Protein Folding Simulation by Particle Swarm Optimization

M. Meissner and G. Schneider*

Johann Wolfgang Goethe-University, Chair for Chem- & Bioinformatics, Frankfurt am Main, Germany

Abstract: This work introduces Particle Swarm Optimization (PSO) to protein structure prediction as a new field of application. Finding the global optimum in the free energy landscape of protein structures is a challenging, non-trivial task and has been subject of research for decades, resulting in many different approaches and methods until today. Here we show that a standard implementation of PSO is capable of optimizing backbone geometries and generating good solutions in refolding studies, yielding near native structures for two small sample proteins. We present a straightforward approach to include secondary structure information in the optimization process and show that results improve. Finally, a first predicted structure from *ab initio* folding by PSO is shown where native topology could be captured with a basic energy function, giving a promising outlook on future research.

INTRODUCTION

For many small proteins and protein domains, prediction of their three-dimensional (3D) or "tertiary" structure from the amino acid sequence ("primary" structure) should be feasible, as many such proteins fold and re-fold independently and spontaneously in an aqueous environment [1, 2]. For example, single-domain proteins adopt their native conformation typically on a millisecond time scale. During the past 40 years, numerous attempts have been made that consider protein folding an optimization problem [3]. The true challenge is *ab initio* folding of an amino acid sequence, either by simulating the folding dynamics or direct structure prediction without knowing the native state. Because of the many degrees of freedom of an amino acid sequence, this cannot be achieved by exhaustive evaluation of all theoretically possible conformations available to a given protein (Levinthal paradox). It is generally assumed that protein folding dynamics follow a directed process rather than random sampling [4]. Various potential functions and prediction methods have been developed for the purpose of protein structure prediction and simulation of folding dynamics [5], and tertiary structure prediction has become feasible for an increasing number of sequences [3, 6]. The energy functions used as quality criteria for folded protein states may also be used to analyze the potential energy landscape for solved protein folding problems [7].

In this study, we demonstrate that particle swarm optimization (PSO) [8, 9] can be applied to predict the tertiary structure of a protein backbone in re-folding experiments. In this setting, an experimentally determined protein tertiary structure serves as the reference for optimization of a random coil conformation or heterogeneous denatured state ensemble of the same protein [10, 11]. In other words, by re-folding a successful optimization process converges at a known optimum. PSO seems an intuitive choice for this purpose as the "native state" of a protein tertiary structure essentially represents an ensemble of low-energy structures [12, 13], which can be mimicked by a swarm population of individual protein backbone conformations. The aim of our study was not to perform realistic forward folding simulation but to introduce PSO as an optimization technique for protein structure prediction. We restrict our discussion to small proteins that fold spontaneously and independently from molecular "chaperones" [14]. We show that PSO yields native-like conformations of simplified protein backbone models in re-folding experiments, and leads to properly folded conformations in a prospective setting using a knowledge-based potential function.

*Address correspondence to this author at the Johann Wolfgang Goethe-University, Chair for Chem- & Bioinformatics, Frankfurt am Main, Germany; E-mail: gisbert.schneider@modlab.de

METHODS

Particle swarm optimization (PSO). In this study, we employed a common variant of PSO, the constriction-type PSO [15]. In the following, we refer to this PSO implementation as "CPSO". Here, each particle was initialized at a random position in search space. The position of particle i is given by the vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ where D is the dimensionality of the problem. Its velocity is given by the vector $v_i = (v_{i1}, v_{i2}, ..., v_{iD})$.

The applied variant of constriction-type PSO, implemented as in reference [16], contains of two kinds of memory that influence the movement of the particles: In the "cognitive memory" $p_i = (p_{i1}, p_{i2}, ..., p_{iD})$ the best previous position is stored that was visited by each individual particle i. The vector $p_{best} = (p_{best1}, p_{best2}, ..., p_{bestD})$, also called "social memory", contains the position of the best point in search space visited by all swarm particles so far. By sharing this information within the swarm population, particles can imitate more successful individuals and improve their own fitness. Over optimization time, the swarm converges in an optimum after having explored the search space.

In each epoch, the velocity of every particle is updated according to Eq. (1):

$$\mathrm{v}_i(t+1) = K \cdot (\mathrm{v}_i(t) + n_1 \cdot r_1 \cdot (p_i - x_i(t)) + n_2 \cdot r_2 \cdot (p_{best} - x_i(t))) \quad (1)$$

where n_1 and n_2 are positive constants called "cognitive" and "social" parameters that weight the influence of the two types of swarm memory; r_1 and r_2 are pseudo-random numbers in [0,1]; K is the constriction factor as defined by Eq. (2).

1874-1991/07

$$K = \frac{2}{|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}|} \quad (2)$$

where $\varphi = n_1 + n_2, \varphi > 4$

For all experiments, we applied $n_1 = n_2 = 2.05$, as recommended by others (e.g. [17]).

The constriction factor *K* controls the magnitude of the particle velocity and can be seen as a dampening factor. It provides the algorithm with two important features [18]: First, it often leads to a faster convergence than standard PSO [18]. Second, the swarm maintains its ability to perform broad movements in search space even when convergence is already advanced but a new optimum is found. CPSO has a potential ability to avoid being trapped into local optima while possessing a fast convergence capability [18].

In our experiments a restriction constant V_{max} was applied to control the maximal velocity of the particles. Velocities exceeding the threshold set by V_{max} were set back to the threshold. Due to our search space (angular degrees), V_{max} was set to 180.

Based on the velocity vector the positions of the particles were updated according to Eq. (3):

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (3)$$

We employed a maximum number of 2000 epochs as termination condition for the algorithm and chose a swarm size of 20 particles.

Particle initialization. In *ab initio* protein structure prediction, every swarm particle represents a distinct backbone conformation. For our simulations we chose a coarse-grained "beads-on-a-string" backbone model, which has been used in various previous studies [19, 20, 21]. The geometry of the protein is hereby described by just two free parameters, the phi (Φ) and psi (Ψ) torsion angles (Fig. **1**). According to this, the number of dimensions *D* to optimize is for each particle is given by Eq. (4):

$$D = N \times 2 - 2 \quad (4)$$

where *N* is the number of amino acids. The first (N-terminal) and last (C-terminal) residue contains only one torsion angle. Each dimension was initialized in the range of [-180, 180], unless secondary structure constraints were exerted.

In order to test whether the inclusion of predicted secondary structure information can reduce the complexity of the search, the following constraints to the dimension ranges were included:

α-helical regions: Φ in [-65, -50], Ψ in [-55, -40].

ß-strand regions: Φ in [-120, -110], Ψ in [125, 135].

Since these restrictions in structural space are based on secondary structure predictions, there is the possibility that important regions in the solutions landscape are not sampled in case of incorrect secondary structure assignments. We used the *PredictProtein* web server (www.predictprotein.org; version May 2007) [22] to predict secondary structure (α-helix, ß-strand) from the amino acid sequence. Note that only a limited degree of flexibility that can occur in secondary structure elements is taken into account with the above defined ranges.

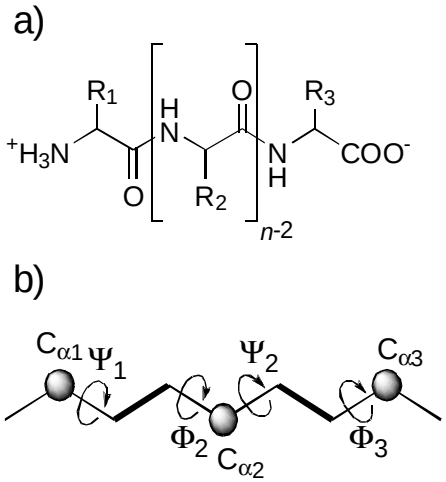

Fig. (1). Protein backbone architecture. Proteins are biopolymers consisting of amino acids, which are connected *via* planar amide bonds (C=O-NH). (**a**) The amino acid side chains are neglected in the beads-on-a-string model of a protein backbone. (**b**) Here, only two parameters determine the fold, the phi (Φ) and psi (Ψ) torsion angles. The planar peptide bond is shown as a bold line. In our study, peptide bonds were considered to be in all-*trans* conformation, and idealized backbone angles and bond lengths were used for structure generation. C_α atoms were used for structure alignment and calculation of the *rmsd* fitness function (see text).

Two small proteins were used as examples, their experimentally determined atom coordinates were taken from the Protein Data Bank (PDB) [23]: i) a sub-domain from chicken protein Villin containing 35 amino acid residues with an all-α-helix fold (PDB identifier: 1vii, NMR model 1) [24], and ii) the immunoglobulin binding domain of streptococcal protein G encompassing 56 amino acid residues with a mixed alpha-beta (ubiquitin-like) fold (PDB identifier: 2gb1, NMR model 1) [25]. Sequences and secondary structure predictions are shown in Fig. (**2**).

Scoring function for protein backbone folding. We performed protein re-folding simulations using the root mean square deviation (*rmsd*) between the actual backbone structure represented as a swarm particle and an experimentally determined structure (from the PDB; interpreted as the "native" structure) as fitness function (Eq. 5).

$$rmsd = \sqrt{\frac{\sum_i d_i^2}{n}} \quad (5)$$

where d_i gives the spatial distance (in Å) between two corresponding backbone atoms of residue *i*, and *n* is the total number of atom-pairs considered. Each generated structure is aligned to the native structure, and the *rmsd* is calculated over all C_α-backbone atoms (Fig. **2b**). We used the *rmsd* as the most obvious scoring function, in order to assess the ability of CPSO to find a good set of torsion angles that models the native geometry best.

Since the *rmsd* value alone can only be taken as a rough measure of fold similarity [26], we also calculated the ratio of the C_α contacts found within a fixed range of 8 Å in the generated structure and the contacts found in the native structure. The distribution of contacts present in a planar

1vii
MLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF
---HHHHHHH----HHHHHH--HHHHHHHHHH----

2gb1
MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
-EEEEEE------EEEEEE---HHHHHHHHHHHH------EEEEE----EEEEE-

Fig. (2). Amino acid sequences of the two model proteins used in this study (PDB identifiers 1vii and 2gb1). Secondary structure predictions are given below each residue (H: helix, E: sheet). Underlined residues indicate actual, experimentally observed secondary structure elements.

backbone chain is used as reference (background) state, so only contacts formed through the protein fold are considered (Eq. (6)).

$$ratio = \frac{nativelike^2}{native \cdot observed} \quad (6)$$

In some of our folding studies, two additional scores were applied: To favor a globular overall shape, a "confinement score" [27] was employed. This score disfavors structures with a radius of gyration (R_g) above a certain expected value typical for a globular protein with a similar number of residues (Eq. (7)).

$$Confinement = \begin{pmatrix} 0, & R_g \le R_{g-glob} \\ \left(R_g - R_{g-glob}\right)^2, & R_g > R_{g-glob} \end{pmatrix} \quad (7)$$

where R_g is the radius of gyration of the template structure and $R_{g\text{-}glob}$ is the expected radius of gyration (Eq. (8)):

$$R_{g-glob} = 2.5 \times N^{0.34} \quad (8)$$

We used the same prefactor value as Fleming *et al.* [20], thus putting slightly more pressure on globularity than in the original confinement score (prefactor = 2.83).

To avoid steric clashes between the modeled backbone atoms, the "soft-debump potential" described by Gong *et al.* [27] was applied. This potential acts as soft-sphere potential when the inter-atomic distance is smaller than the sum of the respective atoms van der Waals radii, otherwise a *pseudo*-energy of zero is applied (behavior of a hard-sphere potential).

To estimate the energy of the generated structures in the preliminary prediction study, a basic statistical contact potential [28, 29] was calculated from a set of PDB-chains taken from the *PDBselect* list [30] of July 2005. A radius of 5 Å between the van der Waals surfaces of the atoms was chosen, all inter-atomic distances between C_α-atoms below this threshold were defined as "contacts". By invoking the inverse Boltzmann approach [28], a *pseudo*-energy is calculated for a new structure, based on the present contacts.

In the prediction run, we used the weighted sum of the *pseudo*-energy, the confinement score and the score from the soft-debump function as energy function.

RESULTS AND DISCUSSION

In our first study, we wanted to assess the ability of CPSO to tune each backbone torsion angle such that the optimized structure resembles the native one best in terms of C_α-backbone atoms. We also compared the CPSO performance in simulations with and without the inclusion of information from secondary structure prediction. For each approach, 50 separate optimization runs were performed. 2,000 iterations and a swarm size of 20 particles were chosen for optimization. The mean *rmsd* value, standard deviation and minimum *rmsd* from the best run for both approaches are listed in Table **1**.

Table 1. Results of CPSO-Based Protein Folding Simulation with the *rmsd* (Å) Fitness Function

	With Secondary Structure		**Without Secondary Structure**	
Protein identifier	1vii	2gb1	1vii	2gb1
Mean *rmsd*	1.61	2.10	2.48	2.80
Standard deviation	0.53	0.37	0.30	0.30
Minimum *rmsd* (best structure)	0.79	1.17	1.77	2.22

In both cases, CPSO was able to converge close to the native conformation. The inclusion of predicted secondary structure elements yielded lower *rmsd* values, as expected. The restriction of conformational space improves the mean *rmsd*. Also, the best structure generated under consideration of the secondary structure predictions is closer to the native structure than the best structure out of the runs without any restraints. Fig. (**3**) shows the best results for both approaches. This outcome demonstrates the usefulness of CPSO for protein folding simulation.

One should bear in mind that the use of idealized backbone angles and bond lengths introduces a systematic error to the simulation that makes exact refolding (*rmsd* = 0) unlikely and can introduce relatively large *rmsd* aberrations [31]. This is one probable explanation why no CPSO run came closer to the perfect alignment with *rmsd* ~ 0.

Then, we analyzed swarm behavior for different numbers of particles. 50 independent CPSO runs over 2,000 iterations

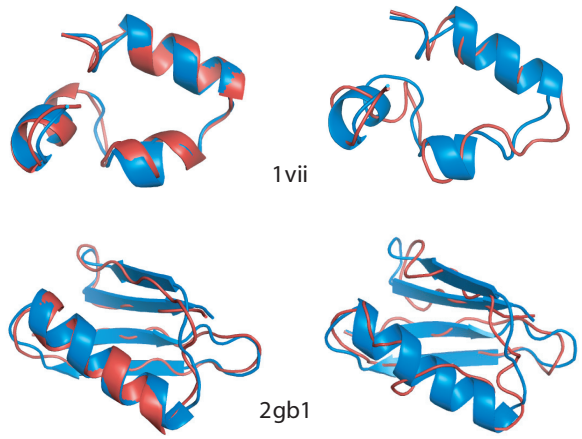

Fig. (3). Superimposed conformations of the best simulated structures (red) with the reference conformation ("native" structure) in blue. On the left, the results from folding with secondary structure constraints are shown, the simulated conformations without secondary structure are shown on the right. Here, only approximative helical elements have evolved, which is a consequence of the lacking consideration of any terms accounting for interaction energy.

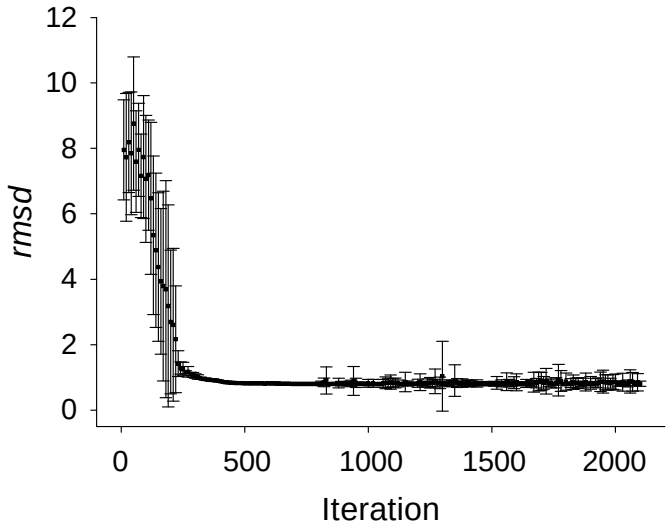

Fig. (4). Course of structure optimization by CPSO for the best final conformation of model protein 1vii. Average *rmsd* values and standard deviations of the swarm population (20 particles) are shown.

were performed with 10, 20, 30, 50, and 100 particles, respectively. Again, reference structure 1vii served as model protein. We found no striking dependency of the average *rmsd* value on the swarm size (Fig. **5**). Runs with secondary structure constraints yielded lower average *rmsd* values than the CPSO without secondary structure constraints in all runs. Notably, overall swarm diversity (expressed as *rmsd* standard deviation) was lower in the unconstrained runs (Fig. (**5**), open circles). This reflects a more localized, fine-grained search compared to conformation sampling with fixed secondary structure elements.

Fig. (**6**) shows the correlation between *rmsd* values and the ratio of correct (native) contacts for both simulated structures. For *rmsd* values < 2 Å (1vii) and < 4 Å (2gb1) a correlation is visible, while there seems to be little or no correlation for greater *rmsd* values. This suggests that below these thresholds the simulated protein conformations are native-like. This result is in agreement with *ab initio* folding simulations performed by Shakhnovich and coworkers, who determined an *rmsd* threshold of 2-6 Å for the lowest energy structures of a representative set of proteins, irrespective of their structural classes [32].

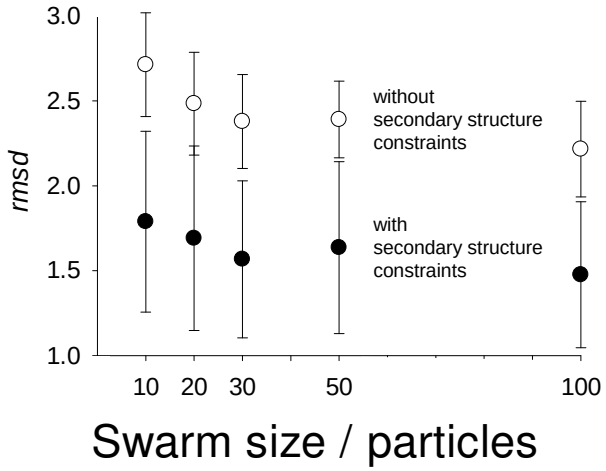

Fig. (5). Dependency of the optimization result (average *rmsd* value) on the swarm size (number of particles) for folding of protein 1vii. Error bars give standard deviations calculated from 50 independent CPSO runs.

Finally, *ab initio* folding of 1vii was performed with a simple *pseudo*-energy function and inclusion of the predicted secondary structure regions. A weighted sum out of the *pseudo*-energy and two terms punishing steric clashes and deviations from an expected radius of gyration was computed. No further optimization of these energy functions was performed; this is subject of ongoing research. We present here one example of a successful optimization run, where topology of the structure 1vii was correctly captured, resulting in a final *rmsd* ~ 4 Å (Fig. **7**). One out of the three α-helical elements exhibits an angular shift, while the other helices are accurately aligned to the native structure. A major fraction of the *rmsd* error origins from derivations in the loop regions as well as on the N- and C-terminal ends of the sequence.

This result is overall promising, but requires further investigation. It is reasonable to assume that more sophisticated scoring functions will improve the outcome of the simulations, and work on this field is in progress. CPSO also requires extensive testing with larger proteins. A major limitation of the scoring function and protein representation used in this study is the total negligence of amino acid side chain interactions [2]. From the simulations with predicted and fixed secondary structure elements one can see that hydrogen bonding between backbone atoms must be taken into account to achieve realistic folded structures. Our simplifying beads-on-a-string representation does not allow for this. Another shortcoming is the imperfect correlation between score and degree of "nativeness" modeled by the scoring function used here. It is evident that with increasing correlation between these two parameters, the optimization should perform better in *ab initio* folding.

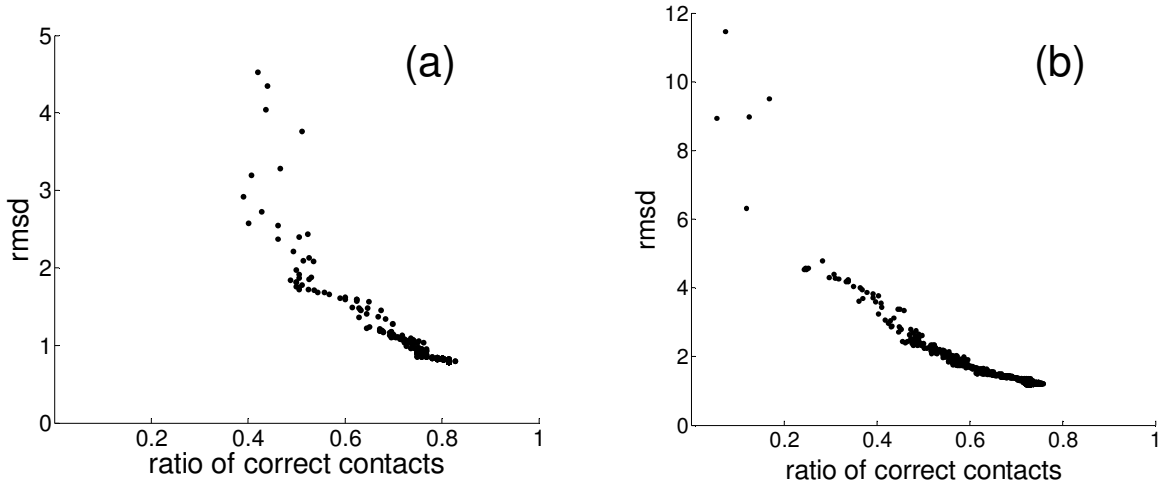

Fig. (6). Correlation between *rmsd* value and the ratio of correct contacts found in the template structures 1vii (**a**) and 2gb1 (**b**).

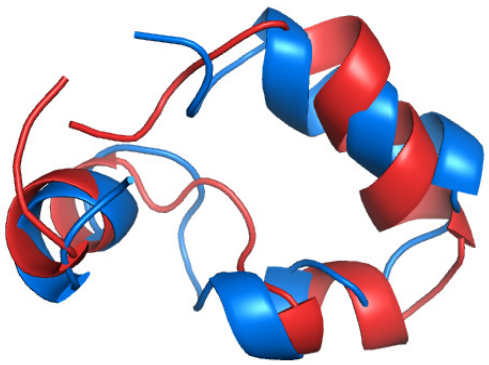

Fig. (7).

CONCLUSIONS

In this work we have introduced PSO as a heuristic to the problem of simulated protein folding and structure prediction. We have demonstrated that CPSO is feasible of successfully tuning protein backbone torsion angles in refolding experiments, producing near-native structures. This implies that with the use of an appropriate energy function *ab initio* protein structure prediction should be feasible. Rigorous statistics for the *ab initio* prediction of a set of small protein structures are required to probe the usefulness of PSO in applied protein structure prediction, along with a comparison of the method to the performance of other common heuristics like Monte Carlo sampling or simulated annealing [32, 33]. Also, modifications to our naïve and straightforward approach might bring improvement. One possible working point for improvements is the way how search space is represented and dealt with. Search space can be limited in more ways than just by including secondary structure constraints, For example, each residue can usually only occupy a limited region in torsion angle space, and by incorporation of this information an additional reduction of search space might be achieved. We are currently working on the development of such more advanced models for protein folding by PSO. Irrespective of the outcome of these particular studies, the amalgamation of PSO and protein folding simulation should provide an interesting alternative to existing methods that are commonly used.

ACKNOWLEDGEMENT

This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt am Main.

REFERENCES

[1] Anfinsen CB. Principles that govern the folding of protein chains. Science 1973; 181: 223-30.

[2] Rose GD, Fleming PJ, Banavar JR, Maritan A. A backbone-based theory of protein folding. Proc Natl Acad Sci USA 2006; 103: 16623-33.

[3] Moult J. Rigorous performance evaluation in protein structure modelling and implications for computational biology. Philos Trans R Soc Lond B Biol Sci 2006; 361: 453-8.

[4] Levinthal, C. Are there pathways for protein folding? Journal de Chimie Physique et de Physico-Chimie Biologique 1968; 65: 44-5.

[5] Buchete NV, Straub JE, Thirumalai D. Development of novel statistical potentials for protein fold recognition. Curr Opin Struct Biol 2004; 14: 225-32.

[6] Petrey D, Honig B. Protein structure prediction: inroads to biology. Mol Cell 2005; 20: 811-9.

[7] Wolynes PG. Energy landscapes and solved protein-folding problems. Philos Transact A Math Phys Eng Sci 2005; 363: 453-64.

[8] Kennedy J, Eberhart RC. Particle swarm optimization. Proceedings of IEEE International Conference on Neural Networks; Piscataway, NJ. 1995; 1942-8.

[9] Eberhart RC, Kennedy J. A new optimizer using particle swarm theory. Proceedings of the Sixth International Symposium on Micromachine and Human Science; Nagoya, Japan 1995; 39-43.

[10] Jaenicke R. What does protein refolding *in vitro* tell us about protein folding in the cell? Philos Trans R Soc Lond B Biol Sci 1993; 339: 287-94.

[11] Caflisch A. Network and graph analyses of folding free energy surfaces. Curr Opin Struct Biol 2006; 16: 71-8.

[12] Hilser VJ. Modeling the native state ensemble. Methods Mol Biol 2001; 168: 93-116.

[13] Lindorff-Larsen K, Rogen P, Paci E, Vendruscolo M, Dobson CM. Protein folding and the organization of the protein topology universe. Trends Biochem Sci 2005; 30: 13-9.

[14] Walter S, Buchner J. Molecular chaperones-cellular machines for protein folding. Angew Chem Int Ed Engl 2002; 41: 1098-1113.
[15] Clerc M. The swarm and the queen: towards a deterministic and adaptive particle swarm optimization. Proceedings of the IEEE Congress on Evolutionary Computation 1999: 1951-7.
[16] Poli R, Kennedy J, Blackwell T. Particle swarm optimization. Swarm Intell 2007; 1: 33-57.
[17] Carlisle A, Dozier G. An off-the-shelf PSO. Proceedings of the 2001 Workshop on Particle Swarm Optimization, Indianapolis, In: Purdue School of Engineering and Technology, IUPUI 2001: 1-6.
[18] Eberhart RC, Shi Y. Comparing inertia weights and constriction factors in Particle Swarm Optimization. Proceedings of the Congress on Evolutionary Computing 2000: 84-8.
[19] Mayewski S. A multibody, whole residue potential for protein structures, with testing by Monte Carlo simulated annealing. Proteins 2005; 59: 152-69.
[20] Fleming PJ, Gong H, Rose GD. Secondary structure determines protein topology. Protein Sci 2006; 15: 1829-34.
[21] Ho BK, Dill KA. Folding very short peptides using molecular dynamics. PLoS Comput Biol 2006; 2: e27.
[22] Rost B, Yachdav G, Liu J. The PredictProtein server. Nucl Acids Res 2004; 32: W321-6.
[23] Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. Nucl Acids Res 2000; 28: 235-42.
[24] McKnight CJ, Matsudaira PT, Kim PS. NMR structure of the 35-residue villin headpiece subdomain. Nat Struct Biol 1997; 4: 180-4.
[25] Gronenborn AM, Filpula DR, Essig NZ, *et al*. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. Science 1991; 253: 657-61.
[26] Borreguero JM, Skolnick J. Benchmarking of Tasser in the ab initio limit. Proteins 2007; 68: 48-56.
[27] Gong H, Fleming PJ, Rose GD. Building native protein conformation from highly approximate backbone torsion angles. Proc Natl Acad Sci USA 2007; 102: 16227-32.
[28] Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J Comp Aided Mol Des 1993; 7: 473-501.
[29] Moult J. Comparison of database potentials and molecular mechanics force fields. Curr Opin Struct Biol 1997; 7: 194-9.
[30] Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci 1992; 1: 409-17.
[31] Holmes JB, Tsai J. Some fundamental aspects of building protein structures from fragment libraries. Protein Sci 2004; 12: 1636-50.
[32] Yang JS, Chen WW, Skolnick J, Shakhnovich EI. All-atom ab initio folding of a diverse set of proteins. Structure 2007; 15: 53-63.
[33] Chen WW, Yang JS, Shakhnovich EI. A knowledge-based move set for protein folding. Proteins 2007; 66: 682-8.

Received: September 12, 2007 Revised: October 29, 2007 Accepted: November 12, 2007