A reduced model of short range interactions in polypeptide chains

Andrzej Kolinski[a)]
Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland
(and Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037)

Mariusz Milik
Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037

Jakub Rycombel
Department of Chemistry, University of Warsaw, Pasteura 1, 02-093, Warsaw, Poland

Jeffrey Skolnick
Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037

(Received 13 April 1995; accepted 7 June 1995)

A simple model of short range interactions is proposed for a reduced lattice representation of polypeptide conformation. The potential is derived on the basis of statistical regularities seen in the known crystal structures of globular proteins. This potential accounts for the generic stiffness of polypeptides, the correlation between peptide bond plates, and the sequence dependent correlations between consecutive segments of the $C\alpha$-trace. This model is used for simulation of the equilibrium and dynamic properties of polypeptides in the denatured state. It is shown that the proposed factorization of the local conformational propensities reproduces secondary structure tendencies encoded in the protein sequence. Possible applications for modeling of protein folding are briefly discussed. © *1995 American Institute of Physics.*

I. INTRODUCTION

Under proper conditions, a globular protein adopts a unique three-dimensional structure that is encoded in its amino acid sequence.[1,2] The theoretical prediction of this structure, and the pathway(s) followed during the folding process make up one of the most challenging, and still unsolved, problems of structural molecular biology.[3–5] Due to the present state of the art of computing techniques, and the time scale of the protein folding process (of the order of milliseconds to seconds), the standard molecular dynamics tools cannot be used for simulations of the folding dynamics of protein systems. This is one of the major reasons for studies of reduced models of protein structure and simplified models of polypeptide chain dynamics. Reduced models usually exploit the concept of a united atom representation of the protein chain.[6–13] This reduces the number of degrees of freedom and may make the problem computationally tractable. In the majority of previous applications, the reduced models employ a single united atom as a representation of the amino acid unit[6,14] or two united atoms per residue[7,15] (one for the main chain unit, and the second for the side group of amino acid). Interactions of the united atoms are usually deduced from regularities seen in a database of solved three-dimensional structures of globular proteins. Further simplifications of the models are frequently achieved by grouping the 20 amino acids into classes according to their properties in proteins.[16,17] Additional reduction of the number of accessible states and the subsequent increase in the speed of computations may be achieved by a lattice discretization of the conformational space. An example of extreme simplification of the protein representation is the so-called HP model[18] studied in great detail recently.[4] The model considers only two classes of amino acid residues, H (hydrophobic) and P (hydrophilic), and a simple lattice representation of the conformational space. The study of the HP model and closely related models provided some very general insights into the protein folding dynamics and thermodynamics.[19,20] On the other hand, the very low resolution of such a representation does not allow questions related to specific sequences and to finer structural detail to be addressed.

A different class of reduced models attempts to reproduce additional details of protein structure.[6,7,12,13,15,21–24] High coordination lattices can reproduce the $C\alpha$ backbone with a level of accuracy close to contemporary experimental measurements.[24] Using full sequence information and the complex set of potential functions of statistical origin, some simple folds of small globular proteins can be predicted.[15,25–28] The accuracy of this prediction varies from a level which allows almost exact full atom reconstruction (as was demonstrated for the coiled coil motif of the leucine zipper of the GCN4 fragment[27]), to low resolution folds of 2–5 Å root mean square deviation (rms) from the native $C\alpha$ trace in several other cases.[15,24,25] Unfortunately, the methodology fails for more complex folds and/or for longer sequences (the longest protein for which the model reproducibly predicts a plausible folded conformation is 120 residues and is a redesigned ROP monomer, which putatively forms a four helix bundle). It appears that a much longer simulation time for a system of this complexity was required, and/or the specificity of the force field was too low. The above calls for a more detailed examination of various interactions that control protein behavior. This could be achieved by dissection of the problem in a manner that would allow for a more precise study of the effects of the various interactions. In addition, such studies may provide insights into the factors controlling protein folding. Such factors may include hydrophobic inter-

[a)] To whom correspondence should be addressed.

0021-9606/95/103(10)/4312/12/$6.00

actions, hydrogen bonding, intrinsic local conformational preferences, and pair and higher order packing preferences.

In this work, we examine short range interactions and their effect on the static and dynamic properties of a high coordination lattice model of protein conformation. Our aim is to construct an interaction scheme which reproduces secondary structure propensities encoded in the sequence of amino acids. In this study, we neglect all the long range interactions; thus the possibility of collapse to a unique state is precluded. Based on the (one dimensional) sequence information alone, it is possible to predict the secondary structure of a protein with an accuracy in the range of 55%–70%, when the three secondary structure (helix, β-strand, loop) classes are taken into account.[29–36] This limitation of accuracy may have a physical origin, and may result from the interplay between the short and the long range interactions in the folded proteins. The long range interactions, due to more favorable packing, electrostatics, etc., may override the secondary tendencies of particular fragments. Indeed, one may find short sequences of residues that adopt completely different secondary structures in different proteins. Therefore, it is very important to design reduced models of protein chains in such a way that the above secondary structure features could be reproduced in the absence of tertiary (long range) interactions. Having such a model, the more difficult design of the tertiary interaction scheme(s) can be controlled and tested in a more rigorous way. In other words, it is our aim to develop a force field for reduced models with a local energetic frustration (local contradictions of secondary with respect to tertiary interactions), on the same level as might be expected for real proteins.

The model proposed here employs a Monte Carlo dynamics scheme, which solves a stochastic equation of motion in a discrete, conformational space. The problem of ergodicity of such models must always be addressed. While it is typically very difficult to show that a model is ergodic in the context of a rather complex potential, the best way to demonstrate ergodicity is to compare its behavior to that of simpler models which are known to be ergodic. Actually, the practical requirements for successful protein folding are somewhat stronger. One needs a model that is not only ergodic, but that is also ''practically ergodic,'' which means that the sampling is fast and that the model explores important regions of ''proteinlike'' conformational space in a reasonable amount of simulation time.

II. METHOD

A. Lattice representation and Monte Carlo scheme

The conformation of the polypeptide chain is represented by a high coordination lattice approximation of $C\alpha$-trace. The lattice chain is built from a sequence of vectors belonging to the following set {[3,1,1],...,[3,1,0],...,[3,0,0],..., [2,2,1],...,[2,2,0],...}. There are 90 vectors in this set. Fitting such a lattice to high resolution, protein structures of the Brookhaven Protein Data Bank[37,38] (PDB), the best fit is obtained when the spacing of the underlying simple cubic lattice is equal to 1.22 Å. (This lattice spacing provides the length of a $C\alpha$–$C\alpha$ segment equal to 3.8±0.3 Å.) The average accuracy of the fit is 0.6–0.7 Å rms, and is almost independent of the angle of rotation of the particular PDB structure with respect to the lattice. Only very short fragments exhibit slight orientational dependence. This is in contrast to low coordination lattice models where the quality of the fit depends dramatically on the rotation angle.[24] Excluding sparse values of the planar angles that may result from some geometrical errors in database, and neglecting the cases of *cis*-proline, the lattice fits can be regularized, with no expense in fitting accuracy. This way, the obtained lattice structures have the same distribution of the planar ($C\alpha$–$C\alpha$–$C\alpha$) angles and dihedral rotation angles as in the original PDB structures. In other words, only ''proteinlike'' sequences of three consecutive reduced backbone vectors are allowed. Consequently, about 30 vectors (instead of 90) are allowed for the third vector when the two preceding vectors are specified. This reflects the short range excluded volume and other interactions that result in the occurrence of prohibited regions of the Ramachandran map.[39] Thus, the effect of the side chains on the short range interactions is implicitly accounted for. Short range interactions between atoms, or groups of atoms, are understood here are those between units which are close to each other in sequence. It has recently been shown that the conformational energy maps generated on the basis of $C\alpha$ traces are *no* less specific than those based on the phi–psi map.[40] The secondary structure conformational propensities can be described both ways.[40] For computational expediency, the $C\alpha$ based description is employed in the Monte Carlo dynamics[41,42] scheme.

The sampling procedure works as follows:

(1) The input data, containing the sequence and a random conformation, is generated subject to the short range restrictions discussed above.

(2) A micromodification of the chain conformation is attempted. The following modifications are considered:

(a) **A two bond modification**, where two vectors are replaced by two new vectors and do not alter the conformation of the rest of the chain.

(b) **A three bond modification**, where up to 168 new three bond fragments can replace the old fragment. This number depends on the old conformation.

(c) **Chain ends modifications**. For each end, two new vectors are randomly selected. Each sequence of n residues is represented as a chain of $n+2$ vertices, or $n+1$ vectors (the two end vertices serve as terminal C- and N-caps).

The set of local moves employed here consists of a subset of moves used previously [see Figs. 2(B)–2(C) and Fig. 3(A) of Ref. 25].

(3) The local geometry is tested, i.e., all the triplets of vectors have to be ''proteinlike.'' If not, a new micromodification is attempted.

(4) The new trial conformation is subject to the Metropolis criterion,[43] according to the assumed interaction scheme with an acceptance probability equal to $P(\text{new/old}) = \exp[-(E_{\text{new}}-E_{\text{old}})/k_BT]$. The energy is expressed in k_BT units, and the temperature is dimensionless.

(5) Steps 2–4 are repeated. The arbitrary time unit is defined as a time required for $n-2$ attempts at two bond modifications, $n-3$ attempts at three bond modification, and

2 attempts at a two-bond, chain end modification. The location in the chain for each kind of move is selected by a pseudorandom algorithm.

B. Interaction scheme

The set of triplets of the chain vectors is restricted to ''proteinlike'' states. For example, since the $C\alpha$ trace of real proteins always exhibit a zig–zag geometry, three consecutive identical vectors are not allowed. There are two kinds of contributions to the short range interactions; those which are generic, and those which are sequence specific. The sequence specific potential of mean force[44] is based on the statistics of the occurrence of particular triplets of $C\alpha$–$C\alpha$ vectors in the database of known protein 3D structures. The conformation of the three virtual $C\alpha$ backbone bonds is strictly defined by two pairs of phi–psi angles for the two central α-carbons. Thus, the identity of the two corresponding central amino acids enters into the sequence specific potential. Of course, there is perhaps a moderating influence of the neighboring residues. However, this effect cannot be taken directly into account because of the too weak statistics for sequential triplets (not to mention quartets) of residues in the available structural database. On the other hand, successive pairs and associated pairwise potentials overlap along the sequence, and therefore, there is a direct influence of the identity of the neighboring residues on the conformational propensity of the fragment under consideration.

An implicit assumption is that the nature of the short range conformational restrictions seen in the native state are similar to that in the denatured state.[45] The major difference is in the long range interactions (stronger in the native state) and in the entropy of the surrounding solvent (larger for the native state). The sequence specific part of the short range interactions can be expressed as follows:

$$E_s = \sum \epsilon(A_i, A_{i+1}, \mathbf{v}_{i-1}, \mathbf{v}_i, \mathbf{v}_{i+1}). \tag{1}$$

To further reduce the numerical desorption of the local conformational propensities, the three vector descriptor is mapped onto the ''chiral'' distance between the ends of corresponding fragments,

$$E_s = \sum \epsilon(A_i, A_{i+1}, r^{2*}_{i-1,i+2}), \tag{2}$$

where A_i is the identity of residue at position i, and $\mathbf{v}_i$ is the $C\alpha$ trace vector from ith to $i+1$th $C\alpha$'s. $r^{2*}_{i-1,i+2}$ is the ''chiral'' square of distance between the corresponding chain vertices. ''Chiral'' means there is a negative sign for the left-handed conformations and a positive sign for the right-handed ones, respectively. The potential is used in the second formulation. The numerical values of the energy of various conformations, grouped into six coarse grained bins of $r^{2*}_{i-1,i+2}$ that correspond to qualitatively different structural classes, are given in Table I. The numerical value 1.0 (this is of the order of the largest absolute values of the statistically significant entries) was arbitrarily assigned to those states which do not occur in the database and to those cases when their frequency was below the level of statistical significance. The reference state for each pair is the expected number of pairs and equals the total number of occurrences of the pair of amino acids of interest times the probability that the given bin is occupied averaged over the identity of all amino acid pairs. Table I contains only the entries which are necessary for the definition of the secondary structure interactions in the 56 residue chain of B1 domain of Streptococcus protein G.[46] The entire data set is available by anonymous ftp.[47] The strength of the statistics (the number of occurrences) in the database of nonhomologous proteins is given for reader convenience.

There are also three generic short range interactions terms. We discuss their effect on behavior of the model in the next section. Here, let us just note that the generic terms provide for a ''proteinlike'' stiffness of the model chain, and penalize against nonproteinlike conformations. The first term is in the following form:

$$E_g = \sum \epsilon_g(\mathbf{v}_{i-1}, \mathbf{v}_i, \mathbf{v}_{i+1}), \tag{3}$$

E_g is defined and implemented in the same spirit as the short range sequence specific contribution. Here, the exact number of occurrences of particular triplets of vectors in the lattice fit of the database structures is used and projected onto six bins via the computation of the chiral end-to-end distance for a particular conformation. The zero of energy corresponds to the average frequency of vector triplets seen in the database. Since the full data set for this contribution is too long, we present in Table II only those values after projection onto the six bins of $r^{2*}_{i-1,i+2}$. The full data set (i.e., the numerical values of the potential of mean force as a function of $\mathbf{v}_{i-1}$,$\mathbf{v}_i$, and $\mathbf{v}_{i+1}$) used for derivation of this potential are also available via anonymous ftp.[47] Note that the straightforward usage of the Eq. (3) takes into account the underlying degeneracy of particular structural bins. This generic part of the potential is meant to suppress the conformational entropy of the lattice chains, which is somewhat higher than the corresponding entropy of real polypeptide chains. Moreover, the statistics for some pairs of amino acids is rather weak. In such cases, the generic (sequence independent) contribution splines the underlying conformational propensities encoded in the sequence specific contribution.

Geometrical correlations generated by the three-vector potentials decay too quickly down the model polypeptide chains. They are not sufficient (at any temperature) to reproduce the conformational stiffness (the relatively large correlation length for the orientation of main chain bonds) of real proteins, which results from interactions of side groups, electrostatic and/or hydrogen bond interactions between peptide bonds, and other short range interactions. This fact must be taken into account. First, we consider the database distribution of the distances between $C_{\alpha_{i-2}}$ and $C_{\alpha_{i+2}}$. This distribution strongly peaks at distances corresponding to helical or turnlike (compact) conformations. Another diffuse peak with very similar weight (the area under the distribution curve), corresponds to expanded conformations (β-strands and open loops). In contrast, the athermal lattice chain distribution is peaked at intermediate distances. Thus, a generic correction term of very simple form is introduced,

$$E_\eta = \sum \eta_i(r^2_{i-2,i+2}), \tag{4}$$

where

TABLE I. Sequence specific short range interactions.

Ai	$Ai+1$	a	1[b]	2	3	4	5	6
MET	THR	72	−0.2421	0.4126	1.0000	−0.0725	0.0561	0.0021
THR	TYR	164	−0.3931	0.2886	1.0000	0.2434	0.0070	−0.1174
TYR	LYS	136	−0.2989	0.1303	−0.0552	0.1385	0.1375	−0.0178
LYS	LEU	371	−0.1273	0.6031	0.1203	−0.1669	−0.1015	0.2468
LEU	ILE	281	−0.2831	0.5752	1.0000	−0.0441	0.0785	−0.0955
ILE	LEU	274	−0.4996	0.5310	0.1148	−0.0574	0.6080	0.0733
LEU	ASN	284	−0.0054	0.3017	0.0933	−0.1590	0.1397	−0.0252
ASN	GLY	296	0.8735	−0.7332	−1.1461	0.6547	0.1680	0.0666
GLY	LYS	394	0.5893	0.0384	−0.0725	0.7241	−1.0363	0.1661
LYS	THR	250	−0.3145	0.6405	0.0105	0.0459	−0.0991	0.0006
THR	LEU	367	−0.2717	0.0485	0.1647	0.0070	0.1663	0.2655
LEU	LYS	413	0.0682	0.3418	1.0000	−0.3496	0.4268	0.1106
LYS	GLY	290	0.6347	−0.4070	−0.4695	0.0972	0.0486	−0.1389
GLY	GLU	324	0.3683	−0.0434	−0.3689	0.4477	−0.7568	0.3000
GLU	THR	219	−0.1226	0.4359	1.0000	−0.2408	0.1759	0.1520
THR	THR	227	−0.2941	0.0536	1.0000	0.2581	−0.1607	0.1105
THR	GLU	200	0.2994	−0.4804	0.0732	−0.0355	0.1207	0.3871
GLU	ALA	346	0.7078	0.5067	1.0000	−0.6166	0.5055	0.3102
ALA	VAL	369	−0.2447	0.7457	1.0000	−0.2496	0.4216	0.0006
ASP	ALA	386	0.8168	−0.3449	0.1747	−0.1849	−0.0478	0.1534
ALA	ALA	596	0.7287	0.5001	0.2781	−0.6178	0.5744	0.4435
ALA	THR	334	−0.1217	0.3391	0.1008	−0.0810	0.0826	−0.0973
THR	ALA	349	−0.0681	−0.0527	1.0000	0.0031	0.0103	0.1147
ALA	GLU	356	0.4586	0.3675	0.1590	−0.5546	0.4587	0.5434
GLU	LYS	296	0.5252	0.2523	0.0359	−0.5112	0.3957	0.3702
LYS	VAL	304	−0.3015	0.3703	0.0847	0.0457	0.1149	−0.1127
VAL	PHE	189	−0.4833	0.4310	1.0000	0.1005	0.4375	−0.1807
PHE	LYS	187	−0.1795	0.2835	0.0657	0.0094	−0.0059	−0.0617
LYS	GLN	145	0.2709	0.2491	−0.0054	−0.2792	−0.0811	0.1759
GLN	TYR	109	−0.0571	0.1653	0.0195	−0.0780	0.0043	0.0408
TYR	ALA	191	−0.1918	0.2204	0.0215	−0.0401	0.0490	0.1195
ALA	ASN	248	0.5174	0.0935	0.0539	−0.3249	−0.0720	0.2760
ASN	ASP	163	0.4784	−0.3339	−0.1996	0.1193	−0.2897	0.2596
ASP	ASN	193	0.6803	−0.1706	−0.2875	0.0454	−0.3789	0.1241
GLY	VAL	431	−0.1405	0.2535	0.0642	0.5903	−0.6581	−0.0975
VAL	ASP	331	−0.1834	0.0687	−0.1771	0.0949	0.2097	−0.1396
ASP	GLY	381	1.1010	−0.9586	−0.8265	0.5769	0.1781	0.2384
GLU	TRP	56	0.1863	0.1934	1.0000	−0.2662	0.0107	0.0291
TRP	THR	60	−0.3193	0.1355	1.0000	0.1562	0.0678	0.0018
TYR	ASP	197	−0.3507	0.1027	−0.0195	0.1085	0.2070	0.1026
ASP	ASP	222	0.9116	−0.3026	−0.1278	−0.0874	−0.3219	0.3228
THR	LYS	213	−0.1293	−0.1414	−0.0103	0.1946	−0.0643	0.1089
THR	PHE	189	−0.3455	0.0563	1.0000	0.3053	0.1826	−0.2696
PHE	THR	185	−0.5174	0.4181	1.0000	0.2018	0.3626	−0.2131
THR	VAL	330	−0.6289	0.3303	1.0000	0.4862	0.3011	−0.2772
VAL	THR	330	−0.6117	0.4940	0.1452	0.4303	0.3827	−0.4107

[a] Number of occurrences in the database.

[b] Ranges of $r^{2*}_{i-1,i+2}$ 1 (−86,−57) extended beta, 2 (−56,−26) loops (left-handed), 3 (−25,0) left-handed helix, 4 (1,25) right-handed helix, 5 (26,56) loops (right-handed), 6 (56,91) extended beta.

$$\eta_i=-1 \quad \text{for} \quad r^2_{i-2,i+2}<35$$

$$\eta_i=-1 \quad \text{for} \quad r^2_{i-2,i+2}>75$$

$$\eta_i=0 \quad \text{otherwise.}$$

TABLE II. Generic three bond potential.

1	2	3	4	5	6
−0.0520	0.1057	2.4738	−0.9866	0.0746	1.0431

All the numbers are given in lattice units and can be easily translated into corresponding distances in real proteins (1 in lattice units is equivalent to 1.22 Å).

The last contribution to the short range interactions is designed to propagate secondary structure and to further contribute to the peptidelike stiffness of the model chain. First, let us note that the three consecutive $C\alpha$ vectors define the orientation of the central polypeptide bond with levels of high accuracy. In other words, the $C\alpha$ trace can be used for full atom reconstruction of the main chain conformation.[48–57] The most straightforward approach is to store the positions of the carbonyl oxygen atom of the ith residue and the nitrogen atom of the $i+1$st residue (they belong to the same peptide bond) in the reference coordinate system defined by the three vectors $\mathbf{v_{i-1}}$, $\mathbf{v_i}$, and $\mathbf{v_{i+1}}$. For all possible local backbone configurations, the width of the distribution does not exceed 0.22 Å (root mean square deviation) for the nitrogen atom position; about 0.25 Å for the carbonyl carbon positions, and about 0.45 Å for the carbonyl oxygen[57] position. Since there is a strong correspondence between the $C\alpha$ trace and phi–psi description of a polypeptide conformation, this is not surprising. Consequently, one may define the direction of peptide plate (we use here the hydrogen-to-oxygen vectors) employing the $C\alpha$-trace as a convenient reference frame (the numerical data are again available via anonymous ftp).[47] The angular error of such a reconstruction of the direction of the peptide bond plate (we assume the typical *trans* conformation of the peptide bonds) does not exceed 15 deg. This seems to be more than adequate for our purposes. When the above method of reconstruction of peptide bonds is applied to the PDB $C\alpha$ reduced structures, almost all hydrogen bonds (~89%) of the main chain (short range and long range) could be identified. The Kabsch–Sander[58] method is used as a reference assignment. This will be used in the forthcoming work as a fast method of computing hydrogen bond interactions in the framework of the reduced lattice model. Having the orientation of the peptide bonds, one can introduce a bias towards regular (helix or β-strand) conformations of the model polypeptides. For these structural elements, the ith peptide bond plate is almost parallel to the $i+2$nd and to the $i+4$th peptide plate. The corresponding potential is of the following form:

$$E_p=\Sigma[\cos(\mathbf{h}_i,\mathbf{h}_{i+2})+\cos(\mathbf{h}_i,\mathbf{h}_{i+4})], \tag{5}$$

where $\cos(\mathbf{h_i},\mathbf{h_j})$ denotes the cosine between the ith and jth vectors defining the orientation of the peptide plates (the vectors from hydrogen bonded to the carbonyl oxygen).

The total energy of the model chain is computed as

$$E=4E_s+E_g+E_\eta+E_p. \tag{6}$$

The scaling of the sequence specific interactions against the generic ones is, to some extent, arbitrary. This particular choice has been made by a trial and error method for various proteins belonging to different structural classes. Let us finally note that instead of the $C\alpha$ vectors, one may use side group vectors as a basis for factorization of the sequence specific conformational propensities.[25] Here, however, we try to keep the models as simple as possible and open for easy implementation of various long range interactions.

III. RESULTS AND DISCUSSION

The proposed model of the short range interactions and dynamics of the protein chain has been tested on several proteins. Two major problems need to be addressed. The first is related to the dynamics of the model and its ergodicity. The second is the problem of reproducing the secondary structure encoded in the amino acid sequence. If the proposed factorization of the secondary structure is correct then the simulations at low temperatures should lead to results that coincide in 55%–70% of the cases with the secondary structure seen in the native state. Due to lack of tertiary interactions, we do not require higher accuracy (see the comments in the Introduction). The secondary structure definition in the model without the long range interactions has to be somewhat modified. It is understood here as a conformation of the main chain which is consistent with the conformation of the chain fragments in the secondary structure seen in the native state. For helical conformations, both definitions are virtually the same (one may introduce a geometrical criterion for detection of hydrogen bonds). For beta structure, very expanded conformations are considered as fragments of a hypothetical β-sheet. This differs from the Kabsch–Sander[58] assignment (which is more frequently used); nevertheless, the present definition has also been previously used.[42–54]

A. Dynamic properties of the model

The dynamics of the proposed model are examined in detail on the example of the B1 domain of Streptococcus protein G, which is a small protein consisting of 56 residues. In spite of its small size, the fold of protein G is exceptionally regular and very stable.[46] The fold consists of four stranded β-sheets and a single α-helix whose topology could be classified as $(-1,+3x,-1)$. In Table III, the sequence and the secondary structure assignment of B1 domain of protein G are found. For the readers' convenience, we present a simplified notation based on the Kabsch–Sander method, which is commonly used in various methods of prediction of secondary structure from sequence of amino acids, where only three outcomes are considered [helix (H), beta (E), and everything else (-)].

At high temperatures, the model should behave as a Rouse chain.[59] Indeed, this is the case. In Fig. 1, the center of gravity autocorrelation function[41,59] (the time averaged square of displacement of the center of gravity, computed for the $C\alpha$ backbone), is plotted vs time on a log–log scale. The results for various temperatures show free diffusion (a straight line with slope equal to 1). With decreasing temperature, the diffusive motion of the model chain slows down. As shown in Fig. 2, where the end-to-end vector autocorrelation functions are plotted vs time in semilog plots, the relaxation of the chain orientation is exponential. In all cases, the initial orientation decays exponentially with the longest relaxation dependent on temperature. For all temperatures, the presented data are generated on the basis of a hundred times longer trajectory than the time range shown in the plots. The error bars are in range of the symbol size (except for a rather irrelevant part of the end-to-end autocorrelation function in the limit of negligible memory of the initial state), and therefore, they are omitted in the pictures.

In conclusion, the model chain behaves as a Rouse chain. Due to the well known ergodicity of the Rouse chains[41,59] this suggests that the present model is also ergodic or at least it belongs to an acceptable ergodicity class. This is not surprising, due to the high coordination number of the lattice. The present model could be considered as a convenient discretization of a continuous (off-lattice) chain. At very low temperatures, the mobility of the model chain is suppressed; however, segmental free diffusion and the relax-

TABLE III. Comparison of accuracy of prediction of secondary structure for ten randomly selected proteins. The first line is the residue number. The second line gives the sequence. The third and the fourth lines denote secondary structure obtained via the Kabsch–Sander method applied to the crystal structure and via the predicted conformational statistics described in this work. The threshold values are the following: $r^2_{i-2,i+2} > 75$ assigns E, $r^2_{i-2,i+2} < 56$ and the chirality is positive, assigns H. Everything else is $(-)$.

1*cd*8			56.1% correctly predicted	
1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0
S Q F R V S P L D R	T W N L G E T V E L	K C Q V L L S N P T	S G C S W L F Q P R	G A A A S P T F L L
- - E E E E - - - -	- - - - - - - E E E	E E E E - - - - - -	- - E E E E E E - -	- - - - - - E E E E
- - E E E E E - - -	E - H H - E E E - H	H - E E E E E - - -	- - - E - - E E E -	- H H H H E E E - H
1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0
Y L S Q N K P K A A	E G L D T Q R F S G	K R L G D T F V L T	L S D F R R E N E G	Y Y F C S A L S N S
E E - - - - - E E -	- - - - - - - E E E	E E E - - E E E E E	E - - - - H H H - E	E E E E E E E E - -
H H H - - - - - H H	H - - - E E E E E -	H H H - E E E E E E	E - - H H H H H - -	E E E E - - H H - E
1 2 3 4 5 6 7 8 9 0	1 2 3 4			
I M Y F S H F V P V	F L P A			
E E E E - - - E E E	- - - -			
E E - - - - - E E E	E - - -			
1*crn*			60.9% correctly predicted	
1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6
T T C C P S I V A R	S N F N V C R L P G	T P E A I C A T Y T	G C I I I P G A T C	P G D Y A N
- E E - - - H H H H	H H H H H H H - - -	- - H H H H H H H H	- - E E - - - - - -	- H H H - -
- E E E E E E - H H	H H H H H H H H - H	H - H H - E E E E E	- - E E E - - - E E	- H H H H -
1*ctf*			58.8% correctly predicted	
1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0
E F D V I L K A A G	A N K V A V I K A V	R G A T G L G L K E	A K D L V E S A P A	A L K E G V S K D D
- E E E E E E E E -	H H H H H H H H H H	H H H H H - - - H H	H H H H H H - - - E	E E E E E E - H H H
- E E E E - - H H -	H H - E E E E - - -	- - - - - - - H H H	H H H H H H H E E -	- E - - - - - - H H
1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8			
A E A L K K A L E E	A G A E V E V K			
H H H H H H H H H H	H - - E E E E -			
H H H H H H H H H H	H - - - E E E -			
1*gb*1			73.2% correctly predicted	
1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0
M T Y K L I L N G K	T L K G E T T T E A	V D A A T A E K V F	K Q Y A N D N G V D	G E W T Y D D A T K
- E E E E E E - - -	- - - E E E E E E -	- - H H H H H H H H	H H H H H - - - - -	- E E E E E - - - -
- E E E E E E E - -	E - - - - E E E - H	H H H H H H H H H H	H H H H H - - - E E	- - E E E E E E E E
1 2 3 4 5 6				
T F T V T E				
E E E E E -				
E E E E E -				
1*mba*			54.8% correctly predicted	
1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0
S L S A A E A D L A	G K S W A P V F A N	K N A N G L D F L V	A L F E K F P D S A	N F F A D F K G K S
- - - H H H H H H H	H H H H H H H H H -	H H H H H H H H H H	H H H H H - H H H H	H H - - - - - - - -
- H H H H H H H H H	- - - - - E E - H H	H H H H - H H H H H	H H H H H H - - - -	H H H H H H - - - -
1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0
V A D I K A S P K L	R D V S S R I F T R	L N E F V N N A A N	A G K M S A M L S Q	F A K E H V G F G V
H H H H H - - - - H	H H H H H H H H H H	H H H H H H H - - -	H H H H H H H H H H	H H H H H H H - - -
- H H H H H H - - H	H H E E - E E E E -	H H - E E - - - H H	H - H H H H H H H H	H H H H H - - - - -
1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6
G S A Q F E N V R S	M F P G F V A S V A	A P P A G A D A A W	T K L F G L I I D A	L K A A G A
- H H H H H H H H H	H H H H H H H - - -	- - - - - H H H H H	H H H H H H H H H H	H H H - - -
- - - H H H H - E -	- - - - - - - - - H	H - - - - H H H H H	- - - - - E E E - H	H H H H - -
1*pcy*			60.6% correctly predicted	
1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0
I D V L L G A D D G	S L A F V P S E F S	I S P G E K I V F K	N N A G F P H N I V	F D E D S I P S G V

TABLE III. (*Continued.*)

```
- E E E E - - - - -   - - - E E - - E E E   E - - - - E E E E E   E - - - - - E - - E   E - - - - - - - - -
- E E E E - - - - -   - - H H H - - - - -   E E - - H H E E E -   - - - - E E E E E E   E - - - - E E E E E
1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0
D A S K I S M S E E   D L L N A K G E T F   E V A L S N K G E Y   S F Y C S P H Q G A   G M V G K V T V N
- H H H H - - - - -   - - E - - - - - E E   E E E - - - - E E E   E E E - H H H - - -   - - E E E E E E E -
- - - - E E E - - H   H - E - - - - - - E   - - H H - - - - E E   E E E E E - - - - -   - E E E E E E E E -

2pab.A                                                52.6% correctly predicted

1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0
C P L M V K V L D A   V R G S P A I N V A   V H V F R K A A D D   T W E P F A S G K T   S E S G E L H G L T
- - E E E E E E E -   - - - E E - - - - E   E E E E E E - - - -   - E E E E E E E E -   - - - - E E - - - -
- - E E E E E E - -   - - - - - - H H - -   E E E E - H H H H -   E E E - - H H - - -   - - - - H H - - E E
1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0
T E E Q F V E G I Y   K V E I D T K S Y W   K A L G I S P F H E   H A E V V F T A N D   S G P R R Y T I A A
- - - - - - - E E E   E E E E - H H H H H   H H H - - - - - - -   - E E E E E E E - -   - - - - - E E E E E
- - H H E E - - - E   - - - - - - - - E -   H H H - E E - - - H   H H H E E E E E E E   E E - - - E E - H H
1 2 3 4 5 6 7 8 9 0   1 2 3 4
L L S P Y S Y S T T   A V V T
E E E - - E E E E E   E E E -
H - - - - - E E E E   E E E -
351c                                                  61.0% correctly predicted

1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0
E D P E V L F K N K   G C V A C H A I D T   K M V G P A Y K D V   A A K F A G Q A G A   E A E L A Q R I K N
- - H H H H H H H -   - H H H - E - - - -   E - - - - - H H H H   H H H H - - - - - H   H H H H H H H H H H
- - - E E E E H H H   - - E E - E E E E E   - E E - - - H H H H   H H H H H - H H - H   H H H H H H H H H H
1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2
G S Q G V W G P I P   M P P N A V S D D E   A Q T L A K W V L S   Q K
- E - - - - - - - -   E - - - - - - H H H   H H H H H H H H H -   - -
- - - - E E E E E E   E - - - - E E - - H   H H H H H H H E E   - -
3fxn                                                  60.1% correctly predicted

1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0
M K I V Y W S G T G   N T E K M A E L I A   K G I I E S G K D V   N T I N V S D V N I   D E L L N E D I L I
- E E E E E - - - -   H H H H H H H H H H   H H H H H - - - - -   E E E E H H H - - -   - - - - - - - E E E
- - E E E E E E - -   - - - H H H H H H H   H - - H H H - - - -   E E E E E E - E E E   - H H H H - - E E E
1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0
L G C S A M G D E V   L E E S E F E P F I   E E I S T K I S G K   K V A L F G S Y G W   G D G K W M R D F E
E E E - - E - - - E   - - - - - H H H H H   H H H - - - - - - -   E E E E E E E E - -   - - - H H H H H H H
E - - - - - - - - -   H H H H H H H - - -   H H - - - E E E - -   - - H H H - - - - -   - - - - - H H H H H
1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8
E R M N G Y G C V V   V E T P L I V Q N E   P D E A E Q D C I E   F G K K I A N I
H H H H H - - - E E   - - - - E E E E - -   - H H H H H H H H H   H H H H H H H H - -
H H H H - - E E E E   E E E E E E E E - -   - - H H H H H H H H   H - H H H H H -
3trx                                                  50.5% correctly predicted

1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0
M V K Q I E S K T A   F Q E A L D A A G D   K L V V V D F S A T   W C G P C K M I K P   F F H S L S E K Y S
- - E E - - - H H H   H H H H H H H - - -   - - E E E E E E - -   - - - - - H H H H H   H H - - H H H H - -
- - - - - H H - - H   H H H H H H H H - H   H - E E E - - - - -   E E - - - - E E E E   E E E E - - - - - -
1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0   1 2 3 4 5 6 7 8 9 0
N V I F L E V D V D   D C Q D V A S E C E   V K C T P T F Q F F   K K G Q K V G E F S   G A N K E K L E A T
- - E E E E E E - -   - - H H H H H H H -   - - - - - E E E E E   E - - E E E E E E E   - - - - H H H H H H
E E E E E E E E - -   - - - - - - - - E E   E E E E E E E E E -   - - - - - E E - E E   - - - - H H H H H -
1 2 3 4 5
I N E L V
H H H H -
E - - - -
```

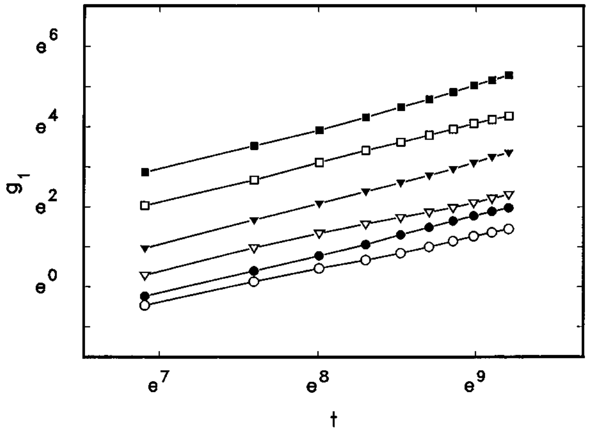

FIG. 1. Log–log plot of the center of mass autocorrelation functions for the protein G chain at various temperatures (open circles, T=1.0; solid circles, T=1.1; open triangles, T=1.25; solid triangles, T=1.5; open squares, T=1.75; solid squares, T=5.0).

ation of the initial internal coordinates is evident at very long times as is shown in Fig. 1.

Figure 3 shows two profiles at two different times of the single residue autocorrelation function (square of displacement) at the relatively low temperature of T=1.1. At this temperature, the secondary structure preferences are already highly visible. For a Rouse chain, a parabolic shape of the profiles is expected. The overall shape of the profiles shown in Fig. 3 are close to parabolic; however, due to the different flexibility of various fragments of the model polypeptide, there are noticeable distortions. For example, residues 50 to 54 tend to move with the same velocity, regardless of their different separation from a more mobile chain end. This is related to a strong preference of this fragment of the chain to adopt very expanded β-strand like conformations (see Table III). The rotational motion of such a rigid fragment is somewhat hindered; however, the translation is even faster than for more flexible fragments. Remarkably, the translational motion of a portion of the helical part of the chain also seems to be faster than expected for a homopolymeric Rouse chain. The above features of the model polypeptide could be examined more closely by the direct analysis of the orientational autocorrelation function for various fragments of the chain. The relaxation of several selected fragments is illustrated in Fig. 4 where the autocorrelation functions, $g_4(t)$ for the $\mathbf{R}_{\mathbf{i-2,i+2}}$ vectors are drawn in semilog plots for two temperatures. While the differences between the speed of local relaxations of various structural elements are rather small at low temperature, the relaxation rate of the two β-fragments next to the central helix is the lowest. The fastest relaxation is

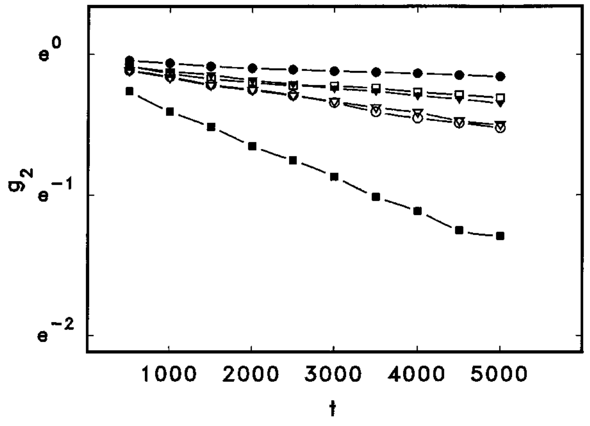

FIG. 2. Semilog plot of the end-to-end vector autocorrelation function of the protein G chain at various temperatures (symbols as in Fig. 1). (Solid circles, T=1; open squares, T=1.1; solid triangles, T=1.25; open triangles, T=1.5; open circles, 1.75; solid squares, T=15.)

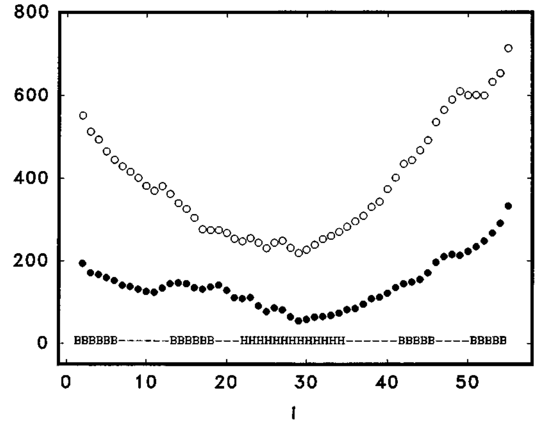

FIG. 3. Average displacement of a single $C\alpha$ atom at two times (t=1000, open circles; t=500, solid circles) as a function of the position in the chain, at temperature T=1.1.

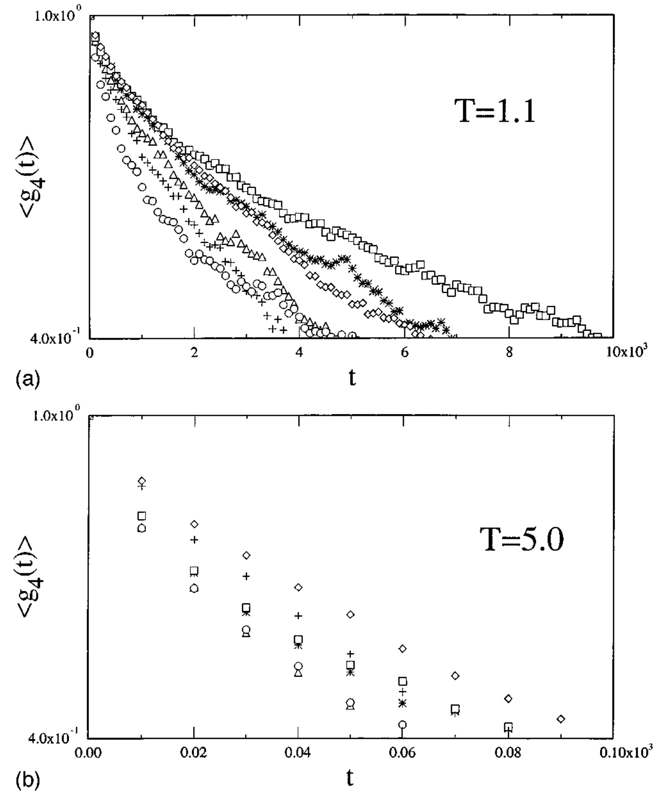

FIG. 4. Semilog plot of the four-bond vector orientational autocorrelation function for various fragments of the protein G chain. The set of curves in (a)[(b)] corresponds to T=1.1; [T=5.0] (the crosses show the autocorrelation function for residues 2–6, the stars for residues 14–18, triangles for residues 23–27, circles for residues 31–35, squares for residues 42–46, and diamonds for the fragments 51–55).

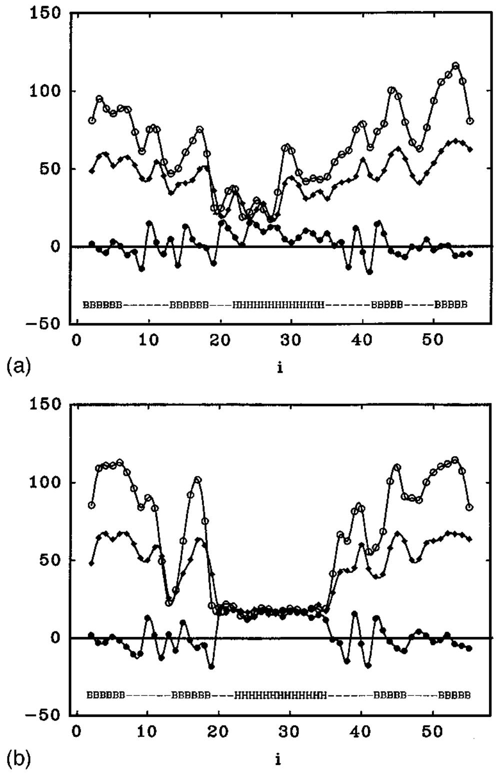

FIG. 5. Profiles of various conformational characteristics of protein G. The upper curve (open circles) shows the time average of the square of the distance between the $i-2$nd and $i+2$nd $C\alpha$'s as a function of position along the chain. The middle curve (solid diamonds) shows the corresponding plot for the square of the distance between the $i-1$th and $i+2$nd $C\alpha$'s. The lowest curve (solid circles) represents the handness of the three bond fragments (see the text for more details) [(a) $T=1.1$; (b) $T=1.0$].

exhibited by the *C*-terminus of the putative helix. At higher temperatures, the relaxation rate of the strongly helical region is closer to the average of the other fragments, however, the helix still relaxes with the highest rate.

The global dynamics of the model chains, which results from a long random sequence of very local conformational jumps, is virtually identical to the dynamics of an ideal polymer chain.[59] On a local level, the dynamics are somewhat moderated by the temperature dependent fluctuating secondary structure. However, even at very low temperatures when some fragments are structured during the entire time of simulation, the diffusive motion is not prohibited.

B. Secondary structure of model polypeptides

To what extent is the secondary structure, seen in its native state, reproduced by the present reduced model without the long range interactions? We discuss in more detail the case of protein G. It should be noted that protein G was not included into the database of the structures used for derivation of the statistical potential of mean force given in Table I. Moreover, there is no sequence or structure homology to any protein from the database. The list of structures used to construct the potential is also available via anonymous ftp.[47] In Fig. 5(a), the three profiles that can be used to deduce the secondary structure are plotted based on the statistics from long runs at low temperature, $T=1.1$. The upper profile corresponds to time average of $r^2_{i-2,i+2}$, the middle curve is to $r^2_{i-1,i+2}$, and the lowest one reflects average handness of the main chain $C\alpha$ trace computed as $(\mathbf{v_{i-1}} \otimes \mathbf{v_i}) \cdot \mathbf{v_{i+1}}$. The reduced secondary structure assignment (compare Table III) of the native protein is included for easy reference. The threshold for the β-typical value of $r^2_{i-2,i+2}$ can be chosen in such a way that the location of all four β-strands (with the possible exception of the second one) can be correctly identified, including the very likely locations of the turns. For all of the data, we use the same threshold given in the caption to Table III. Even the lower peaks at positions 18 and 40 (approximate) have physical meaning. They coincide with very open and relatively long connections between the central α-helix and the neighboring β-strands. Qualitatively, the same picture can be deduced from the $r^2_{i-1,i+2}$ profiles, with, of course, a different threshold value. In both cases, the *C*-terminal β-strand has the strongest prediction. The helix in the native state runs from residue 22 to residue 35. The two upper profiles show a well defined helix between residues 20 and 28 (small distances between the $C\alpha$ atoms), while the *C*-terminal part of the helix, although visible, is less obvious. During the simulations, this part of the helix dissolves after a time, contributing to a somewhat weaker prediction. The handness profile correctly identifies a long stretch of right handed turns from residue 20 to residue 37 that corresponds

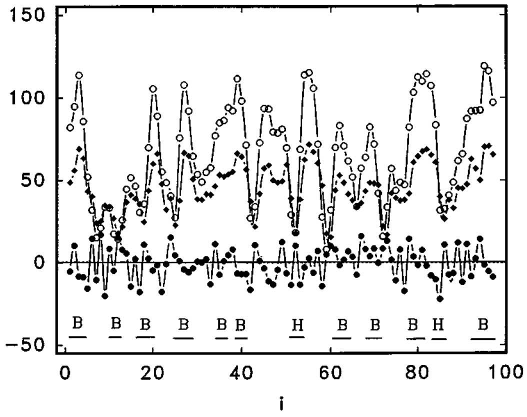

FIG. 6. Profiles of various conformational statistics for the plastocyanin sequence. $T=1.0$ (for additional details, see the caption to Fig. 5).

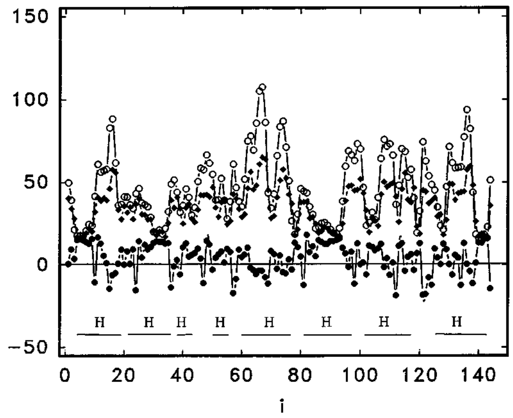

FIG. 7. Profiles of various conformational statistics for the myoglobin sequence. $T=1.0$ (for additional details, see the caption to Fig. 5).

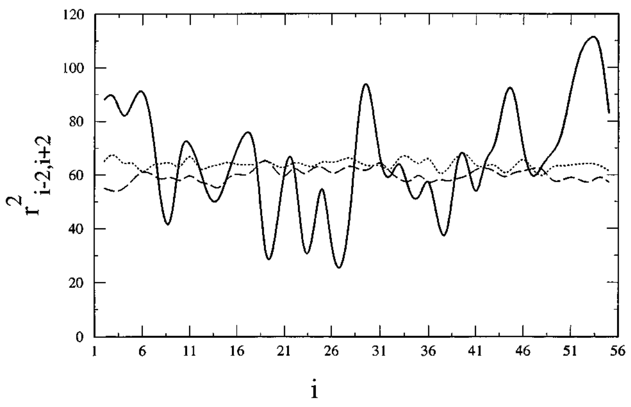

FIG. 8. Comparison of the effect of sequence specific potential and the generic potential on the average values of $r^2_{i-1,i+2}$ as a function of residue position in the sequence of the domain B1 of protein G. Solid line, the case of sequence specific potential without the generic regularized terms; dashed line, the case of the generic potential; the dotted line, the case of a phantom athermal chain.

to the entire helical fragment. The β-fragments can be right-handed or left-handed, and the profile is consistent with the native secondary structure. In summary, for this particular sequence, the compilation of various local conformational characteristics obtained from long Monte Carlo simulation lead to an accurate prediction of secondary structure. The errors of positioning of particular secondary structure elements do not exceed two residues. The results are even clearer when the system is further cooled down to $T=1.0$, as shown in Fig. 5(b).

In Figs. 6 and 7, the analogous profiles are presented for two larger proteins, the 99 residues β-protein plastocyanin (1pcy), and the 146 residues, helical protein myoglobin (1 mba). The same input data (temperature and scaling of specific vs generic interactions) are used in all cases. The resulting secondary structure is correct in most cases, however, some errors are noticeable.

The tests on other sequences show that the present model reproduces secondary structure on the same level of accuracy (that is, 55%–70% for three structural classes under consideration) as obtained by other methods of secondary structure prediction.[29–36]

C. Interplay between specific and nonspecific short range interactions

First, we note (again on example of protein G domain) that the accuracy of the secondary structure prediction drops significantly, by ~5%, when the generic terms are removed. The lower average accuracy and more scattered assignments of secondary structure comes from the absence of "propagation" due to the generic contributions that simulate conformational stiffness and some local cooperativity of polypeptide chains. This is demonstrated in Fig. 8, where the values of $r^2_{i-2,i+2}$ are plotted along the sequence of the protein G domain. The time average values for particular residues are much more scattered than those shown in Fig. 5. For instance, there is a very expanded fragment at positions 29–30 occupied by β-forming Val and Phe. This local β-tendency is balanced by the neighboring amino acids which prefer helix when the collective generic potential is in force. For comparison, Fig. 8 also contains the results for a chain with only generic terms of the short range interactions (the dashed line) and the results for athermal phantom chain (dotted line). Since there are not any sequence specific effects, both profiles are flat. Interestingly, the averages are almost the same, in spite of different distributions that are discussed below.

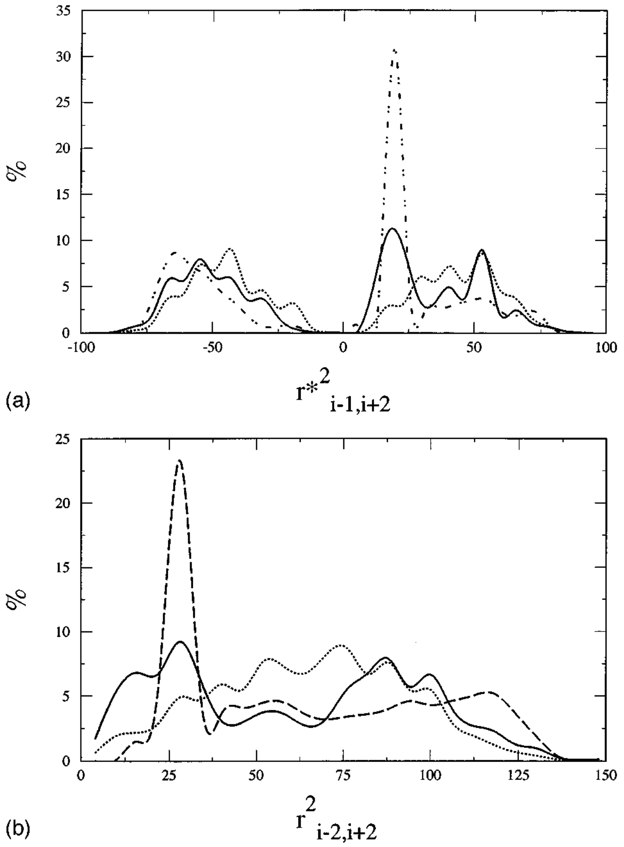

FIG. 9. Comparison of average distribution of the local conformational statistics for the 56 residue lattice chain (sequence irrelevant) with and without generic potentials. In both cases the sequence specific part of the potential is absent. (a) The sample distributions of $r^{*2}_{i-1,i+2}$, averaged over all residues for the system with the generic potential (solid line) and for the athermal phantom chain (dotted line) compared to the PDB distribution (dashed line). (b) The corresponding distributions for $r^2_{i-2,i+2}$.

Further insight into the role of the sequence independent regularizing potential comes from analysis of the distribution of conformations (averaged over entire chain) as measured by the chiral three bond and four bond distances, respectively. In Fig. 9, the distribution for a completely athermal chain (infinite temperature, dotted line) and for the chain with only the generic potential (solid line) are compared to the distributions from the structural data base (dashed line). The distributions for the second case are proteinlike in the sense that there are well defined peaks corresponding to right-handed compact (helical) and expanded (β-type) states. The population of compact left-handed states is negligible. Thus, the generic background potential introduces protein like conformational bias. The sequence specific potential triggers formation of fluctuating secondary structure. Thus, amino acid pair specific propensities are "interpolated" by the generic potential. As a result, the observed secondary structure (time averaged) is partially (as it should due to absence of the interplay between the short and long range interactions) consistent with the secondary structure seen in the native state.

IV. SUMMARY AND CONCLUSIONS

The proposed reduced model of polypeptide conformation employs a high coordination lattice for the $C\alpha$ representation of the main chain backbone. This lattice representation is very accurate. The mean square error of the lattice $C\alpha$ approximation to PDB structures is on the level of 0.6–0.7 Å. What is important is that the quality of the fit does not depend on the angle of rotation with respect to the lattice. Thus, no artificial entropy effects are encountered. The lattice $C\alpha$ trace provides a convenient reference frame for reconstruction of the coordinates of all backbone atoms. The reconstruction requires just a few references to the prefabricated data set and can, therefore, be performed frequently during very long simulations. The model of dynamics is based on a long random sequence of local conformational transitions that preserve ''proteinlike'' backbone geometry. In this work, only the short range interactions and their effect on the behavior of the model are considered. In order to achieve a ''proteinlike'' distribution of conformations, it is necessary to employ a generic (sequence independent) background potential, which introduces a correction to the underlying lattice distribution of states. This generic potential is designed on the basis of general regularities seen in all protein structures. It enforces a ''proteinlike'' distribution of distances between the $C\alpha$ atoms up to the fourth neighbors down the chain. There is also a bias towards the correlated mutual orientations of the polypeptide bonds, which is typical for all regular secondary structure motifs. The sequence specific interactions trigger the specific local secondary structure preferences. The sequence specific part of the potential is derived from the configurational statistics of the high resolution PDB structures. It should be noted that a somewhat similar factorization of the secondary structure propensities for a $C\alpha$-reduced description of polypeptide chains has been recently described by DeWitte and Shakhnovich.[60] They also assumed a sequence specific factorization of the potential that depends on the dihedral angle for four-bond $C\alpha$-backbone fragments and the identity of the two central residues. Their potential was successfully used for sequence-structure matching. However, since they neglect the variability of the planar ($C\alpha$–$C\alpha$–$C\alpha$) angles,[40] this potential cannot be used for explicit simulations of protein geometry on a high-coordination lattice.

The results presented in this work show that it is possible to design a lattice model of a protein which reproduces strong secondary structure propensities and at the same time, exhibits global dynamics which are very similar to that of an ideal Rouse chain. Thus, for all practical reasons, the model is ergodic. Starting from an arbitrary chosen initial state, the Monte Carlo algorithm rapidly achieves thermodynamic equilibrium with fluctuating secondary structure that is typical for the given sequence of amino acids. The model seems to be a plausible candidate for simulations of the long time dynamics of denatured proteins (the algorithm used in the present work is available upon request). At lower temperatures, the observed (time averaged) secondary structure (deduced from the observed short range conformational correlations) is close to that seen in the native state. The accuracy of this method of secondary structure prediction is of the same level as obtained by standard methods (i.e., 55%–70% of residues are correctly assigned). Since the sequence specific part of the short range interactions directly encodes the secondary structure propensities, this is not surprising. The underlying generic contributions to the potential of mean force applied in the reported simulations regularize and propagate the secondary structure. Consequently, the pairwise sequence specific potentials are to some extent ''interpolated'' over relatively long fragments of the model chains, providing consensus secondary propensities. The generic potential plays a similar role as ''filtering'' procedures in more sophisticated applications of computational models of neural networks for secondary structure prediction.[33–36] In principle, we could use a deterministic procedure that generates the prediction of secondary structure according to the proposed factorization of the secondary structure propensities. There is, however, an important advantage of the proposed lattice Monte Carlo model; it carries along its entire geometrical context. Thus, there are straightforward possibilities for considerable improvement of secondary structure prediction and, consequently, for prediction of tertiary structure. For example, tertiary interactions, which moderate secondary propensities, could be introduced. This method was actually employed in our earlier work in the context of a somewhat different (and less accurate) scheme for short range interactions, allowing prediction of several very simple folds of small globular proteins.[25–28]

In the forthcoming work, the various contributions to the tertiary (long range) interactions and the effect on the protein folding process will be examined in the context of the present model of short range interactions and polypeptide chain dynamics.

ACKNOWLEDGMENTS

This work was supported in part by the NIH Grant No. GM37408, by NIH Forgarty International Collaboration Grant No. 1 R03 TW00418, and by University of Warsaw Grant No. BST 472-34/94. Helpful discussions with Dr. Adam Godzik are gratefully acknowledged.

[1] T. E. Creighton, in *Proteins: Structure and Molecular Properties* (Freeman, San Francisco, 1984).
[2] C. B. Anfinsen, Science **181**, 223 (1973).
[3] M. Levitt, Curr. Opinion Struct. Biol. **1**, 224 (1991).
[4] K. A. Dill, Curr. Opinion Struct. Biol. **3**, 99 (1993).
[5] M. Karplus and E. Shakhnovich, *Protein Folding* (Freeman, New York, 1992), pp. 127–195.
[6] C. Wilson and S. Doniach, Proteins **6**, 193 (1989).
[7] M. Levitt and A. Warshel, Nature **253**, 694 (1975).
[8] A. T. Hagler and B. Honig, Proc. Natl. Acad. Sci. USA **75**, 554 (1978).
[9] I. D. Kuntz, G. M. Crippen, P. A. Kollman, and D. Kimelman, J. Mol. Biol. **106**, 983 (1976).
[10] J. Skolnick and A. Kolinski, Annu. Rev. Phys. Chem. **40**, 207 (1989).
[11] J. Skolnick and A. Kolinski, Science **250**, 1121 (1990).
[12] A. Godzik, A. Kolinski, and J. Skolnick, J. Comp. Aided Mol. Des. **7**, 397 (1993).
[13] A. Godzik, A. Kolinski, and J. Skolnick, J. Mol. Biol. **227**, 227 (1992).

[14] D. G. Covell, Proteins **14**, 409 (1992).

[15] A. Kolinski, A. Godzik, and J. Skolnick, J. Chem. Phys. **98**, 7420 (1993).

[16] J. Skolnick, A. Kolinski, and R. Yaris, Proc. Natl. Acad. Sci. USA **86**, 1229 (1989).

[17] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).

[18] H. S. Chan and K. A. Dill, J. Chem. Phys. **99**, 2116 (1993).

[19] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. USA **90**, 7195 (1993).

[20] A. Sali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).

[21] M-H. Hao and H. A. Scheraga, J. Phys. Chem. **98**, 4940 (1994).

[22] J. D. Honeycutt and D. Thirumalai, Biopolym. **32**, 695 (1992).

[23] A. Godzik, J. Skolnick, and A. Kolinski, Proc. Natl. Acad. Sci. USA **89**, 2629 (1992).

[24] A. Godzik, A. Kolinski, and J. Skolnick, J. Comput. Chem. **14**, 1194 (1993).

[25] A. Kolinski and J. Skolnick, Proteins **18**, 338 (1994).

[26] A. Kolinski and J. Skolnick, Proteins **18**, 353 (1994).

[27] M. Vieth, A. Kolinski, C. L. Brooks III, and J. Skolnick, J. Mol. Biol. **237**, 361 (1994).

[28] J. Skolnick, A. Kolinski, C. L. Brooks III, A. Godzik, and A. Rey, Current Biol. **3**, 414 (1993).

[29] P. Y. Chou and G. D. Fasman, Adv. Enzymol. **47**, 45 (1978).

[30] J. Garnier, D. Osguthorpe, and B. Robson, J. Mol. Biol. **120**, 97 (1978).

[31] F. Maxfield and H. Scheraga, Biochem. **18**, 697 (1979).

[32] L. H. Howard and M. Karplus, Proc. Natl. Acad. Sci. USA **86**, 152 (1989).

[33] D. G. Kneller, F. G. Cohen, and R. Langridge, J. Mol. Biol. **214**, 171 (1990).

[34] X. Zhang, J. P. Mesirov, and D. L. Waltz, J. Mol. Biol. **225**, 1049 (1992).

[35] B. Rost and C. Sander, Proteins **19**, 55 (1994).

[36] M. Vieth and A. Kolinski, Acta Biochem. Pol. **38**, 335 (1991).

[37] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Simanouchi, and M. Tasumi, J. Mol. Biol. **112**, 535 (1977).

[38] PDB Q. Newsletter **63** (1993).

[39] G. N. Ramachandran and V. Sasisekharan, Adv. Protein Chem. **23**, 283 (1968).

[40] T. J. Oldfield and R. E. Hubbard, Proteins **18**, 324 (1994).

[41] A. Baumgartner, Annu. Rev. Phys. Chem. **35**, 419 (1984).

[42] A. Kolinski and J. Skolnick, J. Chem. Phys. **97**, 9412 (1992).

[43] *Monte Carlo Methods in Statistical Physics*, edited by K. Binder (Springer, Berlin, 1986).

[44] T. L. Hill, *An Introduction to Statistical Thermodynamics* (Dover, New York, 1960).

[45] H. J. Dyson and P. E. Wright, Current Opinion Struct. Biol. **3**, 60 (1993).

[46] A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore, Science **253**, 657 (1991).

[47] A. Kolinski and J. Skolnick, *Parameters of statistical potential*. Available by ftp from public directory: scripps.edu (pub/andr/MCSP) 1995.

[48] P. E. Correa, Proteins **7**, 366 (1990).

[49] M. Classens, E. van Custem, I. Lasters, and S. Wodak, Protein Eng. **2**, 335 (1989).

[50] L. Holm and C. Sander, J. Mol. Biol. **218**, 183 (1991).

[51] M. Levitt, J. Mol. Biol. **226**, 507 (1992).

[52] A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, Protein Sci. **2**, 1697 (1993).

[53] P. W. Payne, Protein Sci. **2**, 315 (1993).

[54] S. Rackowsky, Proteins **7**, 378 (1990).

[55] L. Reid and J. Thornton, Proteins **5**, 170 (1990).

[56] A. Rey and J. Skolnick, J. Comput. Chem. **13**, 443 (1992).

[57] M. Milik, A. Kolinski, and J. Skolnick (unpublished).

[58] W. Kabsch and C. Sander, Biopolym. **22**, 2577 (1983).

[59] P. G. De Gennes, *Scaling Concepts in Polymer Physics* (Cornell, Ithaca, 1979).

[60] R. S. DeWitte and E. I. Shakhnovich, Protein Sci. **3**, 1570 (1994).