BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

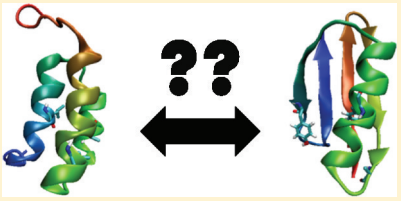
Current Computer Modeling Cannot Explain Why Two Highly Similar Sequences Fold into Different Structures

Jane R. Allison, Maike Bergeler, Niels Hansen, and Wilfred F. van Gunsteren*

Laboratory of Physical Chemistry, Swiss Federal Institute of Technology ETH, 8093 Zürich, Switzerland

Supporting Information

ABSTRACT: The remarkable recent creation of two proteins that fold into two completely different and stable structures, exhibit different functions, yet differ by only a few amino acids poses a conundrum to those hoping to understand how sequence encodes structure. Here, computer modeling uniquely allows the characterization of not only the native structure of each minimally different sequence but also systems in which each sequence was modeled onto the fold of the alternate sequence. The reasons for the different structural preferences of two pairs of highly similar sequences are explored by a combination of structure analyses, comparison of potential energies calculated from energy-minimized single structures and trajectories produced from molecular dynamics simulations, and application of a novel method for calculating free energy differences. The sensitivity of such analyses to the choice of force field is also explored. Many of the hypotheses proposed on the basis of the nuclear magnetic resonance model structures of the proteins with 95% identical sequences are supported. However, each level of analysis provides different predictions regarding which sequence−structure combination should be most favored, highlighting the fact that protein structure and stability result from a complex combination of interdependent factors.



Proteins are synthesized as linear chains of amino acids, but most fold into a well-defined three-dimensional structure to carry out their function. Determining a protein's structure is therefore a first step towards understanding how it functions; correspondingly, one of the holy grails of computational structural biology is to be able to predict the structure of a protein on the basis of its sequence alone. There is a plethora of methods available for protein structure prediction, most of which rely on a combination of template-based modeling, in which databases of known structures are searched for matching sequence fragments, and the use of physics-based force fields or energy functions to assess the relative energies of the predicted structures. Despite this, the fundamental principles underlying the way in which protein sequence determines structure remain poorly understood.

Confounding the difficulty in predicting structure from sequence is the recent design of a pair of proteins that differ by just one residue yet fold into topologically distinct structures.[1] During the switch between the two structures, 85% of the residues change their secondary structure, with just eight residues assuming α-helical structure in both folds. These two proteins, G_A98 and G_B98, are the culmination of a five year project in which two small protein domains were systematically modified to have increasing levels of sequence identity (SI) while retaining different structures and binding functionalities.[1−3] Both were derived from the well-studied binding domains of *Streptococcus* protein G, which binds to serum proteins in the blood. The 45-residue G_A domain binds human serum albumin (HSA), and the 56-residue G_B domain binds the constant (Fc) region of IgG. The naturally occurring versions of these two domains have only 16% SI and fold into 3α and 4β+α structures, respectively.

Two pairs of proteins in the series, G_A95 and G_B95, whose sequences differ by three residues, were the focus of the CASP8 structure prediction competition.[4] Only four of the competing groups correctly predicted the two different folds, with most web servers predicting a 4β+α fold for both sequences, highlighting the difficulty in distinguishing between the subtle conformational preferences encoded by such small differences in sequence.

As part of the characterization of the series of G_A and G_B proteins, NMR model structures were determined for two pairs with 88 and 95% SI,[1,3] providing a useful starting point for computational analyses. Here, the NMR model structures are studied alongside homology models in which each sequence is fitted onto the opposite structure. The latter can be investigated only using simulation. A combination of structural analyses, energies calculated from energy-minimized structures and ensembles generated using molecular dynamics (MD) simulations, and free energies calculated using enveloping distribution sampling (EDS) is used to assess the ability of the physics-based GROMOS force fields to explain the different structural preferences of these minimally different sequences. In doing so, we provide information regarding the relative stabilities of the different sequence−structure combinations that cannot be attained experimentally, and previously suggested hypotheses about the energetic contributions and interactions of specific subsets of residues are investigated. The strengths and weaknesses of computer modeling that become apparent in the context of this

Received: August 12, 2011
Revised: November 11, 2011
Published: November 14, 2011

Table 1. Names and Defining Characteristics of the Different Protein Systems and NOE Data Sets[a]

protein system					NOE data set			
					full		reduced	
name	Seq.	Struct.	SI (%)	PDB	name	number	name	number
$\alpha 88\alpha$	3α	3α	88	2JWS	$\alpha 88\alpha$F	967	$\alpha 88\alpha$R	310
$\beta 88\beta$	$4\beta+\alpha$	$4\beta+\alpha$	88	2JWU	$\beta 88\beta$F	911	$\beta 88\beta$R	260
$\alpha 88\beta$	3α	$4\beta+\alpha$	88	–	–	–	–	–
$\beta 88\alpha$	$4\beta+\alpha$	3α	88	–	–	–	–	–
$\alpha 95\alpha$	3α	3α	95	2KDL	$\alpha 95\alpha$F	830	$\alpha 95\alpha$R	216
$\beta 95\beta$	$4\beta+\alpha$	$4\beta+\alpha$	95	2KDM	$\beta 95\beta$F	1041	$\beta 95\beta$R	239
$\alpha 95\beta$	3α	$4\beta+\alpha$	95	–	–	–	–	–
$\beta 95\alpha$	$4\beta+\alpha$	3α	95	–	–	–	–	–

[a]The Seq. column lists the structure normally associated with this sequence. The Struct. column lists the structure of this construct. The PDB column lists the PDB entries for the NMR model structures. The full (F) NOE data sets contain all NOE distances determined for that protein, and the reduced (R) NOE data sets contain only NOE distances pertaining to backbone atoms.

challenging structure prediction problem are discussed throughout this work.

METHODS

Simulation Setup. All simulations were performed using the GROMOS biomolecular simulation software[5] and the 54A7[6] GROMOS force field. The 45A3[7] and 53A6[8] force fields were also used in the analysis to investigate the sensitivity of the predicted relative stabilities of different sequence−structure combinations to variations in the force field parameters.

The names used to refer to each sequence−structure combination are explained in Table 1. Initial coordinates of $\alpha 88\alpha$, $\beta 88\beta$, $\alpha 95\alpha$, and $\beta 95\beta$ were taken from the NMR model structures deposited in the PDB as entries 2JWS, 2JWU,[3] 2KDL, and 2KDM,[1] respectively. Their amino acid sequences are given in Figure 2. Homology models were created using the I-TASSER server.[9] For instance, to create $\alpha 88\beta$, the sequence of 2JWS was submitted along with the structure of 2JWU. Hydrogens were added to the homology models according to standard geometric criteria. The protonation state of the ionizable residues was chosen according to the pH (7.2) of the NMR experiments. Each sequence−structure combination was subjected to 2000 steps of steepest descent energy minimization in the 45A3,[7] 53A6,[8] or 54A7[6] GROMOS force field.

MD simulations of all sequence−structure combinations were conducted in the 54A7 force field at 298 and 348 K. Each energy-minimized structure was solvated in a rectangular box with a minimal distance from any solute atom to the edge of the box of 1.2 nm. The simple point charge (SPC)[10] water model was used, and periodic boundary conditions were applied. All simulations were initiated with the following equilibration scheme. First, the initial velocities were randomly generated from a Maxwell−Boltzmann distribution at 60 K. All solute atoms were restrained to their positions in the corresponding energy-minimized NMR model or homology-modeled structure through a harmonic potential energy term with a force constant of 2.5×10^4 kJ mol^{-1} nm^{-2}. The system was simulated with these settings for 20 ps, followed by three consecutive 20 ps simulations; prior to each, the temperature was increased by 60 K and the force constant for the positional restraints was reduced by a factor of 10. The position restraints were then removed, and two further 20 ps simulations were conducted, at 298 and 348 K. The final structure saved at each temperature was used as the starting configuration for a 10 ns production run at that temperature. The SHAKE algorithm[11] was used with a geometric precision of 10^{-4} to constrain bond lengths and the water bond angle, allowing for an integration time step of 2 fs. The center of mass motion was removed every 1000 time steps. The temperature and atmospheric pressure were kept constant using a weak-coupling approach[12] with relaxation times τ_T (0.1 ps) and τ_p (0.5 ps) and an isothermal compressibility of 4.575×10^{-4} (kJ mol^{-1} nm^{-3})$^{-1}$. Nonbonded interactions were calculated using a triple-range cutoff scheme. The interactions within a cutoff distance of 0.8 nm were calculated at every step from a pair list that was updated every fifth time step. At this point, interactions between atoms of charge groups within 1.4 nm were also calculated and were kept constant between updates. To account for the influence of the dielectric medium outside the cutoff sphere of 1.4 nm, a reaction field force based on a relative dielectric permittivity ε of 61[13] was added.

The EDS simulations will be described in more detail elsewhere. Briefly, a single-topology approach, in which one topology containing three hybrid building blocks for residues 20, 30 and 45, which differ between G_A95 and G_B95, along with regular building blocks for the remaining amino acids, was used. The single-topology representations of each hybrid building block (Leu/Ala, Ile/Phe, and Leu/Tyr) are given in Figure S1 of the Supporting Information. The energy offset (E^R) and smoothing (s) parameters for the EDS were determined using two different automatic updating schemes.[14,15] Simulation for up to 15 ns was used to obtain the best possible set of parameters (Tables S5 and S6 of the Supporting Information). A 10 ns production run was then performed as described for the normal MD simulations. For some systems, it was necessary to combine two simulations with slightly different energy offsets, because no single parameter combination was able to provide equal sampling of both end states in one EDS reference state simulation (see Tables S4 and S6 of the Supporting Information).

Analysis. The preparation and analysis of structures and simulations were conducted with the GROMOS++ suite of programs.[16] Visualization was done with Visual Molecular Dynamics (VMD).[17] Potential energies were calculated according to the GROMOS 45A3,[7] 53A6,[8] and 54A7[6] force fields. The atom-positional root-mean-square deviation (rmsd) between the Cα atoms of two structures was calculated after superposition of the same set of atoms. The secondary structure of each protein was assigned according to the rules defined by Kabsch and Sander.[18]

RESULTS AND DISCUSSION

The different sequence−structure constructs investigated are outlined in Table 1. The NMR model structures[1,3] and the identity and locations of the residues that differ between the two types of structures with 88 and 95% SI are shown in Figure 1, and the amino acid sequences are given in Figure 2. In addition to studying the NMR model structures, we created homology models in which the sequence that folds into one type of structure was modeled onto the alternate structure. For example, $\alpha 95\beta$ specifies the sequence associated with the 3α structure modeled onto the 95% SI $4\beta+\alpha$ structure. These "crossovers" allowed the energetic cost for each sequence of forming the alternate structure, which is not sufficiently populated to be studied experimentally, to be assessed.

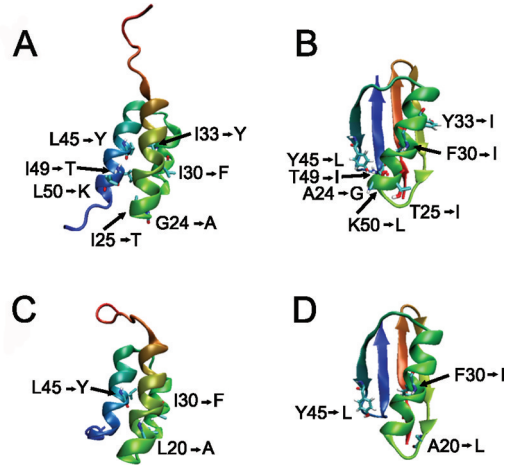


Figure 1. NMR model structures of the experimentally characterized sequence−structure combinations (A) $\alpha 88\alpha$, (B) $\beta 88\beta$, (C) $\alpha 95\alpha$, and (D) $\beta 95\beta$. Coloring is from red to blue from the N- to C-terminus, respectively, for the protein backbone and according to atom type for individual residues. The residue number and the change in residue type upon switching structure (A ↔ B and C ↔ D) or upon creation of homology models (A) $\beta 88\alpha$, (B) $\alpha 88\beta$, (C) $\beta 95\alpha$, and (D) $\alpha 95\beta$ are labeled.

```
     1         10        20        30        40        50
α88  TTYKLILNLKQAKEEAIKELVDAGIAEKYIKLIANAKTVEGVWTLKDEILTFTVTE
β88  TTYKLILNLKQAKEEAIKELVDAATAEKYFKLYANAKTVEGVWTYKDETKTFTVTE

α95  TTYKLILNLKQAKEEAIKELVDAGTAEKYIKLIANAKTVEGVWTLKDEIKTFTVTE
β95  TTYKLILNLKQAKEEAIKEAVDAGTAEKYFKLIANAKTVEGVWTYKDEIKTFTVTE
```

Figure 2. Aligned amino acid sequences. Residues that differ between two proteins with a given SI are colored red.

Verification of the Homology Models. The crossover structures are very similar to their template structures, other than the disordered termini of the 3α structures (Figure 3). To confirm their veracity, the NOE distances were back-calculated and compared to the experimental data used to determine the NMR model structures (Tables S1−S4 of the Supporting Information). From each of the four NOE data sets (one corresponding to each NMR model structure), a reduced data set containing only NOE distances between backbone atoms was extracted (indicated by asterisks in Tables S1−S4 of the Supporting Information). Because these are not specific to residue type, they can be applied to either sequence. This meant that for each crossover, two sets of NOE distances could be calculated: one pertaining to the sequence and one pertaining to the structure (Table 1). For comparison, for each NMR model structure, the NOE distances corresponding to that structure and the reduced set used to determine the alternate structure at the same SI were calculated. Only the results for the 88% SI proteins are shown here (Figure 4), as this is the more stringent test of the homology modeling procedure; the results for the 95% SI proteins are similar (Figure S2 of the Supporting Information).

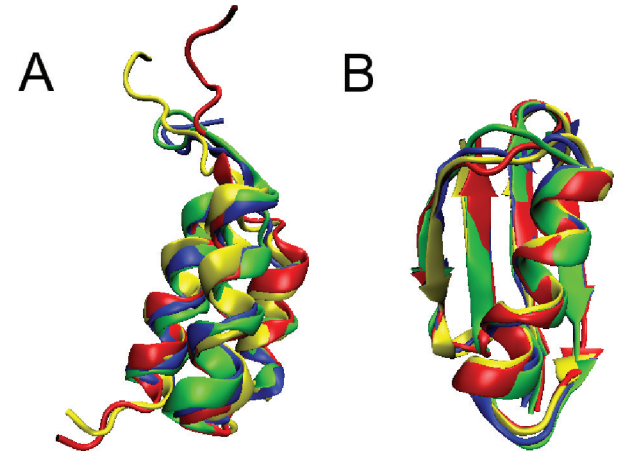


Figure 3. Overlaid structures of (A) $\alpha 88\alpha$ (red), $\beta 88\alpha$ (yellow), $\alpha 95\alpha$ (green), and $\beta 95\alpha$ (blue) and (B) $\beta 88\beta$ (red), $\alpha 88\beta$ (yellow), $\beta 95\beta$ (green), and $\alpha 95\beta$ (blue). The structures are the NMR model or homology-modeled structures, energy-minimized using the GROMOS 54A7 force field[6] and then superimposed via minimization of the atom-positional rmsd between the Cα atoms of residues 9−53.

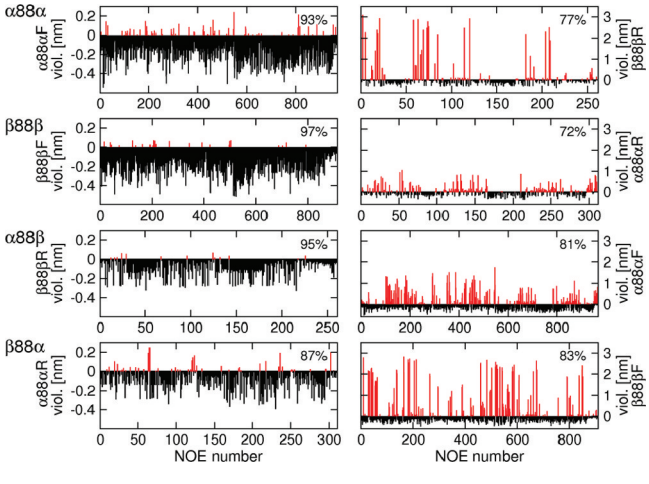


Figure 4. Deviations from the experimentally derived NOE upper distance bounds as a function of the NOE sequence number for the different sequence−structure combinations (indicated at the top left of the four horizontally aligned pairs of panels) and data sets for the sequences at 88% SI as labeled and as outlined in Table 1: (left) NOE data that match the structure and (right) NOE data for the other structure. The NOE distances were calculated from the NMR model or homology-modeled structures after EM using the GROMOS 54A7 force field.[6] The percentage of the data that are satisfied (negative violation) is shown at the top right of each plot.

As expected, when the NOE data set matches the structure, there are few violations, and the violations that do occur are mostly below 0.2 nm (Figure 4). The proportion of the $\beta 88\beta$R NOE data satisfied by $\alpha 88\beta$ (95%) is similar to the proportion of the $\beta 88\beta$F data satisfied by $\beta 88\beta$ (97%), but the proportion of the $\alpha 88\alpha$R NOE data satisfied by $\beta 88\alpha$ (87%) is smaller than

dx.doi.org/10.1021/bi2015663 | Biochemistry 2011, 50, 10965−10973

the proportion of the $\alpha 88\alpha$F data satisfied by $\alpha 88\alpha$ (93%); the positive violations are larger, suggesting that while the α sequence is compatible with the β structure, the β sequence is less suited to the 3α structure. This is in keeping with the results of CASP8, in which none of the algorithms predicted both sequences to be in the 3α structure. This was surmised to be due to steric clashes between A20 and F30 of the G_B95 sequence when it was applied to the 3α structure.

Overall, the crossover structures are in reasonable agreement with the reduced NOE data sets, confirming that at the backbone level, they are similar to the experimentally determined structures. Although the fact that the proportion of positive violations is not more than 28% when the NOE data set for one structure is back-calculated from the alternate structure (e.g., $\alpha 88\alpha$R NOEs calculated from $\beta 88\beta$) might seem to suggest that the information contained in the NMR data is not sufficient to completely define the structure, this effect is largely due to the local nature of most NOE distances. Indeed, the violations that do occur when the data set does not correspond to the structure are much larger in magnitude than those that occur for matching data set−structure pairs.

SASA. According to experiment, despite their high SIs, the distinct folded structures of $\alpha 88\alpha$ and $\beta 88\beta$ are 99.9 and 97% populated, respectively, and the folds of $\alpha 95\alpha$ and $\beta 95\beta$ are both >99% populated,[1] meaning that the alternate sequence−structure combinations cannot be studied experimentally. The advantage of working in silico, therefore, is that all possible sequence−structure combinations can be characterized. To gain some initial insight into why such similar sequences have such different structural preferences, the NMR model structures and the crossover structures, energy-minimized (EM) in the GROMOS 54A7 force field,[6] were analyzed in terms of the solvent accessible surface area (SASA), sorted according to atom and residue type. Some favorability of the native sequence−structure combinations was expected. For instance, hydrophilic amino acids in the α sequence should be more likely to be solvent-exposed in the 3α structure than in the $4\beta+\alpha$ structure, so that the change in the SASA (ΔSASA) upon moving from, for instance, $\alpha 88\alpha$ to $\alpha 88\beta$ should be negative for polar and charged residues and positive for nonpolar residues.

The pattern of ΔSASA with respect to residue number is similar for the 88 and 95% SI constructs (Figure 5), although there are some differences in the magnitude of ΔSASA, indicative of slight differences in the structures formed at each SI level. As noted by Alexander et al.,[1] residues that tend to be exposed in one type of structure are buried in the other. The N- and C-termini are in all cases more exposed in the 3α structure, where they are disordered, than in the $4\beta+\alpha$ structure. This exposure is potentially favorable for the many polar and charged residues present in the termini. Interestingly, the key residues whose mutation precipitates the change in conformation have only relatively small ΔSASA values, although this does not necessarily mean that the nature of the environment is the same in each structure.

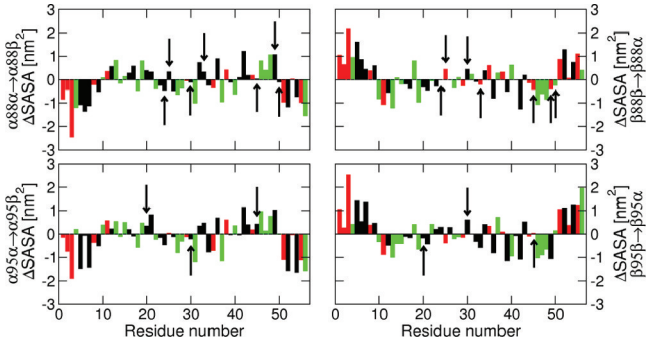


Figure 5. Change in the solvent accessible surface area (ΔSASA) of each residue between pairs of sequence−structure combinations as labeled. The ΔSASA values are colored according to residue type: nonpolar (black), polar (red), and charged (green). The residues that differ at each level of SI are labeled with arrows. The SASAs were calculated from the NMR model or homology-modeled structures after EM using the GROMOS 54A7 force field.[6]

Sorting the SASA according to atom type, rather than residue type, reveals that there is a large decrease in the SASA of hydrophilic atoms and an increase in the SASA of hydrophobic atoms in the crossovers relative to the native combinations (Table 2). The exposure of hydrophobic atoms is in general greater for the 3α structure, especially for the crossovers, in part because of the disordered termini. In the $4\beta+\alpha$ structure, these residues form part of the hydrophobic core. The exposure of hydrophobic atoms is in general disfavored by the simple solvation models that are commonly used in structure prediction algorithms; thus, this result may partially explain why the $4\beta+\alpha$ structure was often the preferred structure for both sequences in CASP8. The apparent contradiction between the (favorable) exposure of hydrophilic residues and the unfavorable exposure of hydrophobic atoms in the termini of the α structure can be reconciled by considering that an amino acid residue classified as hydrophilic may still contain a number of (hydrophobic) carbon atoms. This means that use of a classification of hydrophobicity versus hydrophilicity at the rather coarse-grained level of whole amino acid residues will result in a limited accuracy of fold prediction compared to use of a more fine-grained, atomic level of modeling.

Potential Energies. While the analysis of the SASA gives some insight into the observed structural preferences, it is clear that many other factors must also play a role. Alexander et al.[1] hypothesized that a key determinant of the preferred fold is the stabilization of the $\beta 3-\beta 4$ hairpin in the $4\beta+\alpha$ structure by interactions among residues 45, 47, and 52. They also discussed the markedly different roles of residues in the N- and C-termini in the two different structures, as was already noted in the analysis of the SASA. The advantages of using computational modeling to investigate these hypotheses are that all sequence−structure combinations can be studied, and the potential energy may be calculated for any chosen combination of atoms or residues to elucidate their role in determining the relative stability of each construct. The caveat, of course, is that the precise value of the potential energy calculated depends on the choice of force field. Therefore, the sensitivity of the results to the choice of force field was first tested by calculating the intraprotein potential energy for each sequence−structure combination after energy minimization (EM) in a vacuum for each of three different GROMOS force fields, 45A3,[7] 53A6,[8] and 54A7.[6] The differences in potential energy (ΔE_{EM}) between pairs of related sequence−structure combinations at a given SI are shown in Figure 6.

Each force field produces the same trends in ΔE_{EM}, but the magnitudes differ. This is particularly noticeable for the change from $\beta 88\beta$ to $\beta 88\alpha$, where the sign of ΔE_{EM} differs between force fields. The known destabilization of α helices in the 53A6 force field compared to the 45A3 and 54A7 force fields[19−21] is not observed here. For each SI (88 and 95%), the order of the potential energies (Figure 6) roughly correlates with the total

Table 2. SASAs of Each Sequence−Structure Combination[a]

	total			hydrophobic			hydrophilic		
name	SASA	N_{atoms}	$SASA_{at}$	SASA	N_{atoms}	$SASA_{at}$	SASA	N_{atoms}	$SASA_{at}$
$\alpha 88\alpha$	47.7	445	0.107	30.6	290	0.106	17.1	155	0.110
$\beta 88\beta$	40.2	456	0.088	22.3	296	0.075	17.9	160	0.112
$\alpha 88\beta$	38.0	445	0.085	27.1	290	0.094	10.9	155	0.070
$\beta 88\alpha$	45.1	456	0.099	32.5	296	0.110	12.6	160	0.079
$\alpha 95\alpha$	48.2	445	0.108	30.0	288	0.104	18.2	157	0.116
$\beta 95\beta$	41.9	449	0.093	23.9	291	0.082	18.0	158	0.114
$\alpha 95\beta$	39.2	445	0.088	28.2	288	0.098	11.0	157	0.070
$\beta 95\alpha$	47.2	449	0.105	33.0	291	0.113	14.2	158	0.090

[a]The SASA is in square nanometers. N_{atoms} is the number of atoms of a given type. $SASA_{at}$ is the SASA per atom. The SASAs were calculated for the NMR model or homology-modeled structures after EM using the GROMOS 54A7 force field.[6]

SASA per atom (Table 2): SASA($\alpha 88\beta$) < SASA($\beta 88\beta$) < SASA($\beta 88\alpha$) < SASA($\alpha 88\alpha$), and $E_{EM}(\alpha 88\beta) < E_{EM}(\beta 88\beta) = E_{EM}(\beta 88\alpha) < E_{EM}(\alpha 88\alpha)$. The same was true for the 95% constructs. In other words, the more compact the protein, the lower (more favorable) its intraprotein potential energy. Because the 3α structure is less compact than the $4\beta+\alpha$ structure, the stability of the former will be more dependent on interactions with the solvent than on intraprotein interactions. Further evidence that the protein−solvent interactions need to be taken into account is the fact that in all three force fields, it is the crossovers and not the native sequence−structure combinations that are the most favorable in terms of intraprotein potential energy, as shown by the negative ΔE_{EM} with a change from a native construct to a crossover.

To investigate the key role of residue 45 as a conformational switch, we calculated its interaction (potential) energy with residue 47, residue 52, and all residues in the protein, for all sequence−structure combinations (Table 3). In the $\beta 95\beta$ NMR model structure, Y45 was observed to form a strong hydrogen bond with D47 and be closely packed against F52.[1] The stabilization of the $\beta 3-\beta 4$ hairpin by these interactions was observed in previous studies of protein G mutants.[22] During the derivation of the series of G_A/G_B pairs, it was seen that

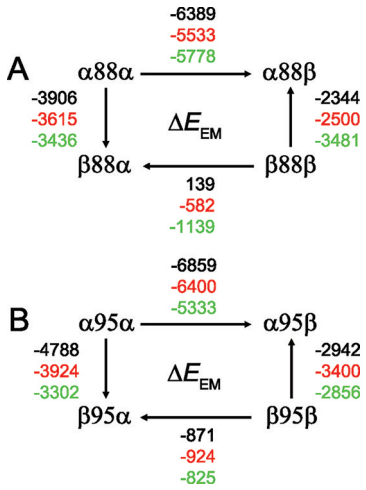


Figure 6. Differences in the intraprotein potential energy ΔE_{EM} (kilojoules per mole) when moving in the direction of the arrows between the different sequence−structure combinations as labeled and as outlined in Table 1 for (A) 88% SI and (B) 95% SI. The energies were calculated for the NMR model or homology-modeled structures after EM in the (black) 45A3,[7] (red) 53A6,[8] or (green) 54A7[6] GROMOS force field.

when residue 45 is Y and residue 52 is F, the $\beta 3-\beta 4$ hairpin is favored, but if one of them has another type, the 3α structure is formed instead. In this way, local structural preferences ultimately influence the overall tertiary structure of the protein.

At the 95% SI level, each sequence has a more favorable $E^{tot}_{EM}(45{:}47{,}52)$ when in its native structure, in agreement with the key role of residue 45 in switching between the two possible conformations. At 88% SI, residue 45 actually has more favorable interactions overall when the sequence and structure are mismatched; however, other residues in the vicinity of the $\beta 3-\beta 4$ hairpin also differ at this level of SI (see Figure 2), complicating the analysis.

The hydrogen bond between residues 45 and 47 noted by Alexander et al.[1] is clearly evident in $\beta 95\beta$ [negative $E^{crf}_{EM}(45{:}47)$], but not in $\beta 88\beta$, where the hydroxyl group of Y45 is oriented away from residue 47 and instead interacts with the solvent, which is not included in this analysis. There are also favorable interactions between residues L45 and D47 in $\alpha 88\beta$, indicating that both sequences can provide favorable interactions for residue 45 in the β structure. Additionally, the favorable packing interactions between residues 45 and 52 occur when either sequence is in the $4\beta+\alpha$ structure at both the 88 and 95% SI levels, indicating that this interaction is not dependent on residue 45 being aromatic.

Overall, although there is some evidence that residue 45 forms favorable interactions that may stabilize the $\beta 3-\beta 4$ region of the β structure, their presence in both the native and mismatched sequence−structure combinations prevents this from being a key indicator of structural preference.

Residues 9−51 are structured in both folds. Only residues 27−33 retain the same (α-helical) secondary structure in both types of structure, however, with all other residues undergoing an alteration in secondary structure upon changing structure. Residues 1−8 and 52−56 are unstructured in the 3α structure and form a β strand in the $4\beta+\alpha$ structure at 95% SI. When the sequence switches from the 3α structure to the $4\beta+\alpha$ structure, the hydrophobic interactions in the core of the helix bundle are released and replaced by hydrophobic interactions between the preserved central α helix and the central β strands, which are formed from the N- and C-termini. Alexander et al.[1] proposed that the competition between forming the helix bundle, in which the N- and C-termini are disordered, and recruiting the termini to form the β sheets of the $4\beta+\alpha$ structure is mediated by the preferred binding partners of residues 20, 30, and 45 in each sequence. According to their hypothesis, when the α sequence forms the 3α structure, L20, I30, and L45 prefer to bind to residues A16, I33, and I49 (3α core), whereas when the β

Table 3. Noncovalent Intraprotein Potential Energies of Residue 45 with All Other Residues[a]

	residue 47			residue 52			all residues		
name	E_{EM}^{lj}	E_{EM}^{crf}	E_{EM}^{tot}	E_{EM}^{lj}	E_{EM}^{crf}	E_{EM}^{tot}	E_{EM}^{lj}	E_{EM}^{crf}	E_{EM}^{tot}
$\alpha 88\alpha$	−4	1	−3	0	−1	−2	−52	−144	−196
$\beta 88\beta$	−9	1	−8	−12	−4	−16	−74	−155	−229
$\alpha 88\beta$	−10	−15	−25	−13	−1	−14	−55	−176	−230
$\beta 88\alpha$	−4	−4	−8	0	−1	−1	−83	−159	−242
$\alpha 95\alpha$	−3	4	0	0	−1	−1	−66	−126	−192
$\beta 95\beta$	−10	−57	−67	−11	−5	−16	−63	−183	−246
$\alpha 95\beta$	−5	5	0	−12	−1	−14	−56	−114	−170
$\beta 95\alpha$	−4	4	0	0	−2	−2	−68	−142	−210

[a]All energies are in kilojoules per mole. lj denotes the Lennard-Jones energy, crf the Coulomb and reaction field energy, and tot the total nonbonded potential energy. The energies were calculated for the NMR model or homology-modeled structures after EM using the GROMOS 54A7 force field.[6]

sequence forms the $4\beta+\alpha$ structure, A20, F30, and Y45 prefer to bind to residues Y3, L5, L7, F52, and V54 ($4\beta+\alpha$ core).

The interaction energies between these sets of residues are listed in Table 4. Supporting the conjecture of Alexander et al.[1]

Table 4. Noncovalent Intraprotein Potential Energies of Residues 20, 30, and 45 with Residues Comprising the Hydrophobic Core[a]

	3α core			4$\beta+\alpha$ core		
name	E_{EM}^{lj}	E_{EM}^{crf}	E_{EM}^{tot}	E_{EM}^{lj}	E_{EM}^{crf}	E_{EM}^{tot}
$\alpha 88\alpha$	−21	−44	−65	−1	−2	−3
$\beta 88\beta$	−9	−15	−24	−45	−4	−49
$\alpha 88\beta$	−8	−7	−15	−38	3	−35
$\beta 88\alpha$	−38	−49	−87	−1	−1	−2
$\alpha 95\alpha$	−18	−34	−52	0	−1	−1
$\beta 95\beta$	−11	−11	−22	−40	−3	−43
$\alpha 95\beta$	−8	−10	−18	−37	2	−35
$\beta 95\alpha$	−16	−48	−63	0	−2	−2

[a]All energies are in kilojoules per mole. lj denotes the Lennard-Jones energy, crf the Coulomb and reaction field energy, and tot the total noncovalent potential energy. 3α core refers to residues 16, 33, and 49, and 4$\beta+\alpha$ core refers to residues 3, 5, 7, 52, and 54. The energies were calculated for the NMR model or homology-modeled structures, energy-minimized using the GROMOS 54A7 force field.[6]

with regard to the identity of the residues involved, the total noncovalent potential energy E_{EM}^{tot}(core) is more favorable for the 3α core when the 3α structure is formed and for the 4$\beta+\alpha$ core when the 4$\beta+\alpha$ structure is formed at both 88 and 95% SI. Interestingly, however, E_{EM}^{tot}(core) for the 3α core is more favorable for $\beta 88\alpha$ than for $\alpha 88\alpha$ and for $\beta 95\alpha$ than for $\alpha 95\alpha$, suggesting that the β sequence is better able to stabilize the α core than the α sequence. In comparison, the interactions between residues 20, 30, and 45 and the 4$\beta+\alpha$ core, which are dominated by E_{EM}^{lj}(core), are more favorable for the β sequence and virtually nonexistent for the 3α structure because of the disordered nature of the termini. From these energies, it can be concluded that while the van der Waals interactions that stabilize the 4$\beta+\alpha$ structure can be formed only by the 4$\beta+\alpha$ sequence, the interactions comprising the α core can be formed by either sequence.

MD Simulations. Comparing properties calculated from individual structures gives an idea about whether force fields used as scoring functions are capable of distinguishing native sequence−structure combinations. Proteins do not exist as single structures, however, but as an ensemble of structures; thus, it is also important to examine their behavior during MD simulations, in which a Boltzmann-weighted ensemble of structures is sampled. MD simulations of the protein in water were run using the GROMOS 54A7 force field[6] at 298 and 348 K for all sequence−structure combinations. As a simple estimate of the structural stability during the simulation, the atom-position root-mean-square deviations (rmsds) of the Cα atoms of each sequence−structure combination from the first structure of the trajectory were calculated (Figure 7). The

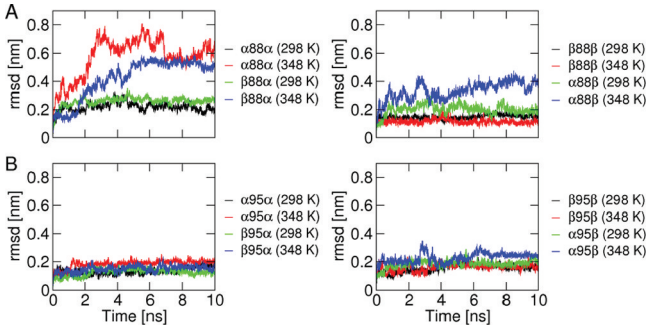


Figure 7. Time series of the atom-positional rmsd of the Cα atoms from the first structure of the trajectory during 10 ns MD simulations of the various sequence−structure combinations in water using the 54A7 force field[6] at different temperatures and with (A) 88% SI or (B) 95% SI. The superposition and rmsd were determined for the Cα atoms of all residues of the 4$\beta+\alpha$ structure and of residues 10−50 of the 3α structure.

nature of the structural changes corresponding to the increases in rmsd can be seen in the secondary structure analysis of the different sequence−structure combinations (Figure 8). Only the analysis of the 298 K simulations of the 88% SI proteins are shown because of the high stability of the 95% SI proteins (see below and Figures S3−S5 of the Supporting Information).

At the 88% SI level, the 3α structure is the least stable, exhibiting a larger rmsd than the 4$\beta+\alpha$ structure in all cases, even though the rmsd was only calculated for the residues in regular secondary structure (10−50). Both $\alpha 88\alpha$ and $\beta 88\alpha$ have rmsds in excess of 0.5 nm at 348 K, which might suggest that both sequences are equally stable in the α structure. The causes of the high rmsd are different, however. The three α helices of $\alpha 88\alpha$ are extremely stable, with the rmsd caused by displacements of the helices relative to one another, whereas for $\beta 88\alpha$, the ends of the helices disintegrate and the first α helix fluctuates throughout the simulation. For the β structure, $\beta 88\beta$ is stable at 298 and 348 K, but $\alpha 88\beta$ diverges from the initial structure at both temperatures. At 298 K, there is fraying of the

dx.doi.org/10.1021/bi2015663 | Biochemistry 2011, 50, 10965−10973

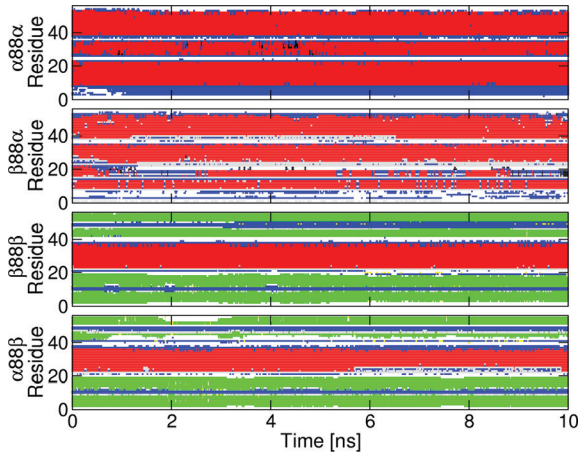


Figure 8. Time series of the secondary structure [(black) 3_{10} helix, (red) α helix, (cyan) π helix, (blue) bend, (yellow) β bridge, (green) β strand, and (gray) turn] of the various 88% SI sequence–structure combinations during 10 ns MD simulations of the protein in water using the 54A7 force field[6] at 298 K as labeled and as described in Table 1.

N-terminus of the α helix and disruption to the secondary structure in the region around the $\beta3-\beta4$ hairpin. This is in agreement with the hypothesis of Alexander et al.[1] that favorable interactions among residues 45, 47, and 52 of the $3\beta+\alpha$ sequence are required to stabilize the hairpin structure, but in contrast to the interaction energies calculated from the EM structures, highlighting the additional insight gained from including the solvent and conducting simulations.

At the 95% SI level, all sequence–structure combinations exhibit a similar stability during the MD simulations at both temperatures. This increased thermal stability compared to that of the 88% SI constructs is in contrast to the experimentally determined melting temperatures (T_M), which decrease with increasing SI.[1] The similar rmsd of the native and crossover sequence–structure combinations highlights the fact that even with state-of-the-art force fields, the different structural preferences of these very similar sequences are difficult to distinguish from analysis of 10 ns simulation trajectories.

According to the potential energies averaged over the 10 ns MD simulations, the $4\beta+\alpha$ structure is favored in terms of intraprotein ΔE^{p}_{MD} at both 88 and 95% SI levels (Figure 9).

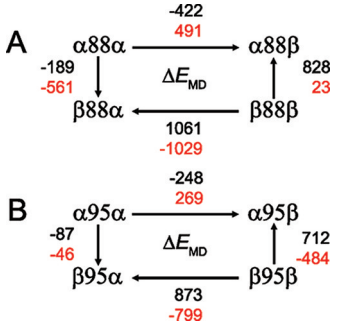


Figure 9. Differences in the (black) intraprotein ΔE^{p}_{MD} and (red) intraprotein and protein–solvent noncovalent potential energies ΔE^{p+ps}_{MD} (kilojoules per mole) when moving in the direction of the arrows between the different sequence–structure combinations as labeled and as outlined in Table 1 for (A) 88% SI and (B) 95% SI. The ΔE_{MD} values are between the averages over the entire 10 ns MD simulations of the protein in water using the 54A7 force field.[6]

This is in keeping with the negative ΔE_{EM} for $\alpha88\alpha \rightarrow \alpha88\beta$ and $\alpha95\alpha \rightarrow \alpha95\beta$ transitions but contradicts the negative ΔE_{EM} for the $\beta95\beta \rightarrow \beta95\alpha$ transition (Figure 6). Once the protein–solvent interactions are taken into account, however, it is the 3α structure that is more favorable for both SI values. This is partially due to favorable interactions with the solvent of the residues in the disordered N- and C-termini, which are instead involved in the hydrophobic core in the $4\beta+\alpha$ structure. This is in keeping with the predictions made on the basis of the analysis of the SASA. The higher rmsd for simulations starting from the 3α structures may therefore be caused by the protein exchanging internal interactions for favorable interactions with water.

Considering changes in sequence for a given structure, it is more favorable for the β sequence to form the $4\beta+\alpha$ structure than the α sequence at the 88% SI level, but not at the 95% SI level. However, it is more favorable for the β sequence to form the 3α structure than the α sequence at both SI levels. Combined with the overall greater stability of the 3α structure once protein–solvent interactions are taken into account, this would imply that both sequences should fold into the 3α structure. The difference between this result and the predictions of CASP8 reveals how important it is to include a realistic description of solvation effects in structure prediction methods. Clearly, there are further factors, not captured by the modeling procedures, that are important for determining the structural preferences of these highly similar sequences observed experimentally.

One significant factor not captured in the average potential energies calculated from the MD simulations, however, is entropy. We therefore also calculated the Gibbs free energy, or free enthalpy, differences and their differences between the 3α and $4\beta+\alpha$ structures for all possible combinations of the three residues that differ between the 95% SI sequences using the EDS methodology[14,23–25] (Table 5). This method eliminates the need to define a priori a reference state or the pathway between different states.

Table 5. Gibbs Free Energy Differences (ΔG) and Their Differences ($\Delta\Delta G$) upon Changing Residues 20, 30, and 45[a]

change in sequence (L20I30L45 →)	ΔG		$\Delta\Delta G$
	3α structure	$4\beta+\alpha$ structure	
A20I30L45	7.7 ± 2.0	−11.7 ± 1.7	−19.4 ± 2.6
L20**F30**L45	−18.3 ± 1.6	−17.4 ± 1.2	0.9 ± 2.1
L20I30**Y45**	−76.5 ± 0.8	−92.2 ± 1.4	−15.7 ± 1.6
A20F30Y45	−90.0 ± 2.1	−125.9 ± 2.0	−35.9 ± 2.9

[a]All free energies are in kilojoules per mole. They were calculated from MD simulations of the proteins in aqueous solution. The errors in $\Delta\Delta G$ were estimated as the square root of the sum of the squares of the errors in ΔG.

The changes in free energy resulting from changing each amino acid individually are roughly additive, indicating that the effects of the changes are largely independent of one another. This is not overly surprising considering the spatial distances in both structures between the three residues that are changed (Figure 1). The Gibbs free energy differences (ΔG) for the single mutations show that changing residue 20 from Leu (α sequence) to Ala (β sequence) is unfavorable in the 3α structure and favorable in the $4\beta+\alpha$ structure, whereas changing residue 45 from Leu (α sequence) to Tyr (β sequence) is favorable in

both structures but most favorable in the 4β+α structure. The preferred nature of residue 30 is Phe (β structure) in both structures. The free energy difference upon introduction of all three changes at once, that is, changing from the α sequence to the β sequence (L20I30L45 → A20F30Y45), is negative for both structures but is most negative for the 4β+α structure, resulting in a $\Delta\Delta G$ of −36 kJ/mol. Thus, overall, the inclusion of entropy results in conclusions slightly different from those obtained from only considering the energy, in that the β sequence is more favorable in both structures, and most favorable in the 4β+α structure (note that this was the finding at the 88% SI level from the MD simulations, but not at the 95% SI level).

CONCLUSIONS

Predicting protein structure from amino acid sequence is one of the greatest challenges of biomolecular modeling. While force field quality continues to improve, proteins such as those studied here, where the change of just a few key residues causes a whole-sale change in the structure, present an extremely difficult task. Analysis of the native and crossover sequence−structure combinations showed that factors that intuitively seem important, such as changes in the hydrophobic versus hydrophilic SASA at the coarse-grained residue level, explain little about the structural preferences, whereas considering local interactions at the finer-grained atomic level provides stronger indications of fold preferences. Simple energy calculations in vacuum showed that the mismatched sequence−structure combinations are the most favorable in all GROMOS force fields in terms of intraprotein noncovalent potential energy, highlighting the importance of including solvent effects. The variation in the magnitudes of the energy differences among the three force fields gave some indication as to the sensitivity of structure prediction to the choice of force field. Looking in more detail at the interaction energy between groups of residues proposed to play an important role in the structural preferences mostly supported the hypotheses proposed by Alexander et al.[1] but did not provide sufficient grounds for predicting the structural preference of each sequence.

The MD simulations highlighted the importance of including the effects of solvent, as the intraprotein noncovalent potential energies of the 4β+α structure are more favorable, but when the protein−solvent energy is accounted for, the 3α structure is more favorable. In addition, the limitations of current force fields in terms of predicting the effects of small changes in the sequence are apparent from the almost indistinguishable behavior of all sequence−structure combinations with 95% SI at both 298 and 348 K. For sequences that are only 88% similar, however, differences between the native and mismatched sequence−structure combinations begin to emerge, and the locations of the structural instability during the simulations were in keeping with hypotheses about the key residues responsible for determining the preferred fold.

The free energy calculations, in which the entropy is also taken into account, showed that the three residues that differ at the 95% SI level act essentially independently. According to the Gibbs free energies, the β sequence is preferable to the α sequence in both structures and is most favorable in the 4β+α structure, in keeping with some, but not all, of the conclusions from the energy analysis of the MD simulations.

Overall, this study reveals that in addition to the difficulty of predicting the structure of a protein given only the sequence, widely used force fields and modeling and analysis techniques are often not able to differentiate between alternative structures for a given sequence, although a fine-grained atomic-level analysis that accounts for solvent effects and entropic contributions to the differential stability of proteins by considering Boltzmann ensembles of structures does better than a coarse-grained residue-level analysis of single protein structures in a vacuum. Admittedly, the sequence−structure changes considered here comprise a particularly difficult case, in which the two sequences differ by only a few residues yet fold into topologically quite different structures. Despite this, it is clear that continued improvements in biomolecular modeling tools are required.

ASSOCIATED CONTENT

Supporting Information

Tables of the NOE data, tables of the final parameters for the EDS simulations, hybrid single topologies used for EDS simulations, plot of the NOE violations for the 95% SI proteins, plots showing the secondary structure during the MD simulations of the 88% SI proteins at 348 K and the 95% SI proteins at 298 and 348 K. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: wfvgn@igc.phys.chem.ethz.ch. Phone: +41 44 632 5501. Fax: +41 44 632 1039.

Funding

This work was supported by the National Center of Competence in Research (NCCR) in Structural Biology, by Grant 200020-121913 from the Swiss National Science Foundation, and by Grant 228076 from the European Research Council.

ABBREVIATIONS

ΔE, difference in potential energy; EM, energy minimization; MD, molecular dynamics; NMR, nuclear magnetic resonance; PDB, Protein Data Bank; rmsd, root-mean-square deviation; SASA, solvent accessible surface area; SI, sequence identity; SPC, simple point charge.

REFERENCES

(1) Alexander, P. A., He, Y., Chen, Y., Orban, J., and Bryan, P. N. (2009) A minimal sequence code for switching protein structure and function. *Proc. Natl. Acad. Sci. U.S.A. 106*, 21149−21154.

(2) Alexander, P. A., He, Y., Chen, Y., Orban, J., and Bryan, P. N. (2007) The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc. Natl. Acad. Sci. U.S.A. 104*, 11963−11968.

(3) He, Y., Chen, Y., Alexander, P., Bryan, P. N., and Orban, J. (2008) NMR structures of two designed proteins with high sequence identity but different fold and function. *Proc. Natl. Acad. Sci. U.S.A. 105*, 14412−14417.

(4) Horst, J., and Samudrala, R. (2009) Diversity of protein structures and difficulties in fold recognition: The curious case of protein G. *F1000 Biol Reports 1*, 69.

(5) Christen, M., Hünenberger, P. H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D. P., Heinz, T. N., Kastenholz, M. A., Kräutler, V., Oostenbrink, C., Peter, C., Trzesniak, D., and van Gunsteren, W. F. (2005) The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem. 26*, 1719−1751.

(6) Schmid, N., Eichenberger, A., Choutko, A., Riniker, S., Winger, M., Mark, A., and van Gun-steren, W. (2011) Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J. 40*, 843−856.

(7) Schuler, L. D., Daura, X., and van Gunsteren, W. F. (2001) An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem. 22*, 1205−1218.
(8) Oostenbrink, C., Villa, A., Mark, A. E., and van Gunsteren, W. F. (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem. 25*, 1656−1676.
(9) Zhang, Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinf. 9*, 40.
(10) Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., and Hermans, J. (1981) In *Intermolecular Forces* (Pullmann, B., Ed.) pp 331−342, Reidel: Dordrecht, The Netherlands.
(11) Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys. 23*, 327−341.
(12) Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys. 81*, 3684−3690.
(13) Heinz, T. N., van Gunsteren, W. F., and Hünenberger, P. H. (2001) Comparison of four methods to compute the dielectric permittivity of liquids from molecular dynamics simulations. *J. Chem. Phys. 115*, 1125−1136.
(14) Christ, C. D., and van Gunsteren, W. F. (2009) Simple, efficient, and reliable computation of multiple free energy differences from a single simulation: A reference Hamiltonian parameter update scheme for Enveloping Distribution Sampling (EDS). *J. Chem. Theory Comput. 5*, 276−286.
(15) Hansen, N., Dolenc, J., Knecht, M., Riniker, S., and van Gunsteren, W. (2011) Assessment of Enveloping Distribution Sampling to calculate relative free enthalpies of binding for eight netropsin-DNA duplex complexes in aqueous solution. *J. Comput. Chem.*, in press.
(16) Eichenberger, A., Allison, J., Dolenc, J., Geerke, D., Horta, B., Meier, K., Oostenbrink, C., Schmid, N., Steiner, D., Wang, D., and van Gunsteren, W. (2011) GROMOS++ software for the analysis of biomolecular simulation trajectories. *J. Chem. Theory Comput. 7*, 3379−3390.
(17) Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD: Visual Molecular Dynamics. *J. Mol. Graphics 14*, 33−38.
(18) Kabsch, W., and Sanders, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*, 2577−2637.
(19) Cao, Z., Lin, Z., Wang, J., and Liu, H. (2009) Refining the description of peptide backbone conformations improves protein simulations using the GROMOS 53A6 force field. *J. Comput. Chem. 30*, 645−660.
(20) Zagrovic, B., Gattin, Z., Lau, J., Huber, M., and van Gunsteren, W. (2008) Structure and dynamics of two β-peptides in solution from molecular dynamics simulations validated against experiment. *Eur. Biophys. J. 37*, 903−912.
(21) Matthes, D., and de Groot, B. L. (2009) Secondary structure propensities in peptide folding simulations: A systematic comparison of molecular mechanics interaction schemes. *Biophys. J. 97*, 599−608.
(22) Sari, N., Alexander, P., Bryan, P. N., and Orban, J. (2000) Structure and dynamics of an acid-denatured protein G mutant. *Biochemistry 39*, 965−977.
(23) Christ, C. D., and van Gunsteren, W. F. (2007) Enveloping distribution sampling: A method to calculate free energy differences from a single simulation. *J. Chem. Phys. 126*, 184110.
(24) Christ, C. D., and van Gunsteren, W. F. (2008) Multiple free energies from a single simulation: Extending enveloping distribution sampling to nonoverlapping phase-space distributions. *J. Chem. Phys. 128*, 174112.
(25) Christ, C. D., and Van Gunsteren, W. F. (2009) Comparison of three enveloping distribution sampling Hamiltonians for the estimation of multiple free energy differences from a single simulation. *J. Comput. Chem. 30*, 1664−1679.