Freire E. J. Mol. Biol 1994;242:62–80. [PubMed: 8078072] (f) Baldwin RL. Proc. Natl. Acad. Sci. USA 1986;83:8069–8072. [PubMed: 3464944] (g) Lee KH, Xie D, Freire E, Amzel LM. Proteins: Struct. Funct. Genet 1994;20:68–84. [PubMed: 7824524] (h) Hebermann SM, Murphy KP. Protein Sci 1996;5:1229–1239. [PubMed: 8819156] (i) Luque I, Mayorga OL, Freire E. Biochemistry 1996;35:13681–13688. [PubMed: 8885848]

17. Pearson K. Philosophical Magazine 1901;2:559–572.

18. (a) Brooks BR, Karplus M. Proc. Nat. Acad. Sci. USA 1983;80:6571–6575. [PubMed: 6579545] (b) Lazaridis T, Karplus M. Proteins 1999;35:133–152. [PubMed: 10223287] (c) Rod TH, Radkiewicz JL, Brooks CL III. Proc. Nat. Acad. Sci. USA 2003;100:6980–6985. [PubMed: 12756296] McCammon, JA.; Harvey, S. Cambridge University Press. Cambridge: 1987. Brooks, CL., III; Karplus, M.; Pettitt, BM. Advances in Chemical Physics. New York: Wiley; 1988. (f) Benkovic SJ, Hammes-Schiffer S. Science 2003;301:1196–1202. [PubMed: 12947189] (g) Grunberg R, Nilges M, Leckner J. Structure 2006;14:683–693. [PubMed: 16615910]

19. Weng ZG, Rickles RJ, Feng SB, Richard S, Shaw AS, Schreiber SL, Brugge JS. Mol. Cell. Bio 1995;15:5627–5634. [PubMed: 7565714]

20. Momany FA, McGuire RF, Burgess AW, Scheraga HA. J. Phys. Chem 1975;79:2361–2381.

21. Lovell SC, Word JM, Richardson JS, Richardson DC. Proteins: Struct. Funct. Genet 2000;40:389–408. [PubMed: 10861930]

22. Bondi A. J. Phys. Chem 1964;68:441–451.

23. Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, Onufriev A. Nucleic Acids Res 2005;33:W368–W371. [PubMed: 15980491]

24. Anandakrishnan R, Onufriev A. Journal of Computational Biology 2008;15:165–184. [PubMed: 18312148]

25. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Proteins 1995;21:167–195. [PubMed: 7784423]

26. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Acta Crystallographica Section D-Biological Crystallography 1998;54:905–921.

27. Powell MJD. Mathematical Programming 1977;12:241–254.

28. Sims GE, Choi IG, Kim SH. Proc. Nat. Acad. Sci. USA 2005;102:618–621. [PubMed: 15640351]

29. (a) Ramachandran GN, Ramakrishnan C, Sasisekharan V. J. Mol. Bio. Adv. Protein Chem 1963;7:95–99. (b) Ramachandran GN, Sasisekharan V. Adv. Protein Chem 1968;23:283–438. [PubMed: 4882249]

30. Bauer F, Sticht H. FEBS Lett 2007;581:1555–1560. [PubMed: 17382937]

31. Lee B, Richards FM. J. Mol. Bio 1971;55:379–400. [PubMed: 5551392]

32. van Buuren A. Gromacs Documentation. http://www.gromacs.org/.

33. Murphy KP, Xie D, Thompson KS, Amzel LM, Freire E. Proteins: Struct. Funct. Genet 1994;18:63–67. [PubMed: 8146122]

34. (a) Becker OM. Proteins: Struct. Funct. Genet 1997;27:213–226. [PubMed: 9061786] (b) Becker OM. J. Comp. Chem 1998;19:1255–1267. (c) Gower JC. Biometrika 1968;55:582–585.

35. Fukunaga; Keinosuke. Introduction to Statistical Pattern Recognition. Elsevier; p. 1990

36. Cullen, CG. Matrices and Linear Transformations. Addison-Wesley, Reading Mass; 1972.

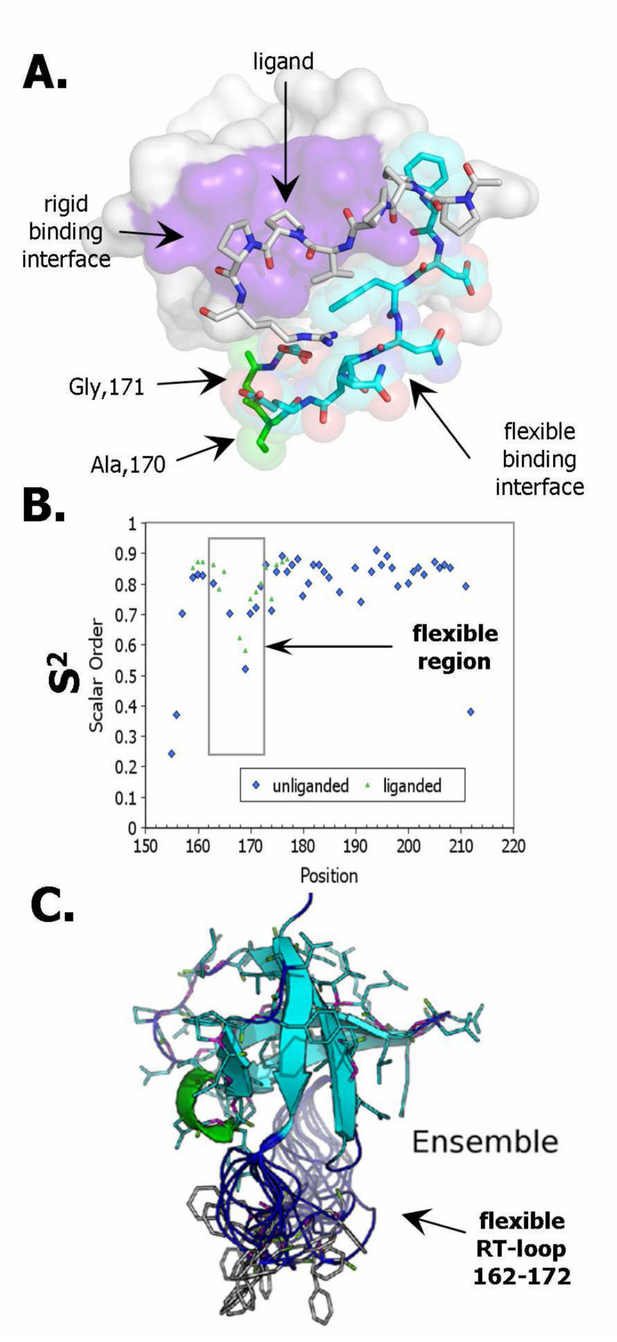

Figure 1.

Conformational fluctuations in the RT loop region of SEM5 C-SH3. **A)** View of the Sos binding site in SH3 [13]. The Sos peptide is shown by a stick cartoon representation and SH3 is shown by space-fill spheres. The regions near the binding site considered to be rigid were colored blue, whereas the flexible regions were colored purple. Those regions of SH3 that do not contact the Sos ligand were colored white. Mutated positions 170 and 171 are shown at the tip of the RT loop (green). **B)** NMR order parameters versus position for the liganded and unliganded SH3 [13]. Positions 162–172 show low order parameters indicating conformational heterogeneity. **C)** Overlay of 20 randomly selected conformations, from among 500 successful conformations generated by an HSC model, showing the conformational diversity of the RT loop.

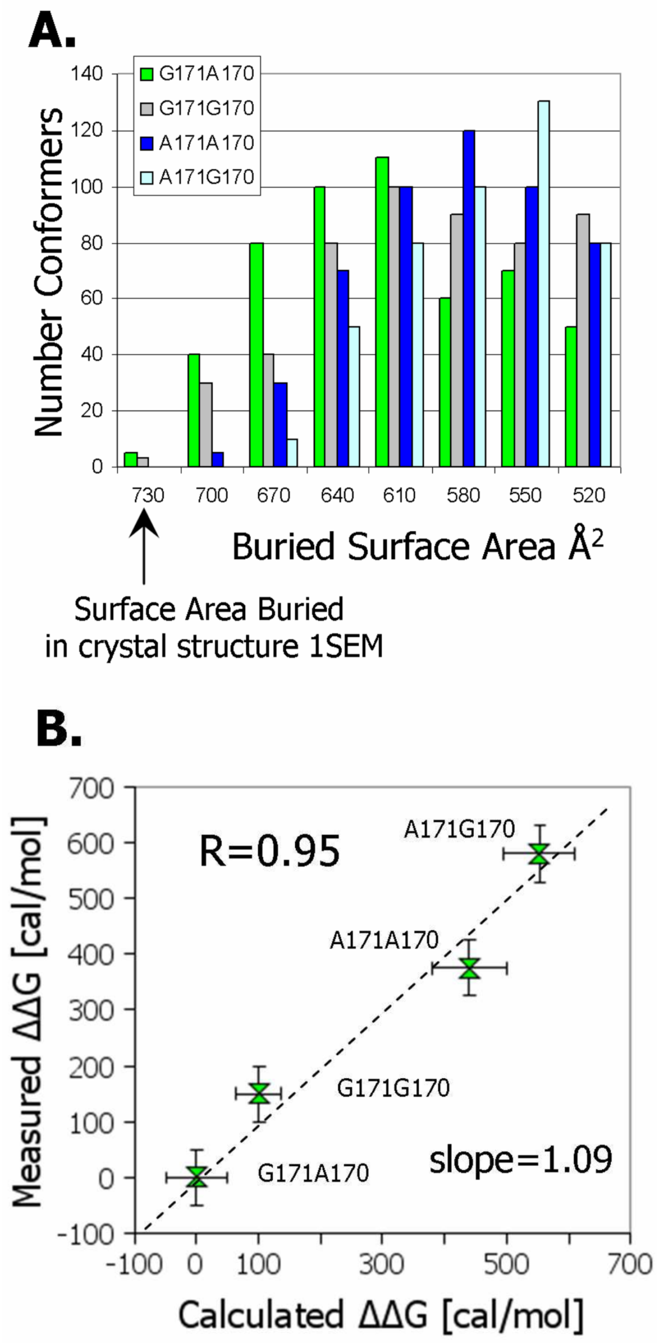

Figure 2.

The effects of substitutions in the RT loop on SEM5 C-SH3 binding energetics. **A)** The buried surface area due to docking the ligand in each SH3 ensemble was dependent on the Gly to Ala substitution. Shown is a histogram in which the states were binned according to the amount of buried surface area upon docking the Sos peptide, with each column colored according to the SH3 variant, as indicated in the figure. **B)** Correlation of calculated difference binding free energies (abscissa) to measured (ordinate) [12]. The error bars for the experimental values represent the range of experimental error in the measured energies, as determined from multiple measurements of the binding energies by ITC. The error bars for the calculated values represent the variation resulting from random removal of 10% of the ensemble states. The correlation between the calculated and measured values was 0.95.

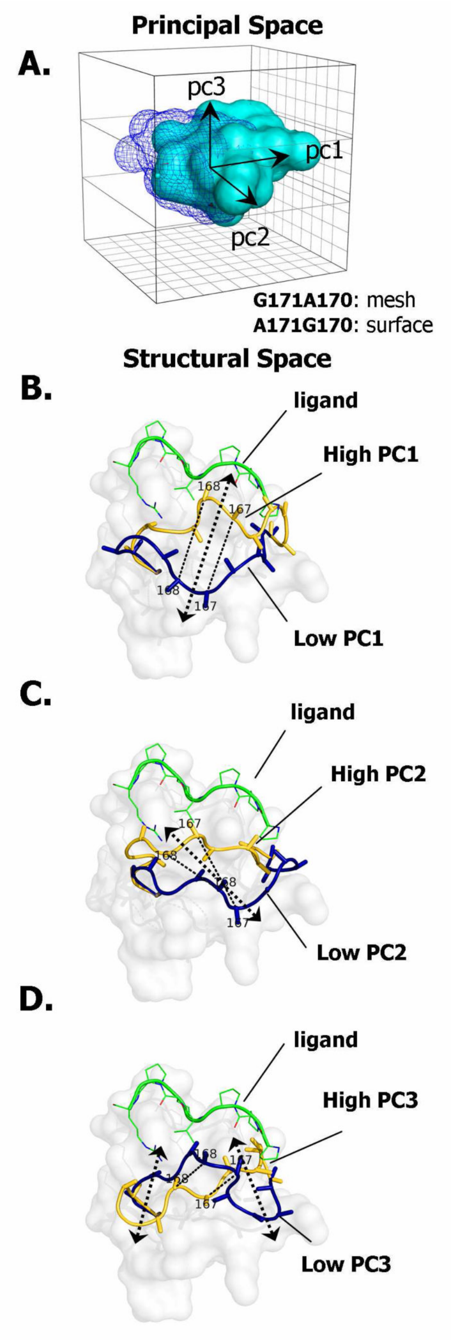

Figure 3.

Characterizing the fluctuations in the RT loop by PCA. **A)** Comparison of conformational envelopes in principal space (showing the first three principal modes) of mutants with the highest (mesh) and lowest (surface) binding affinity. Shown are surface maps representing the regions of principal space where the density of states was greater than at least 10% of the maximum observed density. Variances and corresponding structural extrema along principal axes of the first three principal modes (pc1–pc3) for the G171A170 variant is shown in panels **B–D**.

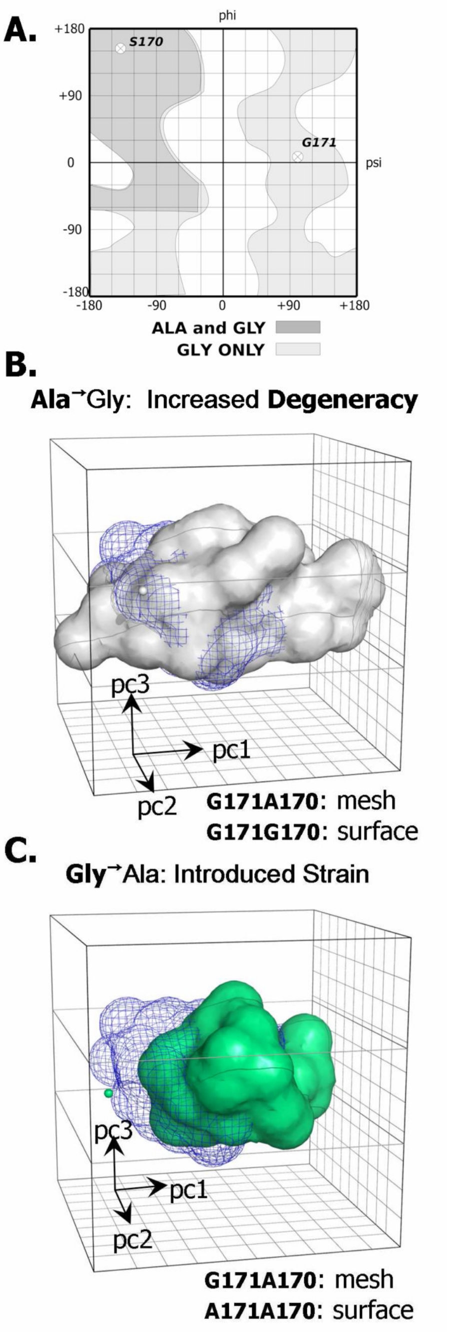

Figure 4.

The effects of substitutions in the RT loop on conformational space. **A)** Comparative plots of the allowed regions for Ala and Gly illustrating the difference in allowed space for a single amino-acid position in a polypeptide chain. The φ/ψ coordinates of positions 170 and 171 found within the crystal structure for the wild-type SH3 are labeled. Note that position 171 falls within the positive phi region for the complexed crystal structure. **B)** Conformer density plots for the allowed conformational space for the RT loop in principal space is given for the Ala to Gly substitution at position 170, and demonstrates a clear increase in degrees of freedom for the RT loop associated with this sequence change. **C)** For the Gly to Ala substitution at position

171, Ala in position 171 shifts the allowable region in principal space, reflecting the effect of conformational strain on the ensemble.

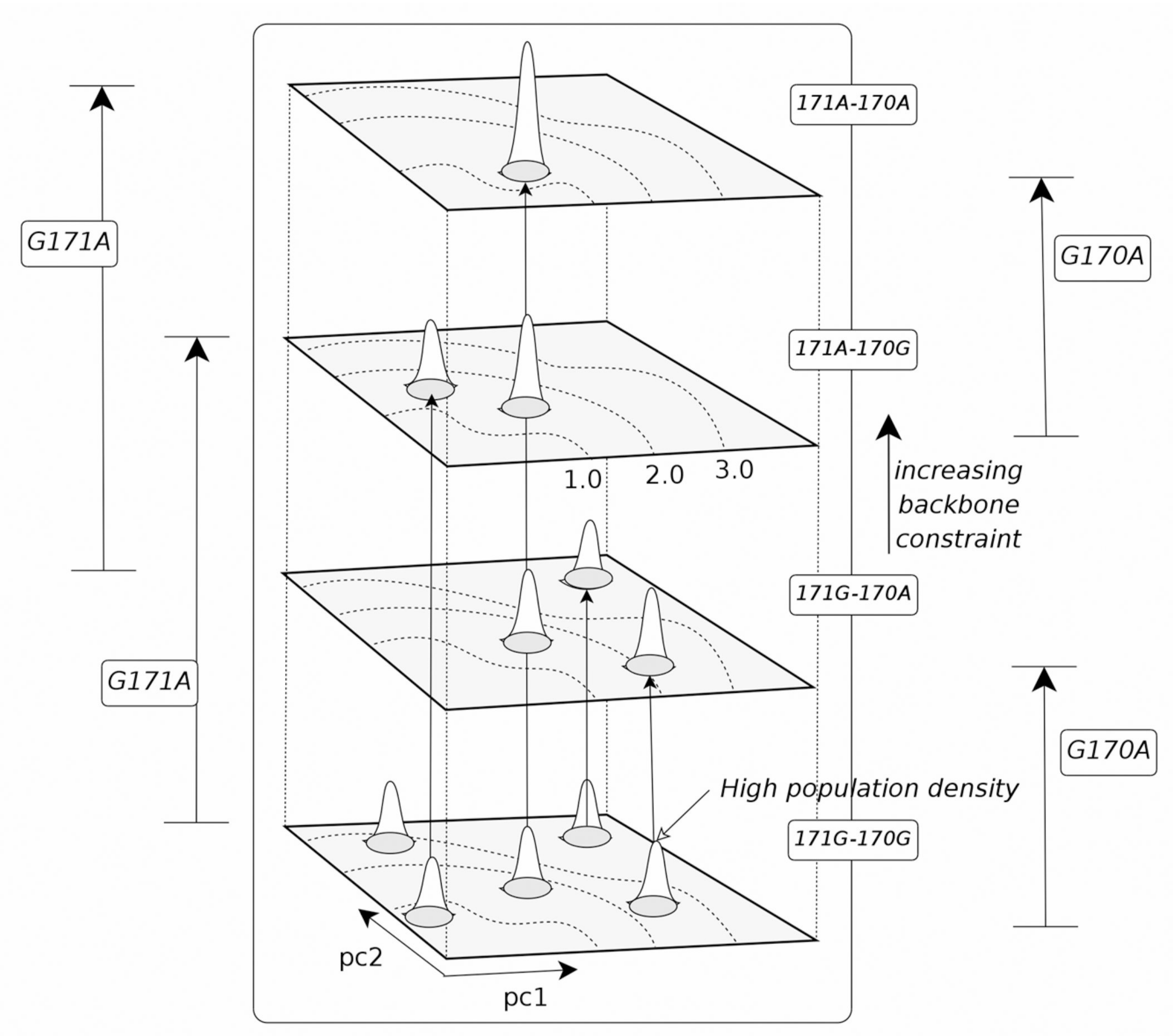

Figure 5.

Comparison of energy landscapes among the SH3 variants. The fractional occupancy (Boltzmann probability) plotted against the first two principal coordinates is shown for each mutant, presented in increasing order of conformational constraint (bottom-top). Energetically favorable clusters are seen to appear and disappear as the allowable region is changed through mutation. Contours in each two-dimensional plane represent increasing structural (*rmsd*) distance from the crystal structure complex located at origin. The *rmsd* between two structures (a,b) is given by $rmsd_{a,b}=\left(\frac{1}{N}\sum_{i}^{N}(r_{a,i}-r_{b,i})^2\right)^{1/2}$, where N is the number of atoms in each protein, $r_{a,i}$ is the cartesian position of atom i within structure a and $r_{b,i}$ is the position of atom i within structure b.

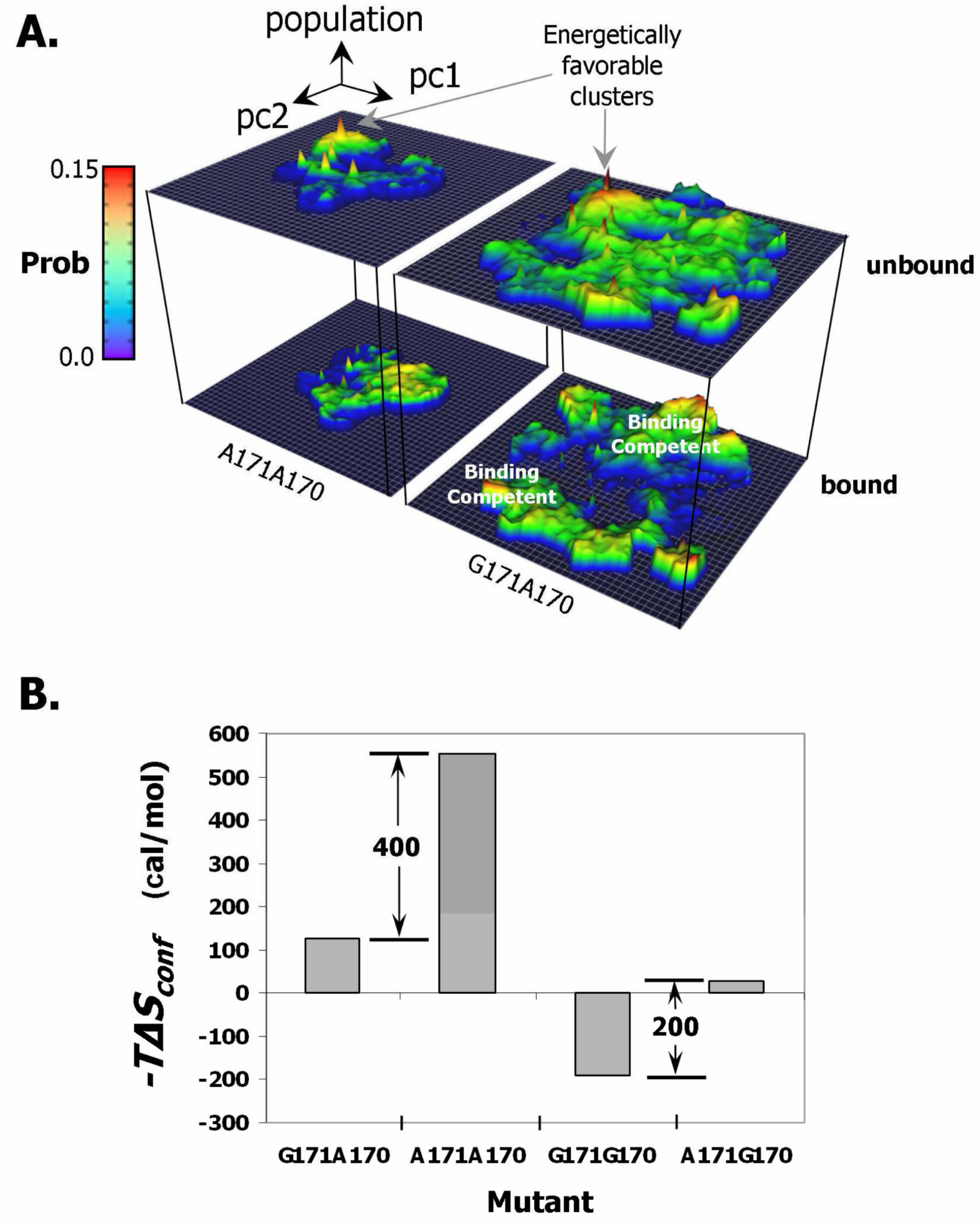

Figure 6.

Comparison of energy landscapes between the unbound and bound ensembles of SH3 variants G171A170 and A171A170. **A)** The fractional occupancy of the unbound (upper) and bound (lower) ensembles for the two SH3 variants was mapped onto the first two principal modes. The region delineating the allowed space has been raised in the figure for clarity. The common regions of both these mutants have similar, but not identical, terrain features, which suggest that the predominant differences in the simulated binding thermodynamics were caused by the non-overlapping region of G171A170. **B)** The change in conformational entropy for the ensembles upon binding ligand was determined from Eq. 3. The constriction of conformational space (T*ΔS_{conf}) due to the Gly to Ala substitution at position 171, which is readily apparent in panel A, was calculated to be ~ 400 cal/mol and was observed to be strongly dependent on the amino acid type in position 170.

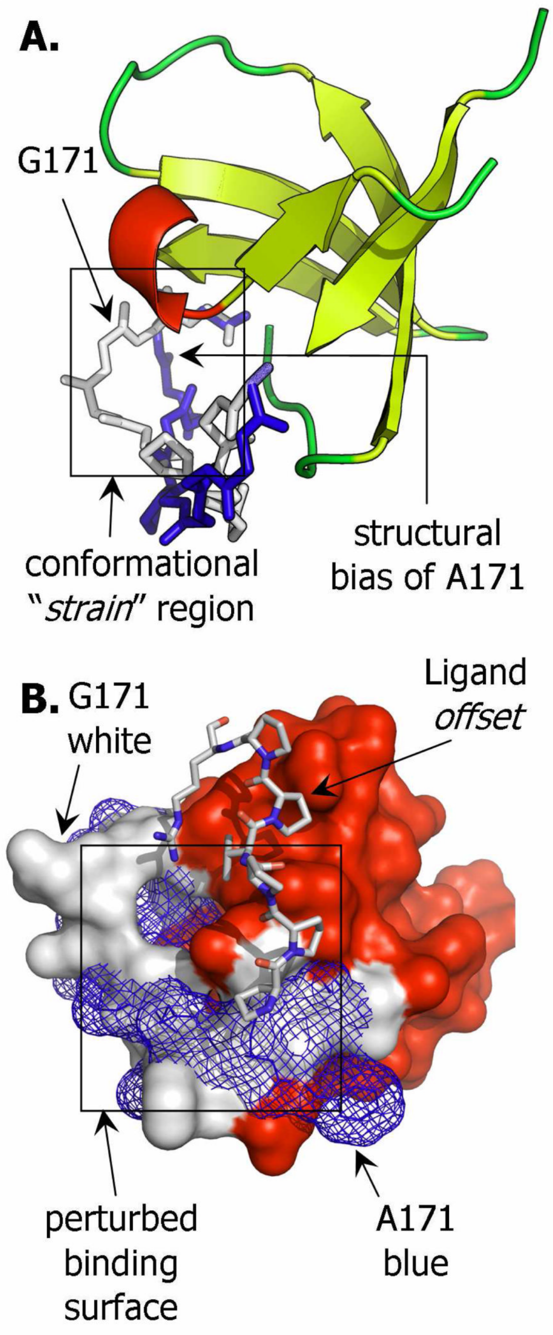

Figure 7.

Structural description of the Gly to Ala substitution at position 171 in the RT loop. **A)** Cartoon rendering of the most highly weighted conformers (within the unbound ensembles) of mutants G171A170 (white) and A171A170 (blue) showing the effect of the G171A mutation on the backbone. The A171A170 mutant shows a pronounced conformational displacement in the RT loop away from the structure of the G171A170 mutant. **B)** The corresponding molecular surface diagrams show a significant difference in the binding surface. The surface of the A171A170 state from panel A (blue mesh) was markedly different from the surface of the G171A170 state (white surface), which corresponded to differences in buried surface areas associated with binding (see Fig 2A), and ultimately to differences in binding energies between the two SH3 variants (see Fig. 2B). The ligand was offset slightly to the right in panel B (relative to its position in the high-resolution complex structure [13]) to better show the binding surface.