LU TP 03-14
June 25, 2003

Sequence-Based Study of Two Related Proteins with Different Folding Behaviors

Giorgio Favrin, Anders Irbäck and Stefan Wallin*

Complex Systems Division, Department of Theoretical Physics
Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden
http://www.thep.lu.se/complex/

Submitted to *Proteins*

Abstract:
Z_{SPA-1} is an engineered protein that binds to its parent, the three-helix-bundle Z domain of staphylococcal protein A. Uncomplexed Z_{SPA-1} shows a reduced helix content and a melting behavior that is less cooperative, compared with the wild-type Z domain. Here we show that the difference in folding behavior between these two sequences can be partly understood in terms of an off-lattice model with 5–6 atoms per amino acid and a minimalistic potential, in which folding is driven by backbone hydrogen bonding and effective hydrophobic attraction.

Keywords: protein folding, folding thermodynamics, three-helix bundle, unstructured protein, Monte Carlo simulation.

*E-mail: favrin, anders, stefan@thep.lu.se

1 Introduction

It is becoming increasingly clear that unstructured proteins play an important biological role [1, 2]. In many cases, such proteins adopt a specific structure upon binding to their biological targets. Recently, it was demonstrated that the *in vitro* evolved Z_{SPA-1} protein [3] exhibits coupled folding and binding [4].

Z_{SPA-1} is derived from the Z domain of staphylococcal protein A, a 58-amino acid, well characterized [5] three-helix-bundle protein. Z_{SPA-1} was engineered [3] by randomizing 13 amino acid positions and selecting for binding to the Z domain itself. Subsequently, the structure of the Z:Z_{SPA-1} complex was determined both in solution [4] and by crystallography [6]. In the complex, both Z_{SPA-1} and the Z domain adopt structures similar to the solution structure of the Z domain. However, in solution, Z_{SPA-1} does not behave as the Z domain; Wahlberg *et al.* [4] found that uncomplexed Z_{SPA-1} lacks a well-defined structure, and that its melting behavior is less cooperative than that of the Z domain.

The Z domain is a close analog of the B domain of protein A, a chain that is known to show two-state folding without any meta-stable intermediate state [7, 8]. The folding behavior of the B domain has been studied theoretically by many different groups, including ourselves, using both all-atom [9–12] and reduced [13–17] models. In many cases, it was possible to fold this chain, but to achieve that most models rely on the so-called Gō prescription [18]. Our model [17] is, by contrast, sequence-based. This makes it possible for us to study both Z_{SPA-1} and the wild-type Z domain and compare their behaviors, using one and the same model.

The purpose of this note is twofold. First, we check whether our model can explain the difference in melting behavior between Z_{SPA-1} and the wild-type sequence. Second, using this model, we study the structure of Z_{SPA-1}.

2 Materials and Methods

2.1 Model

The model we study [17] is an extension of a model with three amino acids [19–21] to a five-letter alphabet. The five amino acid types are hydrophobic (Hyd), polar

(Pol), Ala, Pro and Gly. Hyd, Pol and Ala share the same geometric representation but differ in hydrophobicity. Pro and Gly have their own geometric representations.

The Hyd, Pol and Ala representation contains six atoms. The three backbone atoms N, C_α and C' and the H and O atoms of the peptide unit are all included. The H and O atoms are used to define hydrogen bonds. The sixth atom is a large C_β that represents the side chain. Gly lacks the C_β atom but is otherwise the same. The representation of Pro differs from that of Hyd, Pol and Ala in that the H atom is replaced by a side-chain atom, C_δ, and that the Ramachandran angle ϕ is held fixed at $-65°$.

The degrees of freedom of our model are the Ramachandran torsion angles ϕ and ψ, with the exception that ϕ is held fixed for Pro. All bond lengths, bond angles and peptide torsion angles (180°) are held fixed.

The interaction potential

$$E = E_{\rm loc} + E_{\rm ev} + E_{\rm hb} + E_{\rm hp} \tag{1}$$

is composed of four terms. The first term is a local ϕ, ψ potential. The other three terms represent excluded volume, backbone hydrogen bonds and effective hydrophobic attraction, respectively (no explicit water). For simplicity, the hydrophobicity potential is taken to be pairwise additive. Only Hyd-Hyd and Hyd-Ala C_β pairs experience this type of interaction. In particular, this means that Ala is intermediate in hydrophobicity between Hyd and Pol. The amino acids in the Hyd class are Val, Leu, Ile, Phe, Trp and Met, whereas those in the Pol class are Arg, Asn, Asp, Cys, Gln, Glu, His, Lys, Ser, Thr and Tyr. A complete description of the model, including numerical values of all the parameters, can be found in our earlier study [17].

In this earlier study, the model was applied to the 10–55-amino acid fragment of the B domain of protein A. Despite the simplicity of the potential, this sequence was found to have the following properties [17] in the model:

- It does make a three-helix bundle with the native three-helix-bundle topology,[†] although the suppression of the wrong topology is not very strong. All helices are right-handed, as they should.
- Energy minimization restricted to the thermodynamically favored (native) topology gives a structure with a root-mean-square deviation (RMSD) of 1.8 Å from the NMR structure [22] (calculated over all backbone atoms).

[†]There are two possible three-helix-bundle topologies; if we let the first two helices form a U, then the third helix can be either in front of or behind this U.

Z	QQN	AFY	EIL	HLP	NLN	EEQ	RNA	FIQ	SLK
Z_{SPA-1}	LSV	AGR	EIV	TLP	NLN	DPQ	KKA	FIF	SLW

Table 1: Amino acids 9 to 35 for Z_{SPA-1} and the wild-type Z domain.

- The collapse transition is much more cooperative for this sequence than for random sequences with the same composition. Moreover, chain collapse and helix formation occur at approximately the same temperature.

The relative order of chain collapse and helix formation depends strongly on the relative strength of the hydrogen bonds and the hydrophobic attraction, so the last conclusion may seem somewhat arbitrary. However, the chain does not fold to a helical bundle if the hydrogen bonds are too strong, and it does not fold in a cooperative manner if the hydrogen bonds are too weak [20]. As a result, with our ansatz for the potential, there is not much freedom left in the choice of these parameters, if the chain is to fold to a compact helical bundle in a cooperative manner.

In the present study, we apply the same model, with unchanged parameters, to Z_{SPA-1} and the Z domain of protein A. Following previous calculations for the B domain [9,10,12–17], we consider the 9–54-amino acid fragments of these two sequences (corresponding to the 10–55-amino acid fragment of the B domain). It should be mentioned that we also performed calculations for the 4–54-amino acid fragments of Z_{SPA-1} and the Z domain, with similar results.

The amino acid sequences of Z_{SPA-1} and the Z domain differ at 13 positions, all of which are located in the section 9–35. Table I shows this part of the sequences.

2.2 Numerical Methods

To simulate the thermodynamic behavior of this model, we use simulated tempering [23, 24], in which the temperature is a dynamic variable. Details on our implementation of this method can be found elsewhere [25]. For a review of simulated tempering and other generalized-ensemble techniques, see [26].

In conformation space we use two different elementary moves: first, the pivot move in which a single torsion angle is turned; and second, a semi-local method [27] that works with seven or eight adjacent torsion angles, which are turned in a coordinated manner.

The non-local pivot move is included in our calculations in order to accelerate the evolution of the system at high temperature, whereas the semi-local method improves the performance at low temperature.

Our simulations are started from random configurations. All statistical errors quoted are 1σ errors obtained by analyzing data from eight independent runs.

The temperatures studied range from $0.87\,T_{\rm m}$ to $1.43\,T_{\rm m}$, $T_{\rm m}$ being the melting temperature for the wild-type Z domain. The experimental value of this temperature is $T_{\rm m} = 75°{\rm C}$ [4]. Hence, the lowest and highest temperatures studied correspond to 31°C and 225°C, respectively. In the dimensionless energy unit used in our earlier study [17], $T_{\rm m}$ is given by $kT_{\rm m} = 0.630 \pm 0.001$, k being Boltzmann's constant. In the model we define $T_{\rm m}$ as the maximum of the specific heat.

3 Results and Discussion

Using the model and methods described in the previous section, we study the 9–54-amino acid fragments of $\rm Z_{SPA-1}$ and the wild-type Z domain. The latter sequence differs only by a one-point mutation from the previously studied 10–55-amino acid fragment of the B domain. Our results for the Z domain are similar to those for the B domain [17] summarized in the previous section. Figure 1 shows the free energies $F(\Delta, E)$ and $F(\Delta, E_{\rm hb})$ for the Z domain at $T = 0.87\,T_{\rm m}$, where Δ denotes RMSD from the NMR structure [5] (PDB code 2SPZ, model 1). Two major minima can be seen, with similar hydrogen-bond energies. These minima correspond to the two possible three-helix-bundle topologies. Both topologies are significantly populated, but the average total energy is slightly lower for the native topology, and this topology is the thermodynamically favored one. We also performed an energy minimization for the native topology, by applying simulated annealing combined with a conjugate gradient method to a large number of low-temperature conformations. The minimum-energy structure obtained this way is schematically illustrated in Fig. 2. It has an RMSD of $\Delta = 1.7\,$Å from the NMR structure. The corresponding result for the B domain was, as mentioned earlier, $\Delta = 1.8\,$Å.

Let us now compare the behavior of the Z domain with that of the engineered $\rm Z_{SPA-1}$ sequence. By CD, Wahlberg *et al.* [4] found $\rm Z_{SPA-1}$ to be less helical than the wild-type Z domain, the mean residue ellipticity for $\rm Z_{SPA-1}$ being 60% of that for the wild-type sequence. Furthermore, they found that the helix formation sets in at a lower temperature and is less cooperative for the engineered sequence. Figure 3a shows the

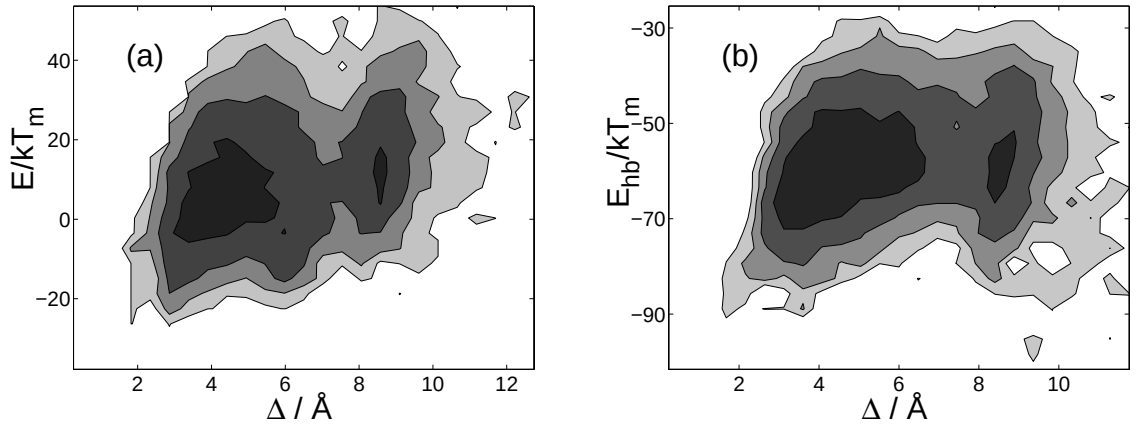

Figure 1: Level diagrams showing the free energies (a) $F(\Delta, E)$ and (b) $F(\Delta, E_{\rm hb})$ for the 9–54-amino acid fragment of the Z domain at $T = 0.87\,T_{\rm m}$. E is the total energy [see equation (1)], $E_{\rm hb}$ is the hydrogen-bond energy, and Δ denotes RMSD from the NMR structure. The separation between adjacent contour lines is $1\,kT$. The darkest regions correspond to $4 < F/kT < 5$ and the white regions to $F/kT > 8$.

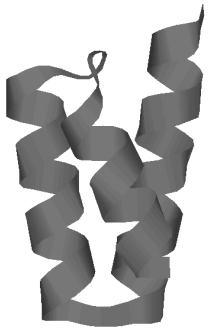

Figure 2: Schematic illustration of the structure obtained by energy minimization restricted to the thermodynamically favored topology (see text) for the 9–54-amino acid fragment of the Z domain.

helix content against temperature in our model, for both sequences.[‡] In agreement with the experimental results, we find that $\rm Z_{SPA-1}$ has a lower helix content, and that the helix formation is shifted toward lower temperature for this sequence. Figure 3b shows the temperature dependence of the radius of gyration. We find that $\rm Z_{SPA-1}$

[‡]We define helix content in the following way. Each amino acid, except the two at the ends, is labeled h if $-90° < \phi < -30°$ and $-77° < \psi < -17°$, and c otherwise. The two amino acids at the ends are labeled c. An amino acid is said to be helical if both the amino acid itself and its nearest neighbors are labeled h. The total number of helical amino acids is denoted by $N_{\rm h}$. The maximum value of $N_{\rm h}$ is $N - 4$ for a chain with N amino acids.

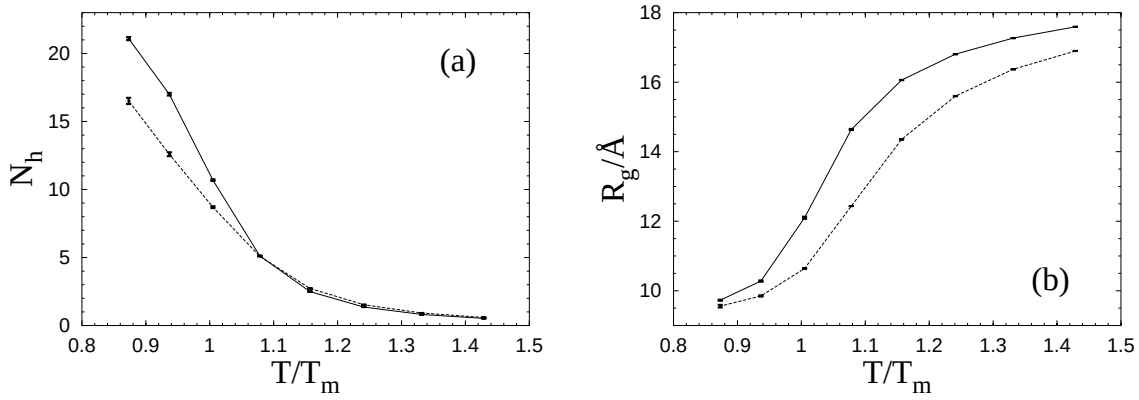

Figure 3: Helix formation and chain collapse for the 46-amino acid fragments of Z_{SPA-1} (dashed line) and the Z domain (full line). (a) The number of helical amino acids, N_h, against temperature. (b) The radius of gyration (calculated over all backbone atoms), R_g, against temperature. T_m denotes the melting temperature for the Z domain. The NMR structure for the Z domain has $N_h = 29$ and $R_g = 9.0$ Å.

is more compact than the Z domain. A comparison with Fig. 3a shows that, in our model, chain collapse occurs before helix formation for Z_{SPA-1}. From Fig. 3 it can also be seen that the melting behavior is less cooperative for Z_{SPA-1} than for the Z domain. This conclusion is supported by our data for the specific heat (not shown); the peak in the specific heat is more pronounced for the Z domain than for Z_{SPA-1}.

That the model predicts Z_{SPA-1} to be more compact than the Z domain is not surprising, given that the number of hydrophobic amino acids is larger for Z_{SPA-1} (14) than for the Z domain (11). In addition, Z_{SPA-1} has one more Pro than the wild-type sequence, which does change the local properties of the chain and could affect the overall size, too. It should be pointed out that the effect of a Pro on the overall size may be poorly described by the model because all peptide bonds, also those preceding a Pro, are held fixed (trans).

The reduced helix content of Z_{SPA-1} shows that this sequence does not make a perfect three-helix bundle, but does not tell how its structure differs from a three-helix bundle. It could be that one of the three helices is missing and that the other two are still there, but it could also be that the disorder is more uniform along the chain, so that all three helices are present but partially disordered. The NMR analysis of Z_{SPA-1} [4] does not exclude any of these two possibilities.

Figure 4 shows how the helix content varies along the chains in our model. A com-

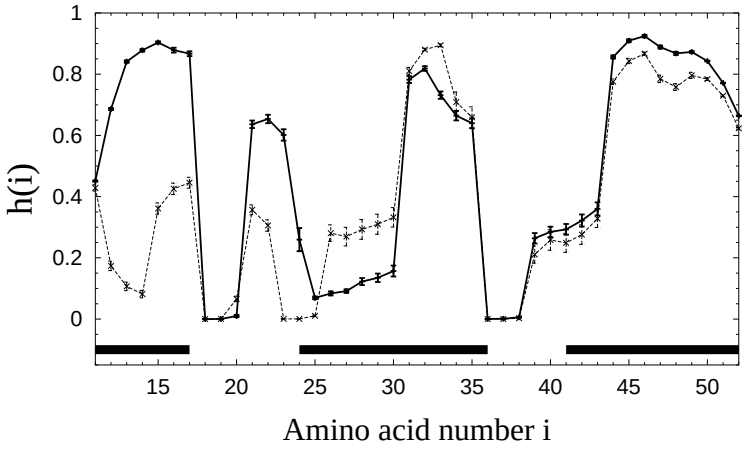

Figure 4: Helix content along the chain, $h(i)$, for the 46-amino acid fragments of Z_{SPA-1} (dashed line) and the Z domain (full line) at $T = 0.87\,T_m$, where T_m is the melting temperature for the Z domain. $h(i)$ denotes the probability that amino acid i is helical (for the definition of helical, see footnote). Thick horizontal lines indicate helical parts of the NMR structure [5] for the Z domain.

parison with experimental data for the Z domain [5] shows that the first half of helix II is somewhat distorted in the model. As a result, it is possible that the model underestimates the structural change produced by the mutation Glu25Pro (see Table 1), which should have a helix-breaking effect. Our results for helices I and III of the Z domain are, by contrast, in good agreement with experimental data. These two helices respond very differently to the mutations leading to Z_{SPA-1}. Our results suggest that helix III, which itself is free from mutations, remains stable in Z_{SPA-1}, whereas helix I, which contains seven mutations (see Table 1), turns unstable. Two possible explanations why the model predicts helix I to become unstable are that the hydrophobicity pattern of helix I is less helical in Z_{SPA-1}, and that one of the mutations, Phe13Gly, increases the flexibility of this part of the chain.

To further investigate how the mutations affect different parts of the chain, we also perform an RMSD-based analysis. For each conformation, we compute two RMSD values, Δ_1 and Δ_2, for the first and second halves of the chain, respectively. The two parts of the chain are separately superimposed on the NMR structure. Figure 5 shows the probability distributions of Δ_1 and Δ_2 both for Z_{SPA-1} and the Z domain. In line with the results in Fig. 4, we find that the two Δ_2 distributions are similar, although the distribution for Z_{SPA-1} is slightly wider. The two Δ_1 distributions differ, by contrast, markedly, the mean being significantly higher for Z_{SPA-1} than for the wild-type sequence. These results confirm that, in our model, the disorder of Z_{SPA-1}

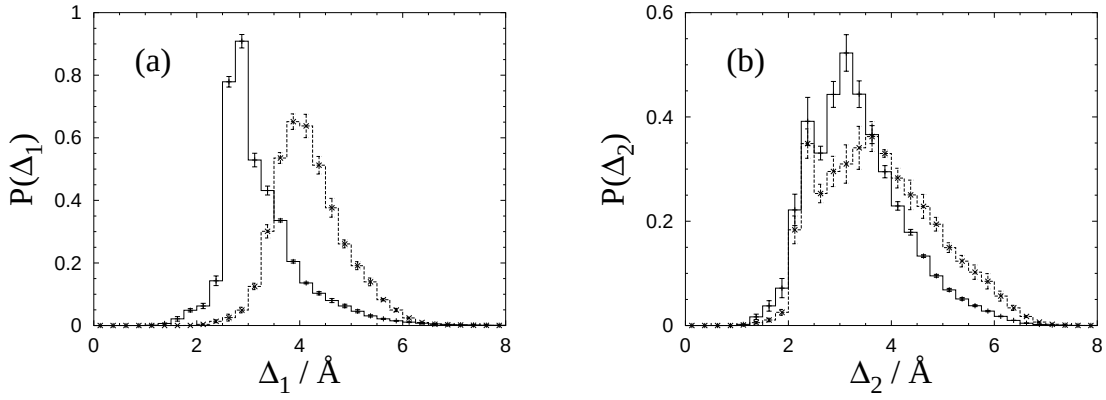

Figure 5: RMSD distributions for the 46-amino acid fragments of Z_{SPA-1} (dashed line) and the Z domain (full line). (a) The distribution of Δ_1 (amino acids 9–31). (b) The distribution of Δ_2 (amino acids 32-54). Both Δ_1 and Δ_2 are backbone RMSDs. The temperature is the same as in Fig. 4.

is not uniformly distributed along the chain; the main difference between Z_{SPA-1} and the Z domain lies in the behavior of the first half of the chain.

4 Conclusion

Using a model that combines a relatively detailed chain representation with a simple interaction potential, we have studied the thermodynamic behaviors of an engineered sequence and its parent. The model is sequence-based, which makes it possible to compare the two sequences in a straightforward manner. Despite the simplicity of the potential, we found that the model is able to capture important effects of the mutations; the mutated sequence, Z_{SPA-1}, shows a reduced helix content and a melting behavior that is less cooperative, compared with the wild-type sequence. We also found that chain collapse occurs before helix formation sets in for Z_{SPA-1}, and that the main difference between the two sequences lies in the behavior of the first half of the chain, which is less stable for Z_{SPA-1}. To decide whether or not these two predictions are correct requires further experimental data.

Acknowledgments

We thank Torleif Härd for a helpful discussion. This work was in part supported by the Swedish Research Council.

References

[1] Wright PE, Dyson HJ. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 1999; 293: 321–331.

[2] Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* 2002; 12: 54–60.

[3] Eklund M, Axelsson L, Uhlén M, Nygren P-Å. Anti-idiotypic protein domains selected from protein A-based affibody libraries. *Proteins* 2002; 48: 454–462.

[4] Wahlberg E, Lendel C, Helgstrand M, Allard P, Dincbas-Renqvist V, Hedqvist A, Berglund H, Nygren P-Å, Härd T. An affibody in complex with a target protein: Structure and coupled folding. *Proc. Natl. Acad. Sci. USA* 2003; 100: 3185–3190.

[5] Tashiro M, Tejero R, Zimmerman DE, Celda B, Nilsson B, Montelione GT. High-resolution solution NMR structure of the Z domain of staphylococcal protein A. *J. Mol. Biol.* 1997; 272: 573–590.

[6] Högbom M, Eklund M, Nygren P-Å, Nordlund P. Structural basis for recognition by an *in vitro* evolved affibody. *Proc. Natl. Acad. Sci. USA* 2003; 100: 3191–3196.

[7] Bai Y, Karimi A, Dyson HJ, Wright PE. Absence of a stable intermediate on the folding pathway of protein A. *Protein Sci.* 1997; 6: 1449–1457.

[8] Myers JK, Oas TG. Preorganized secondary structure as an important determinant of fast protein folding. *Nat. Struct. Biol.* 2001; 8: 552–558.

[9] Boczko EM, Brooks CL III. First-principles calculation of the folding free energy of a three-helix bundle protein. *Science* 1995; 269: 393–396.

[10] Guo Z, Brooks CL III, Boczko EM. Exploring the folding free energy surface of a three-helix bundle protein. *Proc. Natl. Acad. Sci. USA* 1997; 94: 10161–10166.

[11] Kussell EL, Shimada J, Shakhnovich EI. A structure-based method for derivation of all-atom potentials for protein folding. *Proc. Natl. Acad. Sci. USA* 2002; 99: 5343–5348.

[12] Linhananta A, Zhou Y. The role of sidechain packing and native contact interactions in folding: Discrete molecular dynamics folding simulations of an all-atom Gō model of fragment B of staphylococcal protein A. *J. Chem. Phys.* 2002; 117: 8983–8995.

[13] Kolinski A, Galazka W, Skolnick J. Monte Carlo studies of the thermodynamics and kinetics of reduced protein models: Application to small helical, β, and α/β proteins. *J. Chem. Phys.* 1998; 108: 2608–2617.

[14] Zhou Y, Karplus M. Interpreting the folding kinetics of helical proteins. *Nature* 1999; 401: 400–403.

[15] Shea J-E, Onuchic JN, Brooks CL III. Exploring the origins of topological frustration: Design of a minimally frustrated model of fragment B of protein A. *Proc. Natl. Acad. Sci. USA* 1999; 96: 12512–12517.

[16] Berriz GF, Shakhnovich EI. Characterization of the folding kinetics of a three-helix bundle protein via a minimalist Langevin model. *J. Mol. Biol.* 2001; 310: 673–685.

[17] Favrin G, Irbäck A, Wallin S. Folding of a small helical protein using hydrogen bonds and hydrophobicity forces. *Proteins* 2002; 47: 99–105.

[18] Gō N, Abe H. Noninteracting local-structure model of folding and unfolding transition in globular proteins. *Biopolymers* 1981; 20: 991–1011.

[19] Irbäck A, Sjunnesson F, Wallin S. Three-helix-bundle protein in a Ramachandran model. *Proc. Natl. Acad. Sci. USA* 2000; 97: 13614–13618.

[20] Irbäck A, Sjunnesson F, Wallin S. Hydrogen bonds, hydrophobicity forces and the character of the collapse transition. *J. Biol. Phys.* 2001; 27: 169–179.

[21] Favrin G, Irbäck A, Samulesson B, Wallin S. Two-state folding over a weak free-energy barrier. Lund Preprint LU TP 03-07, to appear in *Biophys. J.*

[22] Gouda H, Torigoe H, Saito A, Sato M, Arata Y, Shimada I. Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures. *Biochemistry* 1992; 31: 9665–9672.

[23] Lyubartsev AP, Martsinovski AA, Shevkunov SV, Vorontsov-Velyaminov PV. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.* 1992; 96: 1776–1783.

[24] Marinari E, Parisi G. Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* 1992; 19: 451–458.

[25] Irbäck A, Potthast F. Studies of an off-lattice model for protein folding: Sequence dependence and improved sampling at finite temperature. *J. Chem. Phys.* 1995; 103: 10298–10305.

[26] Hansmann UHE, Okamoto Y. The generalized-ensemble approach for protein folding simulations. In: Stauffer D, editor. Annual reviews of computational physics VI. Singapore: World Scientific; 1999. p 129–157.

[27] Favrin G, Irbäck A, Sjunnesson F. Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space. *J. Chem. Phys.* 2001; 114: 8154–8158.