Scientific paper

Molecular Simulations Find Stable Structures in Fragments of Protein G

Tjaša Urbič,[1] Tomaž Urbič,[2] Franc Avbelj[1] and Ken A. Dill[3]

[1] *National Institute of Chemistry Slovenia, Ljubljana, Slovenia*

[2] *Faculty of Chemistry and Chemical Technology, University of Ljubljana, Slovenia*

[3] *Department of Pharmaceutical Chemistry, University of California, San Francisco*

* *Corresponding author: E-mail: tomaz.urbic@fkkt.uni-lj.si*

Received: 26-01-2008

Abstract

We perform all-atom computer simulations on nearly one hundred 6-, 8-, 10-, and 12-mer peptide fragments of protein G, and look for stable states. We simulated by replica-exchange molecular dynamics using Amber7 with the parm96 force-field and a GB/SA (generalized-Born/solvent accessible) implicit solvent model. We find that useful diagnostics for identifying stable converged structures are the conformational entropy and free energy of each state. A large gap in the ground-state free-energy, and a low entropy indicate convergence to a single preferred peptide conformation. We find that a non-negligible fraction of such structures have some native-like character. Such physics-based modeling may be useful for identifying early nuclei in folding kinetics and for assisting in protein-structure prediction methods that utilize the assembly of peptide fragments.

1. Introduction

A long-standing goal of computational biology has been to devise a computer algorithm that takes, as input, an amino acid sequence and gives, as output, the three-dimensional native structure of a protein. A main motivation is to make drug discovery faster and more efficient by replacing slow expensive structural biology experiments with fast cheap computer simulations. There are many successful bioinformatics approaches to this problem.[1–5] Those methods draw heavily from knowledge bases of known protein native structures. Our interest here is in whether purely physical all-atom force field models are capable of identifying native-like starting points from small peptide fragments.

Short peptide fragments of proteins often have intrinsic propensities for the formation of their native conformations. NMR experiments[6] show that long peptide fragments have native-like conformations.[7–11] Some short peptides in solution have also been shown to adopt their native secondary structures: α helices[12, 13] and β hairpins.[14–18] As a consequence, peptide conformational propensities that are taken from the protein databank (PDB) are now widely used in protein-structure prediction algorithms. Most current protein structure prediction methods make some use of database-derived conformational preferences. A popular set of peptide fragment conformations is the I-sites library of David Baker and his coworkers[19, 20] From the recent CASP (Critical Assessment of Techniques for Structure Prediction) protein structure prediction competition, it was noted that most of the successful de novo methods start with small fragments[21, 22] which are then combined together into a predicted tertiary structure.

However, it would be desirable to achieve high-resolution protein structure prediction in models that are purely physics-based, for various reasons. Such predictions would not rely on information contained in protein structure databases. First, it would put our understanding of protein structures and driving forces on a deeper and more physical foundation. For example, such methods could elucidate the physical routes of protein folding. Second, it would allow the prediction of non-native states, too, those that are of interest for protein folding kinetics and stability. Third, it would allow us to treat induced fit binding of ligands, or other conformational changes.

There is good evidence that physical models can capture these conformational propensities of peptides. Simple physical models can reproduce the structural biases of certain peptide fragments.[23–26] To date, however, such studies have largely focused on selected peptides that are expected to fold. Moreover, several models of protein folding kinetics are premised on the idea that folding routes begin with metastable structures of small peptide fragments.[27–30] In recent work Ho et al[31] simulated 133 peptide 8-mer fragments from six different proteins. In that work, it was found that more than 30 percent of the peptide fragments converge to a preferred structure, some of which are native-like. In the present study, we extend beyond that work, through a systematic exploration of different fragment chain lengths

The present study provides a test of physics-based all-atom simulations – the quality of the force field and solvation model, and the adequacy of current typical levels of sampling. There are well known problems with commonly used molecular mechanics force fields. Yoda et al.[32] conducted multicanonical simulations of several small peptides (the α-helical C-peptide of ribonuclease A and the C-terminal β-hairpin of protein G) by using six common force fields (AMBER94, AMBER96, AMBER99, CHARMM22, OPLS/AA/L, and GROM0S96) in explicit solvent and concluded that all of these force fields have different propensities to form secondary structures because of the differences in backbone torsional energies. Also, it is still largely an open question as to how well these implicit solvent models can predict the thermodynamics as well as the kinetics of protein folding. Zhou[33] tested different combinations of force fields and solvation models on the C-terminal β-hairpin of protein G. He found the best balanced combination to be AMBER96/GBSA, so we use that here.

Here, we study 26 6-mer peptide fragments from protein G, 25 8-mers, 24 10-mers and 23 12-mers. We use replica-exchange molecular dynamics sampling[34] in Amber7[35], with the parm96 parameters[36] and the GB/SA implicit solvent model of Tsui and Case[37]. We are interested in whether this physical model can identify native-like secondary structures in peptide fragments. We systematically generate a series of peptide fragments with overlapping sequences from the original protein sequence. Neighboring fragments have a 4–10 residue overlap (and two-residue gap). We simulate each peptide using 16 replicas for 5 ns/replica.

2. Computational Details

2. 1. Simulation Details

We utilized replica-exchange molecular dynamics (REMD)[34] using a custom PERL script wrapper (http://www.dillgroup.ucsf.edu/~jchodera/code/rex) around the SANDER molecular dynamics program for the Amber7 molecular-modeling package.[35] REMD periodically attempts to exchange conformations between independent molecular dynamics simulations running in parallel at different temperatures, based on a Metropolis-like criterion. This allows individual replicas to heat up to overcome barriers and then cool back down to temperatures of interests. REMD has two advantages. It explores more conformational space then conventional molecular dynamics techniques.[38] And, REMD samples from the canonical ensemble at each temperature. This sampling gives proper estimates of free energies, not just energies. We used 16 replicas exponentially spaced between 270 *K* and 690 *K*. The probability of exchange-acceptance was approximately 50%. Exchanges were attempted every 1 *ps*. Between exchange attempts energy-conserving molecular dynamics was used with a 2 *fs* time step. After each exchange attempt the velocities were randomized from the appropriate Maxwell-Boltzmann distributions to ensure sampling from the canonical ensemble at the appropriate temperature.

The protein is chopped into overlapping fragments 6 to 12 residues in length. Fragments were modeled with the Amber Parm96 force field[36] with GB implicit solvent model and surface area penalty term of 5 $cal \cdot mol^{-1} \cdot Å^{-2}$ of Tsui and Case.[37] All fragments were capped at N and C termini with acetyl and *N*-methylamine blocking groups to avoid undue influence from the zwitterionic termini. The fragments were initialized in extended state. The bonds to hydrogens were constrained with the SHAKE.[39] Simulations were run for 5 *ns* per replica. Configurations were stored every 1 *ps* for analysis.

2. 2. Thermodynamic Properties

The results were analyzed by the weighted histogram analysis (WHAM).[40, 41] In order to extract thermodynamic observables for a target temperature T we must reweight the configurations taken from each temperature T_k in order to combine them into a representative ensemble.[41]

We first calculate the dimensionless free energy f_k for each replica k. We start with an approximate value for f_k and calculate the density of states Ω_{kE} with energy E in replica k as

$$\Omega_{kE} = \frac{N_{kE}}{\sum_{l=1}^{K} N_{kl} \exp\left(f_l - \beta_l E\right)}, \qquad (1)$$

where N_{kE} is number of sampled configurations with energy E from replica k. $\beta_l = \frac{1}{k_B T_l}$ inverse temperature, k_B Boltzmann constant and N_{kl} number of configurations of replica k at temperature T_l. From the updated density of states we can calculate an updated estimate of dimensionless free energy by

$$f_k = -\ln\left[\sum_{E} \Omega_{kE} \exp\left(\beta_k E\right)\right]. \qquad (2)$$

We iterate equations (1) and (2) until the free energy converges. Then we can use the these data to reweight the relative free energy profile F of state i to the inverse target temperature or we can calculate the estimator of the expectation for any observable.[41]

2. 3. Mesostates

To analyze our date, we form discrete bins of the backbone degrees of freedom[31]. This process has a long history, dating back to the original work of Ramachandran et al,[42] who divided the backbone $\phi - \psi$ angles into three district regions, which are known as α, α_L and β. We describe the conformation of the peptide backbone as a string of discrete *mesostates* that we call a *mesostring*. Each mesostring describes a state that is separated by an energy barrier from other mesostrings. This means that each mesostring corresponds to local minima in the conformation space of the peptide backbone. When we know all mesostrings it is easy to find the one with lowest energy and get structure from the lowest energy basin. This partitioning in the terms of discrete regions in the backbone angles has been observed in a molecular dynamics simulation of an α-helical peptide.[43] For mesostring analysis we cannot use the database analysis to define the boundaries of the backbone mesostates because current force fields cannot replicate the database distribution of $\phi - \psi$ angles.[31] Ho and coworkers[31] define the boundaries of the backbone mesostates in the terms of the same force field we use here. They ran replica-exchange simulations of the alanine dipeptide and the glycine dipeptide to define the boundaries of different mesostates. They break up the Ramachandran plot in terms of the following mesostates:

$$[b] : (-180° < \phi < 0°, 45° < \phi < 180°)$$
$$U(-180° < \phi < 0°, -180° < \phi < -135°)$$
$$U(120° < \phi < 180°, 45° < \phi < 180°)$$
$$U(120° < \phi < 180°, -180° < \phi < -135°)$$

$$[a] : (-180° < \phi < 0°, -135° < \phi < 45°)$$
$$U(120° < \phi < 180°, -135° < \phi < 45°)$$

$$[l] : (0° < \phi < 120°, -180° < \phi < 180°)$$
$$U(120° < \phi < 180°, -135° < \phi < 45°)$$

and for glycin

$$[b] : (-180° < \phi < 0°, 45° < \phi < 180°)$$
$$U(-180° < \phi < 0°, -180° < \phi < -135°)$$
$$U(0° < \phi < 180°, 135° < \phi < 180°)$$
$$U(0° < \phi < 180°, -180° < \phi < -45°)$$

$$[a] : (-180° < \phi < 0°, -135° < \phi < 45°)$$

$$[l] : (0° < \phi < 180°, -45° < \phi < 135°)$$

With the use of dimensionless free energies f_k (Eq. (2)) we reweight the free energy profile F_i of mesostring i at the inverse target temperature β as

$$F_x(\beta) = -\frac{1}{\beta} \ln \left\{ \sum_E \left[\frac{\sum_{k=1}^{K} N_{ikE} \exp(\beta E)}{\sum_{l=1}^{K} N_{ilE} \exp(f_l - \beta_l E)} \right] \right\}. \quad (3)$$

After using WHAM to calculate the relative free energies F_i of the mesostrings i we calculate the probabilities p_i of this mesostring by

$$p_i = \frac{\exp(-\beta F_i)}{\sum_j \exp(-\beta F_j)} \quad (4)$$

When each simulation is completed, each fragment will have different populations of the various mesostrings and also different free energies. The ground mesostring is the mesostate with the highest population and the lowest free energy. We then determine whether each peptide finds a metastable structure. We define the existence of the structure in the fragment in terms of two properties of mesostring. First, we use the probability p_i of the ground mesostring. Second, we use the free energy gap ΔF between the ground mesostring and the next mesostring. When the ground mesostring is nearly identical to next mesostring, it indicates a lack of preference of the force field for a particular structure. If the most populated (ground mesostring) differs by only one mesostate, we group them into a consensus mesostring, which contains one indefinite mesostate signified by [–]. When we group similar mesostrings into consensus mesostring we calculate the free energy difference to another mesostring j by

$$\Delta F = -\frac{1}{\beta} \ln \left(\frac{p_{consensus}}{p_j} \right) \quad (5)$$

The backbone entropy is calculated using the Boltzmann formula

$$S = -k_B \sum_i p_i \ln p_i. \quad (6)$$

The backbone entropy is useful for measuring for a given fragment the sharpness of the distribution of probabilities of the mesostring. The more peaked the distribution is and, thus the more favored the mesostring is, the lower is the backone entropy. In this way, the backbone entropy indicates whether any one conformation is substantially favored over the others for given fragment. The fragments having low mesostring entropy can be considered as early folding nuclei to initiate folding.

2. 4. Contact Maps

The three-dimensional structure of a protein may be compactly represented as a symmetrical, square matrix of pairwise, inter-residue contacts, called the contact map.[44,45] The contact map provides useful information about the protein's secondary structure and it also captures non-local interactions giving clues to its tertiary structure.

We define that two residues x_i and x_j in a protein are in a contact if the distance d_{ij} between α-carbon atoms of the residues x_i and x_j is lower then some threshold value.

In our case we chose 7 Å. Distance d_{ij} is defined as

$$d_{ij} = |\mathbf{r_i} - \mathbf{r_j}|, \qquad (7)$$

where r_i and r_j are the coordinates of α-carbon atoms. A contact map for protein with N residues is an $N \times N$ binary matrix S in which the elements S_{ij} are defined as

$$S_{ij} = \begin{cases} 1 & d_{ij} < 7\mathring{A}, |i-j| > 3 \\ 0 & Otherwise \end{cases} \qquad (8)$$

Note that we require a minimum sequence distance of 4 to call it a contact.

3 Results and Discussion

In this study, we chopped protein G in 26 peptide 6-mer fragments, 25 8-mers, 24 10-mers and 23 12-mers. We systematically generate a series of peptide fragments with overlapping sequences from the original protein sequence. Neighboring fragments have a 4–10 residue overlap (and a two-residue gap). Table 1 shows the sequences of the fragments. We simulated each peptide using 16 replicas for 5 *ns* per replica. We clustered our conformations analyzed them using mesostring analysis. We looked for structural biases in each peptide. We used two criteria. First, we use the free energy gap ΔF between the ground mesostring and the next mesostring with higher free energy. Second, we use the probability p_1 to determine the population of the ground mesostring. We identified a stable structure if $\Delta F > 0.6$ *kcal/mol* and $p_1 > 38\%$, which is slightly differently then was used previously by Ho et al.[31] We used a 1 *ns* window for analyzing the mesostrings. Tables 2–5

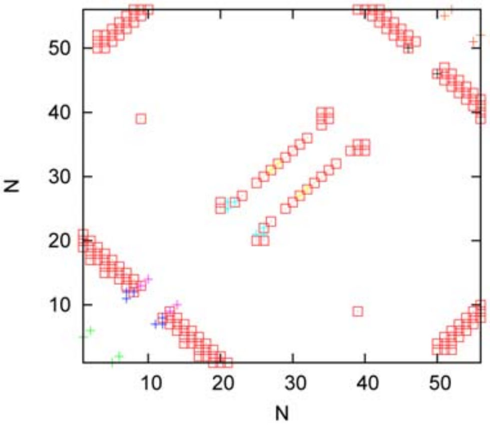


Figure 1: Contact map for different 6-mers. Contact map of native protein G is shown with diamonds, for different structured 6-mers with crosses.

Table 1: Sequences of 6, 8, 10 and 12-mers fragments used in calculations.

Fragment	Sequence of 6-mers	Sequence of 8-mers	Sequence of 10-mers	Sequence of 12-mers
frag1	1-MTYKLI	1-MTYKLILN	1-MTYKLILNGK	1-MTYKLILNGKTL
frag2	3-YKLILN	3-YKLILNGK	3-YKLILNGKTL	3-YKLILNGKTLKG
frag3	5-LILNGK	5-LILNGKTL	5-LILNGKTLKG	5-LILNGKTLKGET
frag4	7-LNGKTL	7-LNGKTLKG	7-LNGKTLKGET	7-LNGKTLKGETTT
frag5	9-GKTLKG	9-GKTLKGET	9-GKTLKGETTT	9-GKTLKGETTTEA
frag6	11-TLKGET	11-TLKGETTT	11-TLKGETTTEA	11-TLKGETTTEAVD
frag7	13-KGETTT	13-KGETTTEA	13-KGETTTEAVD	13-KGETTTEAVDAA
frag8	15-ETTTEA	15-ETTTEAVD	15-ETTTEAVDAA	15-ETTTEAVDAATA
frag9	17-TTEAVD	17-TTEAVDAA	17-TTEAVDAATA	17-TTEAVDAATAEK
frag10	19-EAVDAA	19-EAVDAATA	19-EAVDAATAEK	19-EAVDAATAEKVF
frag11	21-VDAATA	21-VDAATAEK	21-VDAATAEKVF	21-VDAATAEKVFKQ
frag12	23-AATAEK	23-AATAEKVF	23-AATAEKVFKQ	23-AATAEKVFKQYA
frag13	25-TAEKVF	25-TAEKVFKQ	25-TAEKVFKQYA	25-TAEKVFKQYAND
frag14	27-EKVFKQ	27-EKVFKQYA	27-EKVFKQYAND	27-EKVFKQYANDNG
frag15	29-VFKQYA	29-VFKQYAND	29-VFKQYANDNG	29-VFKQYANDNGVD
frag16	31-KQYAND	31-KQYANDNG	31-KQYANDNGVD	31-KQYANDNGVDGE
frag17	33-YANDNG	33-YANDNGVD	33-YANDNGVDGE	33-YANDNGVDGEWT
frag18	35-NDNGVD	35-NDNGVDGE	35-NDNGVDGEWT	35-NDNGVDGEWTYD
frag19	37-NGVDGE	37-NGVDGEWT	37-NGVDGEWTYD	37-NGVDGEWTYDDA
frag20	39-VDGEWT	39-VDGEWTYD	39-VDGEWTYDDA	39-VDGEWTYDDATK
frag21	41-GEWTYD	41-GEWTYDDA	41-GEWTYDDATK	41-GEWTYDDATKTF
frag22	43-WTYDDA	43-WTYDDATK	43-WTYDDATKTF	43-WTYDDATKTFTV
frag23	45-YDDATK	45-YDDATKTF	45-YDDATKTFTV	45-YDDATKTFTVTE
frag24	47-DATKTF	47-DATKTFTV	47-DATKTFTVTE	
frag25	49-TKTFTV	49-TKTFTVTE		
frag26	51-TFTVTE			

.

Table 2: Native and ground mesostring (see the text for definition on state a, b, and l), probability of ground mesostring, free energy difference, entropy and RMSD of 6-mers in 1*ns* time window after 4 *ns* equalibration. Bold indicates stable fragments.

Fragment	*RMSD*	Native	Ground	p_1	ΔF	*TS*
	Å	mesostring	mesostring	%	kcal/mol	kcal/mol
frag1	**3.7**	**bbbbbb**	**-aaaaa**	**50**	**0.71**	**1.20**
frag2	3.9	bbbbba	baaaaa	42	0.37	1.01
frag3	4.1	bbbaba	-aaabb	37	0.82	1.51
frag4	**2.6**	**babaab**	**bb-aab**	**64**	**0.65**	**0.92**
frag5	**2.6**	**baabbb**	**baaaa-**	**49**	**0.92**	**1.38**
frag6	4.1	abbbbb	bbb-aa	20	0.39	1.87
frag7	4.3	bbbbbb	blaabb	20	0.28	1.56
frag8	3.6	bbbbbb	baaaaa	10	0.00	1.82
frag9	3.5	bbbbab	aaaaa-	36	0.43	1.41
frag10	3.1	bbabaa	aaaaa-	34	0.56	1.58
frag11	**0.5**	**abaaaa**	**bbaaa-**	**51**	**1.10**	**1.34**
frag12	3.0	aaaaaa	bbaaa-	23	0.45	1.63
frag13	3.1	aaaaaa	b-aabb	25	0.68	1.67
frag14	**0.4**	**aaaaaa**	**-aaaaa**	**73**	**1.21**	**0.80**
frag15	3.1	aaaaaa	aaaaaa	23	0.43	1.59
frag16	3.5	aaaaaa	babaab	15	0.06	1.53
frag17	3.3	aaaaal	bbbbbb	18	0.40	1.75
frag18	1.5	aaalbb	ba-bbb	33	0.97	1.70
frag19	4.2	albbab	bbbbbb	13	0.03	1.85
frag20	3.0	bbabbb	bbbabb	16	0.01	1.62
frag21	4.7	abbbbb	baaaab	36	0.99	1.56
frag22	1.9	bbbbaa	baaaaa	11	0.14	1.68
frag23	**2.1**	**bbaaal**	**ba-aaa**	**59**	**0.87**	**1.06**
frag24	3.0	aaalbb	baaabb	18	0.18	1.62
frag25	3.8	albbbb	aaaaab	24	0.20	1.45
frag26	**3.9**	**bbbbbb**	**aaaaa-**	**60**	**1.41**	**1.20**

Table 3: Same as table 2, but for 8-mers.

Fragment	*RMSD*	Native	Ground	p_1	ΔF	*TS*
	Å	mesostring	mesostring	%	kcal/mol	kcal/mol
frag1	4.6	bbbbbbba	aaaaaaaa	30	0.43	1.34
frag2	5.5	bbbbbaba	baaaaabb	21	0.01	1.51
frag3	3.8	bbbabaab	bbbbbaab	28	0.31	1.38
frag4	**3.0**	**babaabbb**	**bb-aabbb**	**57**	**1.00**	**1.25**
frag5	4.6	baabbbbb	bblaabbb	11	0.00	1.69
frag6	6.1	abbbbbbb	bbaaabba	12	0.02	1.73
frag7	5.6	bbbbbbbb	bbaaaaaa	25	0.03	1.26
frag8	**4.3**	**bbbbbbab**	**-aaaaaaa**	**55**	**0.68**	**1.18**
frag9	3.8	bbbbabaa	abaaaabb	43	0.48	1.13
frag10	**2.5**	**bbabaaaa**	**b-baaaab**	**63**	**1.29**	**1.22**
frag11	4.0	abaaaaaa	bblaabbb	14	0.08	1.65
frag12	2.3	aaaaaaaa	baaaaaaa	20	0.17	1.48
frag13	4.3	aaaaaaaa	aaaaaaaa	30	0.41	1.34
frag14	3.6	aaaaaaaa	baaaaaaa	44	0.37	1.01
frag15	4.5	aaaaaaaa	lbaaaabb	35	0.47	1.23
frag16	3.7	aaaaaaal	bbaaabbb	21	0.27	1.71
frag17	3.4	aaaaalbb	bbbaabbb	28	0.42	1.46
frag18	3.4	aaalbbab	baababbb	12	0.11	2.08
frag19	4.4	albbabbb	bbaa-bbb	23	0.53	1.77
frag20	**4.4**	**bbabbbbb**	**babaaaab**	**41**	**0.68**	**1.35**
frag21	6.9	abbbbbaa	bbaaaabb	17	0.01	1.51
frag22	3.4	bbbbaaal	aaaaaaaa	30	0.10	1.24
frag23	**3.0**	**bbaaalbb**	**babaaa-b**	**78**	**1.10**	**0.73**
frag24	4.1	aaalbbbb	baaaaaaa	38	0.46	1.31
frag25	6.1	albbbbbb	aaababab	25	0.14	1.42

Table 4: Same as table 2, but for 10-mers.

Fragment	*RMSD*	Native	Ground	p_1	ΔF	*TS*
	Å	mesostring	mesostring	%	kcal/mol	kcal/mol
frag1	**6.2**	**bbbbbbbaba**	**aaaaaaaab-**	**43**	**0.82**	**1.43**
frag2	6.1	bbbbbabaab	bbababbbab	31	0.24	1.22
frag3	4.3	bbbabaabbb	aaabbaabbb	36	0.66	1.42
frag4	5.3	babaabbbbb	bblaaabbab	19	0.15	1.49
frag5	5.4	baabbbbbbb	baaaalbba-	21	0.44	1.93
frag6	7.7	abbbbbbbbb	babbbaaaaa	8	0.02	1.84
frag7	5.7	bbbbbbbbab	bbaaaaaaaa	13	0.10	1.67
frag8	4.6	bbbbbbabaa	baaaaaaaaa	30	0.23	1.30
frag9	4.7	bbbbabaaaa	bbaaaaaaaa	18	0.49	1.86
frag10	3.9	bbabaaaaaa	bbbaabbabb	26	0.39	1.57
frag11	4.3	abaaaaaaaa	aaaaaaaabb	27	0.56	1.61
frag12	3.0	aaaaaaaaaa	baaaaaabbb	35	0.42	1.23
frag13	5.3	aaaaaaaaaa	bbbblbbabb	30	0.38	1.37
frag14	**0.6**	**aaaaaaaaaa**	**aaaaaaaaaa**	**51**	**0.84**	**1.10**
frag15	5.3	aaaaaaaaal	bbabblabbb	18	0.06	1.57
frag16	3.7	aaaaaaalbb	bbbbaaalbb	19	0.37	1.61
frag17	3.3	aaaaalbbab	abaabbabbb	21	0.75	1.94
frag18	5.4	aaalbbabbb	bbbbaabbbb	16	0.15	1.94
frag19	**5.8**	**albbabbbbb**	**bbbbbabbbb**	**38**	**0.77**	**1.59**
frag20	**5.4**	**bbabbbbbaa**	**babaaaaaaa**	**60**	**0.87**	**0.95**
frag21	4.1	abbbbbaaal	baaaaaaaaa	41	0.35	1.14
frag22	3.2	bbbbaaalbb	bababaaaaa	37	0.37	1.14
frag23	3.8	bbaaalbbbb	baaaaaaaaa	39	0.50	1.29
frag24	5.6	aaalbbbbbb	baaaaaabba	27	0.06	1.30

Table 5: Same as table 2, but for 12-mers.

Fragment	*RMSD*	Native	Ground	p_1	ΔF	*TS*
	Å	mesostring	mesostring	%	kcal/mol	kcal/mol
frag1	8.4	bbbbbbbabaab	bababaaaabba	32	0.63	1.49
frag2	5.0	bbbbbabaabbb	baaabbbbabbb	30	0.32	1.49
frag3	4.3	bbbabaabbbbb	aaabbaabbbbb	18	0.37	1.76
frag4	**6.0**	**babaabbbbbbb**	**bblaaabbabbb**	**38**	**0.63**	**1.39**
frag5	7.3	baabbbbbbbbb	bbbbbbbbbbbb	17	0.03	1.89
frag6	**8.9**	**abbbbbbbbbab**	**bbb-bbaaabbb**	**42**	**0.68**	**1.45**
frag7	5.8	bbbbbbbbabaa	babaaaaabbbb	33	0.78	1.62
frag8	5.1	bbbbbbabaaaa	aaaaaaabbabb	33	0.50	1.38
frag9	4.7	bbbbabaaaaaa	baaaaaabbaab	23	0.13	1.57
frag10	4.4	bbabaaaaaaaa	babbaaaabbbb	42	0.13	0.91
frag11	5.8	abaaaaaaaaaa	bbbaaabbbbab	15	0.36	1.68
frag12	**5.7**	**aaaaaaaaaaaa**	**bbabaaababb-**	**42**	**0.66**	**1.42**
frag13	4.2	aaaaaaaaaaaa	baaaabaaabba	12	0.02	1.66
frag14	**5.6**	**aaaaaaaaaaal**	**bbb-aaabbbbb**	**38**	**0.68**	**1.52**
frag15	5.5	aaaaaaaaalbb	bbbbbaabbbbb	15	0.02	1.67
frag16	**4.1**	**aaaaaaalbbab**	**bbbaaaababb-**	**45**	**0.95**	**1.43**
frag17	5.6	aaaaalbbabbb	bbbabbbb-aab	22	0.75	2.02
frag18	6.3	aaalbbabbbbb	babbbbbbbaba	19	0.41	1.87
frag19	6.6	albbabbbbbaa	bbbababbbbaa	13	0.25	1.93
frag20	5.6	bbabbbbbaaal	babaaaaaaabb	24	0.28	1.49
frag21	5.5	abbbbbaaalbb	bbbaaabaaabb	21	0.06	1.53
frag22	4.4	bbbbaaalbbbb	babaaaabbaaa	40	0.17	0.94
frag23	**5.0**	**bbaaalbbbbbb**	**aabaaaaaaaaa**	**53**	**0.60**	**0.95**

show data for different fragments at 4 *ns* from start of simulation in 1 *ns* window. The lines in bold indicate stable structures. A key finding, consistent with that of Ho et al is the persistence of substantial stable structure even in such short peptides, studied here over a more extensive set of peptides and over a systematic series of different chain lengths: in 27% of 6-mer fragments, 20% of 8-mers, 17% of 10-mers and 30% of 12-mers. We classified the structures of stable fragments using the definitions of Ho et al.[31]

Out of 98 simulated fragments, we found 22 that have structural preferences. We found six structured 6-mers each form a helical-turn, and one other forms a reverse turn. Similarly, for 8-mers, three fragments for a helical turn and one other forms a reverse turn. For 10-mers, all but one is in a helical turn structure. The 12-mers can form more complex structures, often a combination of a helical-turn and reverse-turn. Only one of the 12-mers has a stable helical structure. While it has been generally supposed that peptides this short are unlikely to have stable structure, our work is consistent with the conclusions from the previous more limited studies of Ho et al.

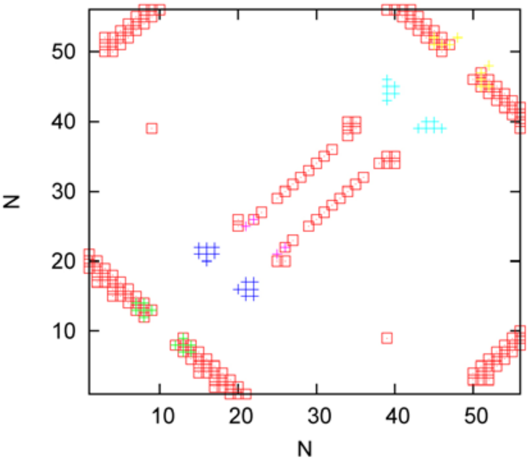


Figure 2: Same as figure 1, but for 8-mers.

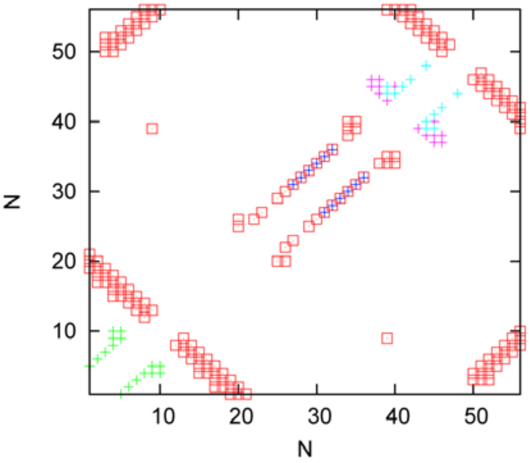


Figure 3: Same as figure 1, but for 10-mers.

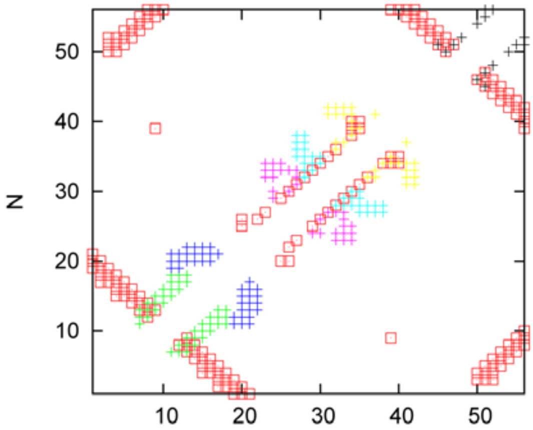


Figure 4: Same as figure 1, but for 12-mers.

In general there are 3 places in protein G where we find structured fragments: the N-terminal β-hairpin, the helix, and the C-terminal β-hairpin. At the N-terminal, the turn in the β-hairpin is predicted by the ground state of the 8-mer called frag4. For several other fragments, this turn is among the highly populated structures. Within the helix, we find two 6-mers and one 10-mer with very native-like structure. Two 12-mers are also stable, but with non-native structures. In the C-terminal β-hairpin, all our stable fragments adopt a helical-turn, the 8-mer of which is the most stable. The isolated C-terminal β-hairpin has been found experimentally to be stable,[14] where this stability is reflected in the structural bias found in the peptide fragments of the hairpin-turn. These structured fragments also are consistent with a study of Minor and Kim,[46] who replaced the α-helix sequence with a sequence based on the C-terminal hairpin. The mutant was able to fold into the same topology, showing that there is a helical propensity in the C-terminal hairpin. In this study we find helical-turns in both the α-helix and the turn of the C-terminal hairpin, which demonstrates the interchangeability of these two sequences in our simulations. Our 6-mers frag10 and frag11 and our 10-mer frag14 all show highly native-like helical structures, with very low RMSD to native and low entropy.

On figures 1–4 we plotted contact maps for native protein G and stable fragments of different sizes. We found a substantial number of stable fragments in the right places and they predict at least some of the contacts correctly. We also see some stable fragments in the places where native protein G does not have contacts. These tend to be locations where the backbone is transitioning from one structural element to another.

We also tested the proposition that keeping the five best mesostrings for a given peptide, based on our entropy and free energy analysis, might capture native-like structures in the fragments. Results are shown in figures 5–8. Figure 5 shows a 10-mer helix for which all the top conformations are native-like. In figure 6 we show struc-

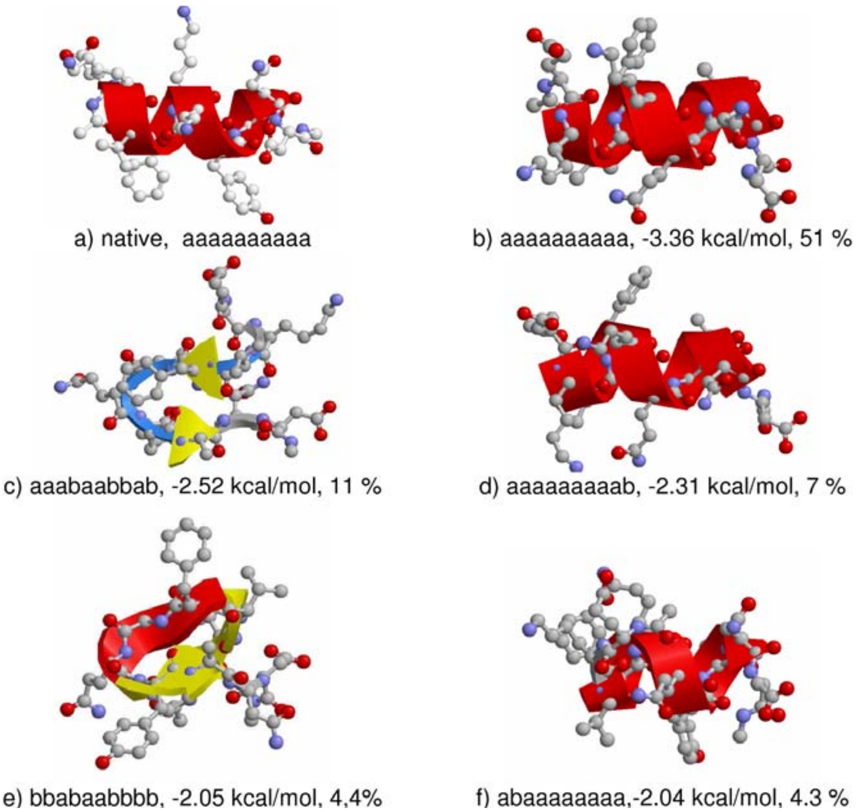

a) native, aaaaaaaaaa
b) aaaaaaaaaa, -3.36 kcal/mol, 51 %
c) aaabaabbab, -2.52 kcal/mol, 11 %
d) aaaaaaaaab, -2.31 kcal/mol, 7 %
e) bbabaabbbb, -2.05 kcal/mol, 4,4%
f) abaaaaaaaa,-2.04 kcal/mol, 4.3 %

Figure 5: Structures of 10-mer fragment 14. a) native structure in protein G, b) ground mesostring, c) 1st higher mesostring, d) 2nd higher mesostring, e) 3rd higher mesostring and f) 4th higher mesostring. For each mesostring, we show the free energy and population.

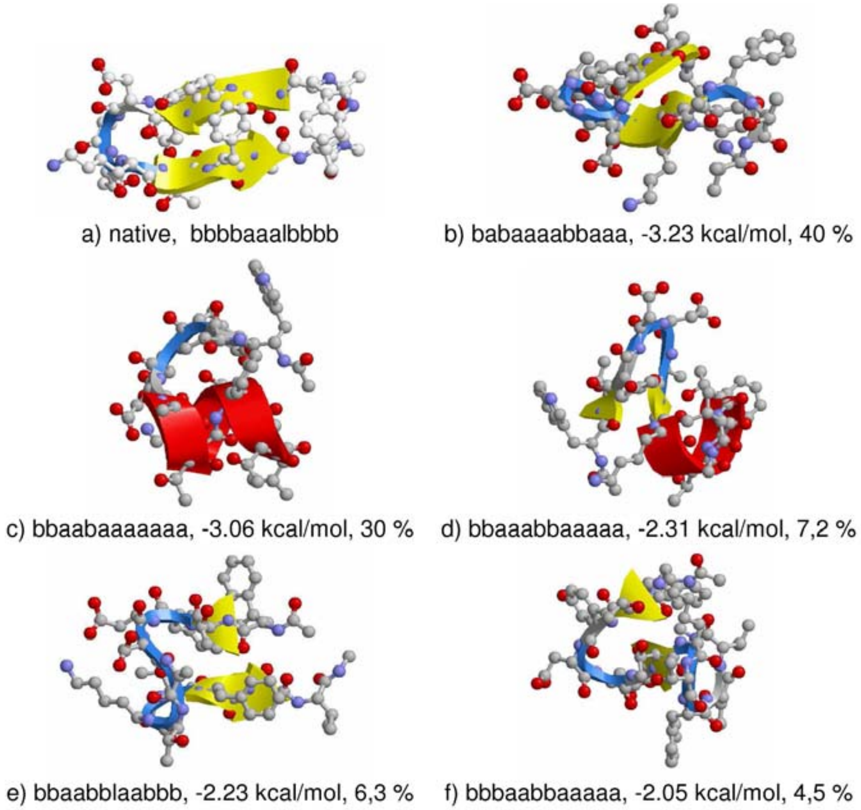
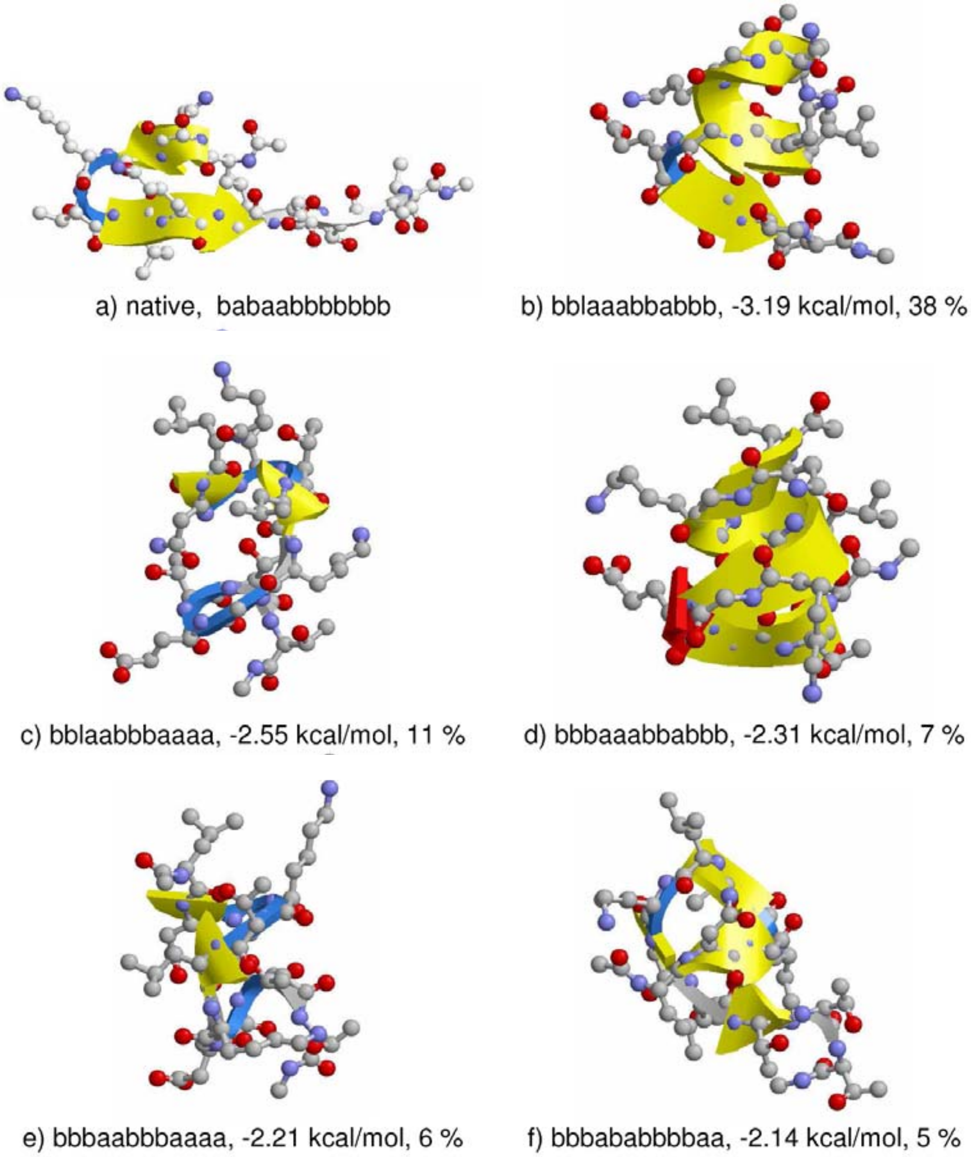

a) native, bbbbaaalbbbb
b) babaaaabbaaa, -3.23 kcal/mol, 40 %
c) bbaabaaaaaaa, -3.06 kcal/mol, 30 %
d) bbaaabbaaaaa, -2.31 kcal/mol, 7,2 %
e) bbaabblaabbb, -2.23 kcal/mol, 6,3 %
f) bbbaabbaaaaa, -2.05 kcal/mol, 4,5 %

Figure 6: Same as figure 5, only structures of 12-mer fragment frag22. This fragment is not stable.

tures of a 12-mer in the C-terminal β-hairpin. Here, we find a competition between helical-turn structures and the correct hairpin. The helix population is about 40% of all mesostrings, the hairpin is about 50% and there is a small population of other conformations. In short, in this case, by keeping the top 2 or 3 of our best conformers, they have the potential to ultimately grow into the native structure. Figures 7 and 8 show an 8-mer and 12-mer from the N-terminal hairpin. In this case, all the top conformations form the correct hairpin. Out of our total of 98 simulated peptides, we find both the correct native topology and structures that are less than 3 Angstroms RMSD from the native in 3 6-mers (frag4, frag10 and frag14), one 8-mer (frag4) and one 10-mer (frag14). Other fragments meet one criterion or the other, but not both. Hence, we find that these short physical simulations have a significant ability to identify useful native-like structures.

4. Conclusions

In this work, we have applied replica-exchange molecular dynamics using AMBER96 force field with the GB/SA solvent model to simulate protein fragments 6–12 residues in length. Out of simulated 98 fragments, 22 were structured. Despite the fact that the simulations are short in time and are on peptides that are very short in length, nevertheless, these physics-based computations are capable of picking out some native-like structures. This may be useful for simulations of protein folding kinetics, or for physics-based native protein structure predictions.


a) native, babaabbbbbbb

b) bblaaabbabbb, -3.19 kcal/mol, 38 %

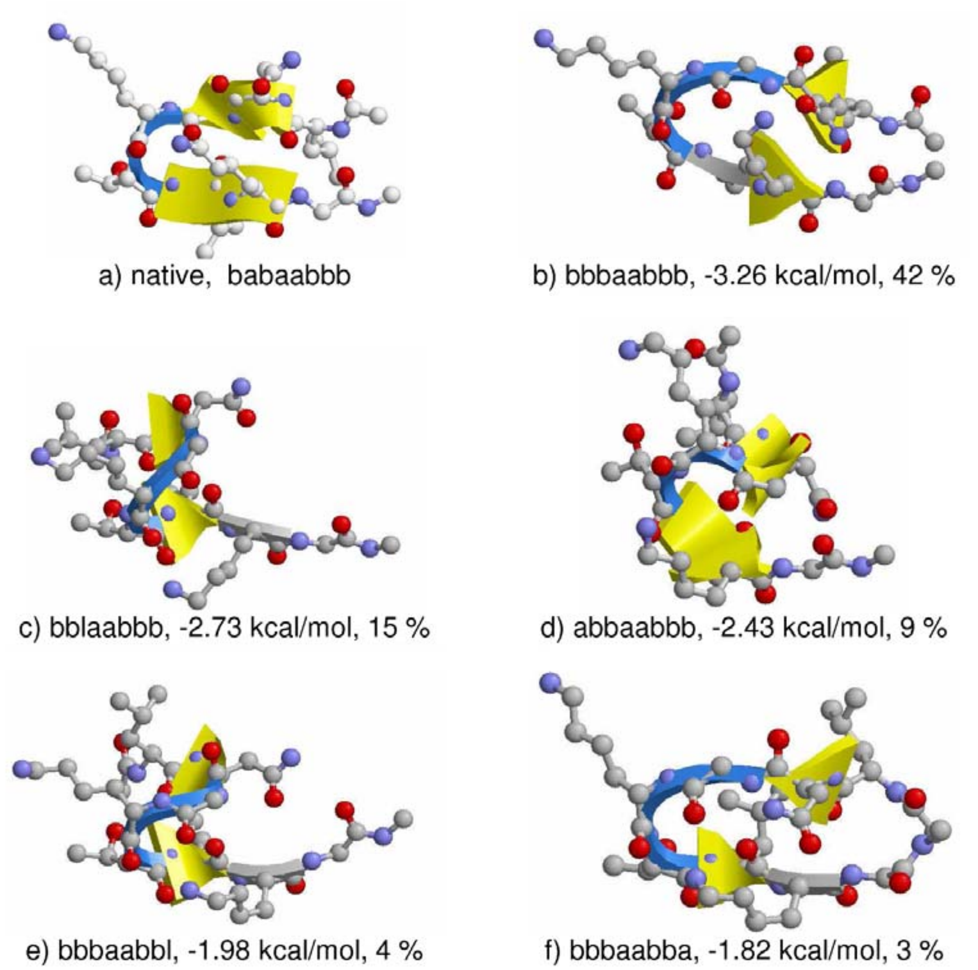
c) bblaabbbaaaa, -2.55 kcal/mol, 11 %

d) bbbaaabbabbb, -2.31 kcal/mol, 7 %

e) bbbaabbbaaaa, -2.21 kcal/mol, 6 %

f) bbbababbbbaa, -2.14 kcal/mol, 5 %

Figure 7: Same as figure 5, only structures of 12-mer fragment frag4.



a) native, babaabbb

b) bbbaabbb, -3.26 kcal/mol, 42 %

c) bblaabbb, -2.73 kcal/mol, 15 %

d) abbaabbb, -2.43 kcal/mol, 9 %

e) bbbaabbl, -1.98 kcal/mol, 4 %

f) bbbaabba, -1.82 kcal/mol, 3 %

Figure 8: Same as figure 5, only structures of 8-mer fragment frag4.

5. Acknowledgments

This work was supported by the Slovenian Research Agency (Research Program P1–0201 and P1–0010) and NIH grant GM34993. T. U. was supported through the Young Researcher Program.

6. References

1. D. Baker and A. Sali, Science 2001, 294, 93–96.
2. A. Liwo, M. Khalili and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **2005**, *102,* 2362–2367.
3. S. Oldziej, C. Czaplewski, H. D. Schafroth, R. Kazmierkiewicz, D. R. Ripoll, J. Pillardy, J. A. Saunders, Y. K. Kang, K. D. Gibson and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **2005,** *102,* 7547–7552.
4. I. A. Hubner, E. J. Deeds and E. I. Shakhnovich, *Proc. Natl. Acad. Sci. USA* **2005,** *102,* 18914–18918.
5. P. Bradley, K. M. Misura and D. *Baker, Science* **2005,** *309,* 1868–1871.
6. H. J. Dyson and P. E. Wright, *Nat. Struct. Biol.* **1998,** *5,* 499–503.
7. H. J. Dyson, G. Merutka, J. P. Waltho, R. A. Lerner and P. E. Wright, *J. Mol. Biol.* **1992,** *226,* 795–817.
8. J. P. Waltho, V. A. Feher, G. Merutka, H. J. Dyson and P. E. Wright, *Biochem.* **1993,** *32,* 6337–6347.
9. M. Ramirez-Alvarado, L. Serrano and F. J. Blanco, P*rotein Sci.* **1997,** *6,* 162–174.
10. D. Eliezer, J. Chung, H. J. Dyson and P. E. Wright, *Biochem.* **2000,** *39,* 2894–2901.
11. R. Mohana-Borges, N. K. Goto, G. J. Kroon, H. J. Dyson and P. E. Wright, *J. Mol. Biol.* **2004,** *340,* 1131–1142.
12. S. Marqusee, V. H. Robbins and R. L. Baldwin, *Proc. Natl. Acad. Sci.USA* **1989,** *86,* 5286–5290.
13. V. Munoz and L. Serrano, Nat. *Struct. Biol.* **1994,** *1,* 399–409.
14. F. J. Blanco, G. Rivas and L. Serrano, *Nat. Struct. Biol.* **1994,** *1,* 584–590.
15. M. S. Searle, D. H. Williams and L. C. Packman, *Nat. Struct. Biol*. **1995,** *2,* 999–1006.

16. R. Zerella, P. A. Evans, J. M. Ionides, L. C. Packman, B. W. Trotter, J. P Mackay and D. H. Williams, *Protein Sci*. **1999,** *8,* 1320–1331.
17. J. F. Espinosa, V. Munoz and S. H. Gellman, *J. Mol. Biol.* **2001,** *306,* 397–402.
18. K. S. Rotondi and L. M. Gierasch, *Biochem.* **2003,** *42* 7976–7985.
19. C. Bystroff, K. T. Simons, K. F. Han and D. Baker, Curr. *Opin. Biotechnol.* **1996,** *7,* 417–421.
20. C. Bystroff and D. Baker, *J. Mol. Biol*. **1998,** *281,* 565–577.
21. P. Aloy, A. Stark, C. Hadley and R. B. Russell, *Proteins* **2003,** *53,* 436–456.
22. J. Moult, Curr. Opin. *Struct. Biol.* **2005,** *15,* 285–289.
23. F. Avbelj and J. Moult, *Proteins* **1995,** *23,* 129–141.
24. R. Srinivasan and G. D. Rose, *Proteins* **1995,** *22,* 81–99.
25. N. Gibbs, A. R. Clarke and R. B. Sessions, ***Proteins*** **2001,** *43,* 186–202.
26. J. L. Klepeis and C. A. Floudas, *J. Comput. Chem.* **2002,** *23,* 245–266.
27. P. S. Kim and R. L. Baldwin, *Annu. Rev. Biochem.* **1982,** *51,* 459–489.
28. K. A. Dill, K. M. Fiebig and H. S. Chan, *Proc. Natl. Acad. Sci. USA* **1993,** *90,* 1942–1946.
29. R. L. Baldwin and G. D. Rose, *Trends Biochem. Sci.* **1999,** *24,* 26–33.
30. S. Banu Ozkan, G. A. Wu, J. D. Chodera and K. A. Dill, *Proc. Natl. Acad. Sci. USA* **2007,** *104,*11987–11992.
31. B. K. Ho and K. A. Dill, *Plos. Comp. Biol.* **2006,** *2,* 1–10.
32. Y. T. Yoda, Y. Sugita and Y. *Okamoto, Chem. Phys.* **2004,** *307,* 269–283.
33. R. Zhou, Proteins 2003, 53, 148–161.
34. Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.* **1999,** *314,* 141–151.
35. D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel and P. Kollman, *Comput. Phys. Comm.* **1995,** *91,* 1–41.
36. P. Kollman, R. Dixon, W. Cornell, T. Vox, C. Chipot and A. Pohorille, *The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data.* In A. Wilkinson, P. Weiner, W. Van Gunsteren (Eds.), Computer Simulation of Biomolecular Systems, Vol. 3, (Elsevier, 1997), pp. 83–96
37. V. Tsui and D. A. Case, *Biopolymers* **2000,** *56,* 275–291.
38. W. Zhang, C. Wu and Y. Duan, *J. Chem. Phys.* **2005,** *123,* 154105–154113.
39. J. P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *J. Comp. Phys.* **1977,** *23,* 327–341.
40. S. Kumar, D. Bouzida, R.H. Swendsen, P. A. Kollman, and J.M. Rosenberg, *J. Comp. Chem.* **1992,** *13,* 1011–1021.
41. J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill, *J. Chem. Theory Comput.* **2007,** *3,* 26–41.
42. G. N. Ramachandran, C. Ramakrishnan and V. Sasisekharan, *J. Mol. Biol.* **1963,** *7,* 95–99.
43. Y. Mu, P. H. Nguyen and G. Stock, *Proteins* **2005,** *58,* 45–52.
44. J. Hu, X. Shen, Y. Shao, C. Bystroff and M. J. Zaki, *Mining protein contact maps.* In M. Zaki, J. Wang and H. Toivonen (Eds.), Proceedings of BIOKDD02: Workshop on Data Mining in Bioinformatics, with SIGKDD02 Conference, AMC press, Edmonton, **2002,** 3–10.
45. N. Gupta, N. Mangel and S. Biswas, *Proteins* **2005,** *59* 196–204.
46. D. L. Minor Jr and P. S. Kim, *Nature* **1996,** *380,* 730–734.

Povzetek

V delu smo naredili računalniške simulacije 98 majhnih peptidov dolžine 6, 8, 10 in 12 aminokislinskih preostankov proteina G. Iskali smo stabilne fragmente. Molekulsko dinamiko zamenjave replik smo izvajali s programom Amber 7. Uporabili smo polje sil parm96 s posplošenim Bornovim modelom topila (GB/SA). Ugotovili smo, da je dober kriterij za določevanje strukturiranih fragmentov konformacijska entropija in prosta energija posameznega stanja. Visoka razlika proste energije med osnovnim in prvim višjim stanjem ter nizka entropija sta značilna za strukturirane peptide. Veliko strukturiranih peptidov ima strukture podobne nativni. Simulacije krajših peptidov so uporabne za ugotavljanje mest, kjer se protein prične zvijati, in lahko pripomorejo k boljšemu napovedovanju strukture proteinov.