Contact interactions method: A new algorithm for protein folding simulations

LUCIO TOMA¹ AND SALVATORE TOMA²

¹ Dipartimento di Chimica Organica, Università di Pavia, Via Taramelli 10, 27100 Pavia, Italy

² Molecular Biology, BioScience Center, R&D/Biopharmaceuticals, Pharmacia, Via per Pogliano, 20014 Nerviano (MI), Italy

(RECEIVED September 12, 1995; ACCEPTED October 16, 1995)

Abstract

Computer simulations of simple exact lattice models are an aid in the study of protein folding process; they have sometimes resulted in predictions experimentally proved. The contact interactions (CI) method is here proposed as a new algorithm for the conformational search in the low-energy regions of protein chains modeled as copolymers of hydrophobic and polar monomers configured as self-avoiding walks on square or cubic lattices. It may be regarded as an extension of the standard Monte Carlo method improved by the concept of cooperativity deriving from nonlocal contact interactions. A major difference with respect to other algorithms is that criteria for the acceptance of new conformations generated during the simulations are not based on the energy of the entire molecule, but cooling factors associated with each residue define regions of the model protein with higher or lower mobility. Nine sequences of length ranging from 20 to 64 residues were used on the square lattice and 15 sequences of length ranging from 46 to 136 residues were used on the cubic lattice. The CI algorithm proved very efficient both in two and three dimensions, and allowed us to localize energy minima not localized by other searching algorithms described in the literature. Use of this algorithm is not limited to the conformational search, because it allows the exploration of thermodynamic and kinetic behavior of model protein chains.

Keywords: conformational search algorithm; hydrophobic interactions; lattice models; protein folding

The determination of the folding pathway followed by a protein in the conformational space from the denaturated to the native structure is still an unsolved problem. Its difficulty is mainly due to the enormous extension of the conformational space and to its exponential growth when the length of the protein chain increases. How a protein solves the Levinthal paradox (Levinthal, 1968), i.e., how it folds to its native structure without an exhaustive search in the conformational space, has been largely debated in literature (for some recent papers, see for example, Fiebig & Dill, 1993; Abkevich et al., 1994a; Covell, 1994; Dill et al., 1995). The nature of the driving forces inducing the collapse to more compact structures has been discussed and the importance of cooperativity in protein folding evidenced (Dill et al., 1993).

Models of globular proteins have been used to get a better and better understanding of these phenomena, from the very simple 2D square lattice or 3D cubic lattice to off-lattice, all-heavy-

atoms, or all-atoms models (Hunt et al., 1994; Šali et al., 1994; Yee et al., 1994; Dill et al., 1995 and references therein cited; Gutin et al., 1995a, 1995b). With the aid of a cubic lattice model, Dill et al. (1993) demonstrated the importance of nonlocal contact interactions and formulated the hydrophobic zipper hypothesis. This hypothesis derives from the concept of cooperativity, i.e., the probability that a chain enters a particular conformation is increased if it had previously been in some particular other predecessor conformations. In consequence of cooperativity, a globally optimal state can be found without an exhaustive search.

The square and cubic lattice HP model maintains only a limited but properly selected number of features of the protein molecules (Dill et al., 1995). It consists of a linear chain molecule configured as a self-avoiding walk (SAW) on a 2D or 3D lattice and takes into account only the hydrophobic interaction as the main driving force in protein folding. In this model, the protein chain is composed of a specific sequence of only two types of amino acids, H (hydrophobic) or P (nonhydrophobic). The sequence is folded on the 2D or 3D lattice; each monomer occupies one lattice site, connected to its chain neighbors and unable to occupy a site filled by any other residue. At each point, the

Reprint requests to: Lucio Toma, Dipartimento di Chimica Organica, Università di Pavia, Via Taramelli 10, 27100 Pavia, Italy; e-mail: burdisso@ipv85.unipv.it.

chain can turn 90° left or right (and up or down in the case of the 3D lattice) or can continue ahead. The energy of the molecule is determined by the summation of the favorable energy contributions of -1 units deriving from each contact between two nonbonded hydrophobic-hydrophobic (HH) amino acids occupying neighboring nondiagonal lattice points. No energy contribution is given by HP or PP contacts. Under these conditions, low-energy conformations are compact with the H residues mainly in the core of the molecule and the P residues outside.

The search in the conformational space of such a kind of model has been performed with the use of different algorithms. The standard Monte Carlo (MC) method (Metropolis et al., 1953) and the Genetic Algorithms (GA) (Sun, 1993; Dandekar & Argos, 1994) have been tested comparatively (Unger & Moulton, 1993) on a 2D square lattice employing pivot moves (Stellman & Gans, 1972; Madras & Sokal, 1988) for a certain number of sequences with a length ranging from 20 to 64 residues; their capability to find low-energy conformers and the global minima in reasonable times has been compared. The GA method proved largely superior in its performances (Unger & Moulton, 1993).

The search of low-energy conformations of sequences modeled as SAWs on a 3D cubic lattice has also been performed with a different approach, named "hydrophobic zipper" (HZ) strategy (Dill et al., 1993), derived from the hydrophobic zipper hypothesis. It is an algorithm that sequentially assembles hydrophobic contacts, leading to a compact chain conformation with one or several hydrophobic cores.

We describe here an alternative procedure, the Contact Interactions (CI) algorithm, that is an attempt to combine the best features of the above cited methods. On the one hand, it relies heavily on the concept of cooperativity, which is the basis of the HZ strategy, and on the other hand, it uses an MC algorithm. Following the idea to assign different conformational freedom to the different residues in the chain, a method was obtained that is even more efficient than GA and HZ in the search of the low-energy regions of the conformational space of sequences modeled as SAWs both on 2D and 3D lattices.

Model

Unger and Moulton (1993) compared the MC and GA methods employing pivot moves. Both methods simulate the evolution of an HP sequence in the conformational space by deriving a new conformation from one or two predecessor conformations. The new conformation is accepted if some conditions are satisfied or it is discarded. In both methods, the criterion of acceptance of a new conformer is based on the energy of the entire molecule.

For example, in the MC method: E_1 is the energy of the old conformation and E_2 is the energy of the new conformation derived from it through a transformation; if $E_2 \leq E_1$, the new conformation is accepted; if $E_2 > E_1$, a criterion of acceptance is applied in the form:

$$\text{Rnd} < \exp[(E_1 - E_2)/c_k],$$

where Rnd is a random number between 0 and 1 and c_k is the "temperature."

Analogously, GA uses the similar criterion of acceptance:

$$\text{Rnd} < \exp[(\bar{E}_{ij} - E_k)/c_k],$$

where E_k is the energy of the new conformation and $\bar{E}_{ij} = (E_i + E_j)/2$, E_i and E_j being the energies of the "parent" conformations from which the new conformer was derived through a "cut and paste" procedure.

These approaches allow the molecule to explore the conformational space making, but also breaking, hydrophobic contacts; according to the temperature of the simulation and to the shape of the potential energy surface, the low-energy regions of the surface are more or less efficiently explored. However, these methods consider the molecule as a whole, do not take into explicit consideration the cooperativity of folding process, and do not put in evidence the real role of the contact interactions among monomers that can be distant in the sequence but that are brought together by earlier events in the folding process.

On the contrary, the HZ strategy gives a prominent role to the contact interactions as it generates conformations according to a sequential assembly of hydrophobic contacts. However, the method suffers, as its main drawback, from the fact that, once a HH contact has been formed, it cannot be broken. The "cascade" that forms the HH contacts is monodirectional and ends when no further contact can be established. Thousands of cascades are necessary to locate global minima or energy minima close to them.

Our approach is an attempt to combine the best features of the MC and GA methods with the concept of cooperativity used in the HZ strategy. It relies on the consideration that, when an HH contact is established between the i th and the j th residue, this creates heterogeneity in the conformational freedom of the different residues of the molecule; all the residues in the loop from i to j have less mobility than the residues outside them (Fig. 1).

This phenomenon is implemented in the following strategy here described in the 2D space: given a protein composed of L residues with a specific sequence of H and P residues, $L - 2$ independent variables are necessary to describe the conformation of the molecule. Whatever position the first and the second residue occupy, the position of the third residue can be described by a coordinate $\theta(3)$ that can assume only the three values $1/0/-1$, according to the fact that the third residue turns left,

Fig. 1. Effects of an HH contact on the mobility of the residues in different portions of an HP sequence. All the residues in the loop defined by the hydrophobic interaction are less prone to change conformation. Black and white boxes represent H (hydrophobic) and P (polar) monomers, respectively.

goes ahead, or turns right with respect to the preceding residues. Similarly, the position of the fourth residue can be described by $\theta(4)$ with reference to the second and third residues, and so on. To each residue from the third to the last one, the CI method associates a value, $f(i)$, that describes its mobility, i.e., the probability for $\theta(i)$ to change. For a residue i whose movement is completely free, $f(i) = 0$; when the same residue is involved in a loop defined by an HH contact and consisting of k bonds, $f(i)$ assumes a negative value and it is lowered by an amount determined by a function $g(k)$, i.e., the i th residue is “cooled” by $g(k)$. If the simulation process begins with an extended molecule, initial values are $E = 0$, $\theta(i) = 0$ and $f(i) = 0$ because there is no HH contact. In a folded conformation in which x HH contacts are present, the energy of the entire molecule is $E = -x$ units, but this energy does not determine the behavior of the molecule. Actually, the molecule evolves according to the values $f(i)$ ’s associated to each residue. The values of $f(i)$ are given by $g(k)$, which depends on the dimensions of the HH loops in which the i th residue is involved. When a residue is involved in more than one loop, the contributions $g(k)$ deriving from the different loops are additive.

See, for example, the three conformations **A**, **B**, and **C**, assumed by the sequence consisting of the seven residues HPPHPHP in Figure 2. In conformation **A**, all $f(i) = 0$, all residues have similar high mobility. In conformation **B**, an HH-loop with $k = 3$ is present; the third and the fourth residues are both cooled by $g(3)$, i.e., $f(3) = f(4) = -g(3)$; these residues are less prone to modify their orientation with respect to the first and the second ones, respectively, than the fifth with respect to the third, the sixth with respect to the fourth, etc. In conformation **C**, two HH-loops with $k = 3$ and $k = 5$, respectively, are present; so, $f(3) = f(4) = -g(3) - g(5)$ because they are involved in two loops defined by a $i/i + 3$ and a $i/i + 5$ contact; $f(5) = f(6) = -g(5)$ because they are involved in a loop defined by a $i/i + 5$ contact; $f(7) = 0$ because the seventh residue is not involved in any loop.

The following algorithm describes the steps of the CI method: (1) Start from an extended structure; (2) random choice of a residue to be moved, for example the i th; (3) use a criterion of mobility to decide if it has to be moved or not: move it if $\text{Rnd} < \exp[f(i)/c_k]$; if not, go to 2 (the criterion is always satisfied for residues not belonging to loops defined by HH contacts); (4) random choice of the movement, i.e., random choice of the value of $\theta(i)$ while taking as invariant all the other θ coordinates (this corresponds to a pivot move); (5) control of the validity of the structure deriving from the movement; if not, go to 2; (6) if the structure is valid, the new conformation is accepted and its energy evaluated; a time step is counted, the temperature is lowered according to a cooling strategy, and the function $f(i)$

assumes the values deriving from the loops present in the conformation; go to 2.

Results and discussion

First of all, our aim was to find the temperature range compatible with our model and proper values for the function $g(k)$. Initially, we chose all $g(k) = 1$. Better values for this function $g(k)$ were determined later and represented a further improvement of the model.

The 20-residue sequence (**1**) reported in Table 1 was chosen as a suitable molecule for the optimization of the system because it has a single native conformation ($E = -10$) (Fig. 3) in which HH-loops of different length are present from the smallest ($i/i + 3$) to the largest one ($i/i + 17$). Several simulations were performed at different temperatures, c_k , and indicated that there is a narrow range of temperature in which the molecule can quickly fold to its native conformation. At high temperatures, the residues move randomly and the structure does not approach the low-energy region. Low temperatures suffer from kinetic bias because the residues can move with difficulty and, once a loop is formed, it can hardly be broken. However, for temperatures in the range of $c_k = 0.2 \sim 0.5$, fast folding was observed. In order to compare quantitatively the folding ability at various temperatures in this range, we performed simulations at constant temperature on sets of 5,000 molecules. The number of time steps necessary for each molecule to reach for the first time the native conformation was evaluated, ending, however, the simulation after 10,000 time steps if the native conformation was not reached. We defined as $t_{1/2}$ the number of time steps necessary for half of the molecules to fold to the native state. In Figure 4A, $t_{1/2}$ is reported as function of c_k . A clear minimum is observed at $c_k = 0.36$ and $t_{1/2} \approx 1,700$.

Further improvement of the CI model consisted of optimization of the function $g(k)$ because it is reasonable to think that the degree of immobilization of the residues in an HH-loop depends on the dimensions of the loop, i.e., small loops are frozen to a higher degree than large ones. The smaller loop that can be formed for an SAW on a square or cubic lattice contains three bonds in consequence of the interaction between residues i and $i + 3$. Larger loops contain 5, 7, 9, etc. bonds in consequence of $i/i + 5$, $i/i + 7$, $i/i + 9$, etc. contacts. The dependence from the dimension of the loop could assume the form $g(k) = c/k$, where c is a constant and k is the already defined number of bonds in the loop. We set $c = 3$ in order to obtain $g(3) = 1$; it follows $g(5) = 3/5 = 0.6$, $g(7) = 3/7 = 0.429$, etc. When this

Fig. 2. Three possible conformations of the HPPHPHP sequence on the 2D square lattice. Conformation **A** has no HH contact; conformation **B** has one HH contact; conformation **C** has two HH contacts.

Fig. 3. Minimum energy (native) conformation of the sequence **1** on the 2D square lattice. Ten HH contacts are present.

Table 2. Comparison of Contact Interactions (CI) with Genetic Algorithm (GA) and Monte Carlo (MC) folding simulations for proteins modeled as SAWs on a 2D lattice

Sequence	MC ^a		GA ^a		CI ^b					CI ^b				
	E^c	Time ^d	E^c	Time ^d	c_k	E^c	Time ^d	Hits ^e	Duration ^f	c_k	E^c	Time ^d	Hits ^e	Duration ^f
2	-9	292,443	-9	30,492	0.300	-9	171	5	10,000	0.300	-9	171	5	10,000
3	-9	2,492,221	-9	30,491	0.300	-9	207	5	10,000	0.250	-9	1,425	5	10,000
4	-8	2,694,572	-8	20,400	0.300	-8	759	4	30,000	0.216	-8	1,132	5	30,000
5	-13	6,557,189	-14	301,339	0.300	-14	4,653	4	300,000	0.267	-14	40,237	2	300,000
6	-20	9,201,755	-22	126,547	0.300	-22	235,780	1	300,000	0.312	-23	204,928	1	300,000
7	-21	15,151,203	-21	592,887	0.300	-21	12,956	5	300,000	0.288	-21	13,464	5	300,000
8	-33	8,262,338	-34	208,781	0.300	-34	111,400	1	500,000	0.430	-35	361,533	1	500,000
9	-35	7,848,952	-37	187,393	0.300	-38	97,220	3	500,000	0.394	-40	461,099	1	500,000

^a Results from Unger and Moulton (1993).^b Our results. Each simulation was repeated five times. The cooling scheme started with the value of c_k indicated lowering it by 0.999999 each time step.^c Lowest energy value found in the most efficient of five runs.^d Number of time steps before the lowest energy value was found.^e Number of runs in the five ones in which the lowest energy value was found.^f Duration of the simulation.

and 64-residue chains having $E = -23$, -35 , and -40 , respectively. The corresponding structures are reported in Figure 5.

The folding ability of the CI algorithm was also analyzed on the cubic lattice model. The CI procedure for the conformational search on a 3D lattice is similar to that described above for a 2D lattice; the same parameters were utilized, for example, the function $g(k)$ was set to $20/(k+17)$. We tested two sequences (Table 1), the 46-residue sequence (10) and the 58-residue sequence (11) already used by Dill et al. (1993) to test hydrophobic zipper (HZ) as strategy for protein folding. By using a cooling strategy starting from $c_k = 0.6$ and progressive cooling by 0.999999 for 1,000,000 time steps, conformations with $E = -32$ and $E = -42$ for (10) and (11), respectively, were found in the best of five simulations (Fig. 6); these conformations are, respectively, three and one unit close to the reported global minima of these two HP sequences (Dill et al., 1993). It should be mentioned that, in a simulation with the CI algorithm using $g(k) = 10/(k+7)$, a conformation with $E = -34$ (Fig. 6) was obtained for (10) in the best of five runs.

The CI algorithm can maintain its searching ability also for longer chains. We modeled, using $g(k) = 20/(k+17)$, three long sequences (12), (13), and (14) (Table 1), designed by Lattman et al. (1994) in simulations of 1,000,000 time steps. Min-

ima with $E = -49$, -58 , and -65 , respectively, were localized in the best of five runs starting with $c_k = 0.3$ and lowering it by 0.9999995 each time step. Figure 7 reports 3D plots and polymer graphs (Fiebig & Dill, 1993) of these conformations. They appear as compact structures with a single hydrophobic cluster.

Recently, Yue et al. (1995) reported a blind test of lattice protein folding algorithms on 10 selected sequences of 48-residue chains. The CI algorithm was applied to such sequences in simulations of 1,000,000 time steps starting from a temperature $c_k = 0.3$ and lowering it by 0.9999998 each time step using $g(k) = 20/(k+17)$. The following minima were determined for the 10 sequences considered in the same order as in Yue et al. (1995): -32 (-31 , -32); -33 (-32 , -34); -32 (-31 , -34); -32 (-30 , -33); -32 (-30 , -32); -30 (-29 , -32); -30 (-29 , -32); -30 (-29 , -31); -32 (-31 , -34); -32 (-33 , -33); the two figures in parentheses represent, respectively, results of the HZ approach (Yue et al., 1995) and the energy of the native conformations determined with the CHCC method (Yue et al., 1995, Yue & Dill, 1995), reported here for comparison.

Conclusions

The CI method is proposed as a new algorithm for the conformational search in the low-energy regions of proteins modeled as SAWs of H and P residues on 2D and 3D lattices. It may be regarded as an extension of the standard MC method improved by the concept of cooperativity.

A major difference with respect to other methods is that criteria for the acceptance of new conformations generated during the simulations are not based on the energy of the entire molecule, but cooling factors associated with each residue define regions of the model protein with higher or lower mobility. These cooling factors determine the probability to change conformation.

The CI algorithm proved very efficient both in two and three dimensions and allowed us to localize energy minima not local-

Fig. 5. Lowest energy conformations of the sequences 6, 8, and 9 found by the CI method on the 2D square lattice.

Fig. 6. Lowest energy conformations of the sequences 10 and 11 found by the CI method on the 3D square lattice. **A:** 3D plots; black and white beads represent H (hydrophobic) and P (polar) monomers, respectively. **B:** Polymer graphs of the same conformations; the straight line with black and white beads represents the covalently linked monomers, the curved lines represent the noncovalent spatial contacts among monomers.

ized by other conformational search algorithms described in the literature. Moreover, the CI approach is easily extendable (L. Toma & S. Toma, unpubl. results) to folding simulations of heteropolymers with many different residue types (Shakhnovich & Gutin, 1993; Abkevich et al., 1994b).

In light of the parallels revealed between some results from exact lattice models and some recent experimental results (Dill et al., 1995; Onuchic et al., 1995), the good performance of the CI algorithm may prove useful in the exploration of the general principles of protein structure, stability, and folding kinetics because use of the CI algorithm allows modeling of longer sequences than the standard MC algorithm. The CI algorithm is now being tested, in the 2D lattice HP model, to analyze its capability to correctly predict thermodynamic and kinetic behavior of the folding sequences. Results obtained with some selected 18–20 monomer sequences show that it can converge to correct thermodynamics (Boltzmann distribution) in a wide range of temperatures. Also, the kinetic behavior is in accordance with

Fig. 7. Lowest energy conformations of the sequences 12, 13, and 14 found by the CI method on the 3D square lattice. **A:** 3D plots. **B:** Polymer graphs.

the results obtained by other authors (Camacho & Thirumalai, 1993; Abkevich et al., 1994b; Socci & Onuchic, 1994).

Acknowledgments

L.T. acknowledges financial support provided by the Italian Ministero dell'Università e della Ricerca Scientifica e Tecnologica, Roma.

References

- Abkevich VI, Gutin AM, Shakhnovich EI. 1994a. Specific nucleus as the transition state for protein folding: Evidence from the lattice model. *Biochemistry* 33:10026–10036.
- Abkevich VI, Gutin AM, Shakhnovich EI. 1994b. Free energy landscape for protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations. *J Chem Phys* 101:6052–6062.
- Camacho CJ, Thirumalai D. 1993. Kinetics and thermodynamics of folding in model proteins. *Proc Natl Acad Sci USA* 90:6369–6372.
- Covell DG. 1994. Lattice model simulations of polypeptide chain folding. *J Mol Biol* 235:1032–1043.
- Dandekar T, Argos P. 1994. Folding the main chain of small proteins with the genetic algorithm. *J Mol Biol* 236:844–861.

- Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS. 1995. Principles of protein folding—A perspective from simple exact models. *Protein Sci* 4:561–602.
- Dill KA, Fiebig KM, Chan HS. 1993. Cooperativity in protein-folding kinetics. *Proc Natl Acad Sci USA* 90:1942–1946.
- Fiebig KM, Dill KA. 1993. Protein core assembly processes. *J Chem Phys* 98:3475–3487.
- Gutin AM, Abkevich VI, Shakhnovich EI. 1995a. Evolution-like selection of fast-folding model proteins. *Proc Natl Acad Sci USA* 92:1282–1286.
- Gutin AM, Abkevich VI, Shakhnovich EI. 1995b. Is burst hydrophobic collapse necessary for protein folding? *Biochemistry* 34:3066–3076.
- Hunt NG, Gregoret LM, Cohen FE. 1994. The origin of protein secondary structure. Effects of packing density and hydrogen bonding studies by a fast conformational search. *J Mol Biol* 241:214–225.
- Lattman EE, Fiebig KM, Dill KA. 1994. Modeling compact denaturated states of proteins. *Biochemistry* 33:6158–6166.
- Levinthal C. 1968. Are there pathways for protein folding? *J Chem Phys* 65:44–45.
- Madras N, Sokal AD. 1988. The pivot algorithm: A highly efficient Monte Carlo method for self-avoiding walk. *J Stat Phys* 50:109–186.
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machine. *J Chem Phys* 21:1087–1092.
- Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND. 1995. Toward an outline of the topography of a realistic protein-folding funnel. *Proc Natl Acad Sci USA* 92:3626–3630.
- Šali A, Shakhnovich EI, Karplus M. 1994. Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. *J Mol Biol* 235:1614–1636.
- Shakhnovich EI, Gutin AM. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proc Natl Acad Sci USA* 90:7195–7199.
- Socci ND, Onuchic JN. 1994. Folding kinetics of proteinlike heteropolymers. *J Chem Phys* 101:1519–1528.
- Stellman SD, Gans PJ. 1972. Efficient computer simulation of polymer conformation. I. Geometric properties of hard-sphere model. *Macromolecules* 5:516–526.
- Sun S. 1993. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Sci* 2:762–785.
- Unger R, Moult J. 1993. Genetic algorithms for protein folding simulations. *J Mol Biol* 231:75–81.
- Yee DP, Chan HS, Havel TF, Dill KA. 1994. Does compactness induce secondary structure in proteins? A study of poly-alanine chains computed by distance geometry. *J Mol Biol* 241:557–573.
- Yue K, Dill KA. 1995. Forces of tertiary structural organization of globular proteins. *Proc Natl Acad Sci USA* 92:146–150.
- Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA. 1995. A test of lattice protein folding algorithms. *Proc Natl Acad Sci USA* 92:325–329.