Beyond the rotamer library: Genetic algorithm combined with the disturbing mutation process for upbuilding protein side-chains

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · NOVEMBER 2002

Impact Factor: 2.63 · DOI: 10.1002/prot.10253 · Source: PubMed

7 AUTHORS, INCLUDING:

Luhua Lai
Peking University
185 PUBLICATIONS **4,569** CITATIONS

SEE PROFILE

PROTEINS: Structure, Function, and Genetics 50:49–62 (2003)

Beyond the Rotamer Library: Genetic Algorithm Combined with the Disturbing Mutation Process for Upbuilding Protein Side-Chains

Zhijie Liu,[1,2] Lin Jiang,[1,2] Ying Gao,[1,2] Shide Liang,[1,2] Hao Chen,[1,2,3] Yuzhen Han,[1,2] and Luhua Lai[1,2,3,*]

[1]*State key Laboratory for Structural Chemistry of Stable and Unstable Species, Beijing, China*

[2]*Department of Chemical Biology, Institute of Physical Chemistry, College of Chemistry, Peking University, Beijing, China*

[3]*Center for Theoretical Biology, Peking University, Beijing, China*

ABSTRACT **The disturbing genetic algorithm, incorporating the disturbing mutation process into the genetic algorithm flow, has been developed to extend the searching space of side-chain conformations and to improve the quality of the rotamer library. Moreover, the growing generation amount idea, simulating the real situation of the natural evolution, is introduced to improve the searching speed. In the calculations using the pseudo energy scoring function of the root mean squared deviation, the disturbing genetic algorithm method has been shown to be highly efficient. With the real energy function based on AMBER force field, the program has been applied to rebuilding side-chain conformations of 25 high-quality crystallographic structures of single-protein and protein–protein complexes. The averaged root mean standard deviation of atom coordinates in side-chains and veracities of the torsion angles of χ_1 and $\chi_1 + \chi_2$ are 1.165 Å, 88.2 and 72.9% for the buried residues, respectively, and 1.493 Å, 79.2 and 64.7% for all residues, showing that the method has equal precision to the program SCWRL, whereas it performs better in the prediction of buried residues and protein–protein interfaces. This method has been successfully used in redesigning the interface of the Basnase-Barstar complex, indicating that it will have extensive application in protein design, protein sequence and structure relationship studies, and research on protein–protein interaction. Proteins 2003;50:49–62.**

Key words: rotamer library; disturbing genetic algorithm (DGA); growing generation amount (GGA) method; protein side-chain; protein–protein interface

INTRODUCTION

Side-chain packing has an essential role in the formation of hydrophobic core and active sites in the protein folding process, and is crucial for protein stability and function. Recently, side-chain conformation predictions have had many successful applications in protein modeling, side-chain redesigning of known proteins,[1] and protein mutation experiments.[2,3] Its increasingly wide applications demand further development of the method.

Side-chain packing seems to be rather esoteric in character because of the size and the flexibility of proteins. The prediction of side-chain conformations is one of the weak points in the comparative modeling procedure,[4,5] the most successful and practical method in protein structures' prediction.[6] Many efforts have been made in recent years to seek a practical solution to these problems, and these methods can be classified into two categories. The first category belongs to exhaustive searching methods, such as the molecular dynamics method,[7] the comprehensive search method,[8,9] and the annealing simulation method (such as the self-consistent ensemble optimization method[10,11]). These methods often start from an arbitrary or selected conformer, and then perform a comprehensive search in the conformational space following certain rules to find the global minimum energy conformation. However, it is clear that exhaustive searches over all possible conformations are beyond the limit of practical computation, so that they can only be applied in small systems. The second category is based on the rotamer library. Through analyzing the statistical characteristics of side-chain conformations in protein X-ray structures, it was reported that most residues favor a relatively small number of conformations determined by their side-chain torsion angles, and these favorable conformations, called rotamers,[12,13] correspond to the constraints of stereochemistry and minimum energy positions. The methods based on the rotamer library greatly reduced comprehensive searches of all possible conformations of protein side-chain to searching the conformations represented by the combinations of rotamers. The rotamer library has become a powerful tool in searching side-chain conformations in protein structure prediction

Grant sponsor: National Natural Science Foundation of China; Grant numbers 29525306 and 20173001; Grant sponsor: Ministry of Science and Technology of China; Grant sponsor: Commission of Science and Technology of Beijing.

The rotamer library and the source codes of the program (in C language) are available by contacting the authors.

*Correspondence to: Luhua Lai, Institute of Physical Chemistry, College of Chemistry and Molecular Engineering, and State Key Laboratory for Structural Chemistry of Unstable and Stable Species, Peking University, Beijing 100871, P.R. China. E-mail: lhlai@pku.edu.cn or lai@mdl.ipc.pku.edu.cn

Received 8 May 2002; Accepted 15 July 2002

since Janin et al.[12] and Bhat et al.[13] performed the pioneering work in this area. Continuous advances were developed by Ponder and Richards,[14] and Tuffery et al.[15–17] through the statistics from more protein crystal structures. In recent years, many groups[18] have extracted more accurate and refined rotamer libraries from detail analyses of protein 3-D structures in the Protein Data Bank. To overcome the incompleteness of the statistical results, Maeyer et al.[19] added many artificial rotamers that were created by rotating certain angles around χ_1 and χ_2 torsion angles subsequently for four aromatic amino acids: HIS, TRP, TYR, and PHE. Kono and Doi[20] made further improvements on extracting more rotamers for long side-chain residues (i.e., LYS and ARG) from statistical analyses. Dunbrack, Karplus, and Cohen developed the backbone-based rotamer library that is widely used in this area.[1,21–24] Most of these methods could work efficiently; however, it is also clear that the calculations will not be reasonable unless the adopted rotamer library has sufficient and accurate rotamers or an additional minimization is performed. Therefore, the completeness and accuracy of the library become the bottleneck for the successful prediction of side-chain conformation.

Recently, some groups focused on improving the rotamer library in the prediction of side-chain conformation. Mendes et al.[25] have developed a flexible rotamer model, where a new rotamer model is described using a continuous ensemble of conformations clustering around classic rigid rotamers. This method can reflect the conformational variance of the side-chain in the real system and predict the free energy precisely by thermodynamical simulation, though the final result is still represented by an approximate rigid rotamer, which may influence the accuracy of the conformation prediction. This flexible rotamer model can be applied to most of the widely accepted approaches for predicting side-chain conformations based on the rotamer library. Mendes et al.[32] have incorporated the model into the self-consistent mean field theory method and obtained reasonable results.

Although the computation is greatly reduced after the rotamer library is adopted, it is still rather computationally expensive, especially when the protein is very large. Thus, the more efficient searching method needs improvement in the speed of the computation. Until now, the conformations of protein side-chains have been rebuilt using different searching algorithms, including the genetic algorithm (GA) (including the selection-mutation-focusing GA and heuristic sparse matrix-driven algorithm[15–17]), Monte Carlo simulation method,[26,27] dead-end elimination method,[19,28–31] fuzzy-end elimination method,[32,33] Hopfield network method,[20] branch and bounding method,[34] branch and terminal method,[21] and the fast method based on the backbone-dependent rotamer library (SCWRL).[22–24]

We have been focusing on two aspects for protein side-chain rebuilding: to overcome the limitations of the rotamer library, and to speed up the computational process. Disturbing genetic algorithm (DGA), a modified GA that combines the disturbing mutation into the normal mutation process, has been developed to search the conformational space around rigid rotamers. The growing generation amount (GGA) method, which can simulate the natural evolution process and makes the sampling procedure more rational, was also introduced to increase the sampling efficiency. The method developed has been tested over examples of side-chain rebuilding of proteins and protein–protein complexes and compared with other side rebuilding methods.

METHODS AND MATERIALS

Rotamer Library

A new database of the solid rotamer library has been generated, which mainly consists of 330 rotamers from the work of Maeyer et al.,[19] and also combines the results from Tuffery et al.[15–17] and Kono and Doi.[20] The cutoff deviations of the side-chain torsion angles χ_1, χ_2, χ_3, and χ_4 are 10°, 20°, 10°, and 10°, respectively; the total deviation of these four torsion angles is 20°. Glycine and alanine are treated to have only one conformer, and proline has the two conformers that are up- and down-type of the conformations minimized by the AMBER force field.[35–38] Thus, we obtained the database of the 525 rotamer library. Considering that side-chains of a few residues, such as HIS, TRP, ASN, and GLN, have asymmetrically planar structures, the rotamer library was enlarged into 549 rotamers finally (see footnote). The accuracy and completeness of our rotamer library will be tested in the section below. In addition, the library can easily be revised by integrating with the flexible rotamer cluster model of Mendes and applied to the thermodynamical free-energy calculation.

DGA

The DGA is proposed, which incorporates the disturbing mutation into the mutation process of GA to search the neighboring region around the certain solid rotamer. Similar to the general GA, our method also includes three phases: initial phase, iterative phase, and judgement phase (see Fig. 1). In the initial phase, a certain amount of arbitrary conformations are generated, then the energy of each conformation is calculated and the lower-energy conformers are chosen as the conformational seeds. The iterative phase consists of five steps: crossover step, general mutation step, disturbing mutation step, energy calculation, and selection of low-energy conformers. The other four steps except the general mutation step are executed using the same methodology described in the general GA. Here, the description of the method only focuses on the disturbing mutation step, the main difference between our DGA and the general GA. The process of the disturbing mutation step is as follows: using the GGL random number generator,[39] a certain proportion of candidates are randomly selected from the proceeding process, and then one torsion angle of one residue for each candidate is randomly disturbed with a user-defined value at one time. Although only one torsion-angle perturbation is performed for each candidate, the corresponding neighboring region in the conformational space will be covered through several generations because of the hereditary link between candidates of different generations in GA. More-

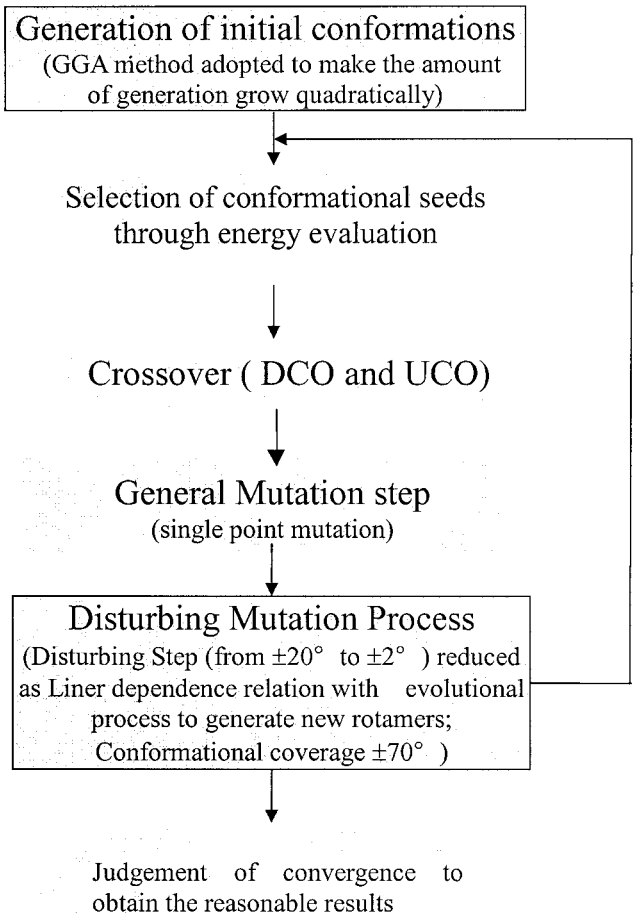

Fig. 1. Flow chart of the disturbing genetic algorithm (DGA).

over, most torsion angles of rotamers, except those related to asymmetrically planar structures, are near 60°, 180°, or −60°. Thus, a disturbing range of −70° to 70° can cover the whole conformation space for one specific torsion angle. In the calculations, the disturbing step is changeable, reducing from ±70° to ±2° with the proceeding of the conformational evolution.

The last phase is the judgment of computational convergence to check whether the system reaches the lowest-energy conformation (LEC). The criterion judging the convergent is that the energy difference between the last two cycles is less than a certain threshold [here they are 0.001 Å for root-mean-squared deviation (RMSD) energy and 0.001 kcal/mol for real energy] for certain times (15 times used in this work), or the computation reaches a maximum number (here it is 1500 times of the number of rebuilt residues).

GGA Method

The GGA is adopted to determinate the candidate number of the generation, including three steps (see Fig. 2). The first step is the initial fast-searching stage, which uses the small candidate number of the generation to reduce the energy of conformational seeds quickly and

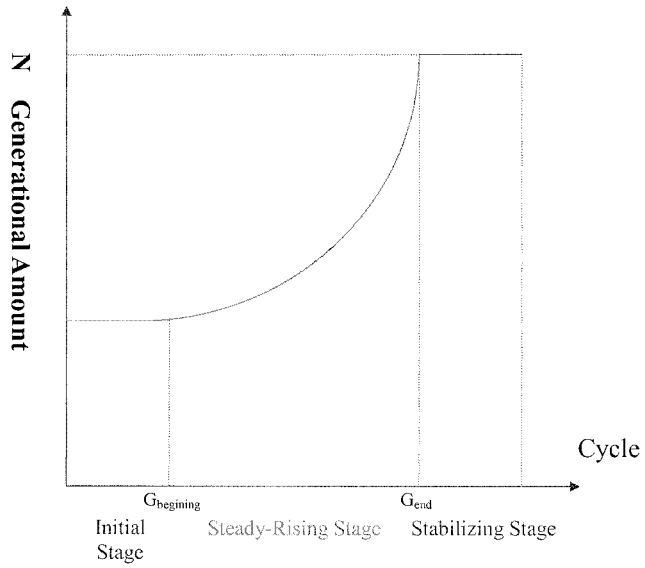

Fig. 2. Curve of generation amount to cycle in growing generation amount method.

dramatically during the first one-tenth of computational cycles. The second one is the steady-rising stage including the four-fifths of computational cycles, where the quadratically growing candidate number is used to screen out all possible conformational combinations. The last step is the stabilizing stage, and the rest computational cycle is performed to obtain the final conformations using the largest value of generation amount.

The detailed growing function is listed below (see Eqs. 1, 2, and 3).

$$Y = \begin{cases} D_1, & X < 0.1 \\ D_1 + (D_2 - D_1)^* Z^2, & 0.1 \le X < 0.9 \\ D_2, & X \ge 0.9 \end{cases} \quad (1)$$

$$X = G_i/G_t \quad (2)$$

$$Z = (G_i - G_{beginning})/(G_{end} - G_{beginning}) \quad (3)$$

D_1 is the initial candidate number of the generation (the small value), D_2 is the final candidate number of the generation (the large one). G_t is the total amount of computational cycles, G_i is the order number of the present running cycle, $G_{\text{beginning}}$ is the beginning number of the steady-rising stage, and G_{end} is the end number of the steady-rising stage.

Energy Function

Conformational energy

During the upbuilding of side-chains, the atoms that belong to the protein backbone and the protein environment are fixed. The total comparative conformational energy Ec is calculated by:

$$Ec = \sum_i \sum_{m,n} E_{i_{mn}} + \sum_{i,j} \sum_{m,n} E_{i_m j_n} + \sum_a \sum_i \sum_m E^a_{i_m}. \quad (4)$$

Here non-rebuilt atoms are denoted by the index a, the residues with the upbuilt side-chain atoms are denoted by the indices i,j, and their side-chain atoms are indexed by m,n. The total energy includes three terms: the inner interaction energy $E_{i_{mn}}$ in the side-chains of each rebuilt residue, the inter interaction energy $E_{i_m j_n}$ between the side-chain atoms of different rebuilt residues, and the outer interaction energy E^a_{im} consisting both of the calculated side-chains with their backbones and with the protein environment.

For each of the energy terms in the Eq. (4), we partly adopted the energy function in the atom-pair form based on the united atom model of the AMBER force field,[36–38] and the energy E' includes three items: van der Waals energy, electrostatic energy, and torsional energy (see Eq. 5).

$$E' = \sum_i \left[\sum_{j<i} \left(k_{ij} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \right) + k'_{ij} \left(\frac{332.1 \cdot Q_i \cdot Q_j}{\varepsilon_{R_{ij}} \cdot R_{ij}} \right) \right) + \sum_{\text{dihedrals}} \frac{V_n}{2} (1 + cos(n\phi - \gamma)) \right]. \quad (5)$$

The definitions and the values of most parameters are directly imported from the AMBER force field, such as standard bond lengths, standard bond angles, van der Waals atom radii (R_i), partial charges (Q_i), nonbonded force constants (ϵ_i) and related A_{ij}, B_{ij}, the constant coefficients of 1–4 interaction to that of >1–4 (k_{ij}, k_{ij}'), torsional potentials ($V_n/2$), numbers of connected bond (n) and phases (γ), etc. The polar hydrogen atoms are deleted, which partial charges are assigned to the corresponding bonded heavy atoms. Here, only the dielectric function $\epsilon_{R_{ij}}$ (see Eq. 6) is a reformulated equation, which is based on a model of the dielectric damping of the electrostatic interaction between two charges in a polar solvent.[40,41] It can ensure the dielectric varying smoothly with the sigmoidal curve of the slope (S) between the plateau value (D, when distance is long), and the minimum value (D', when distance is zero). The parameters D, D', and S are fixed at 78.0, 4.0, and 0.395, respectively. To speed up the computation, the interactions between atoms within the distance range of 15.0 Å are calculated, and this range is wide enough to contain most of the energy contributions:

$$\varepsilon_{R_{ij}} = D - \frac{D - D'}{2} [(R_{ij}S)2 + 2R_{ij}S + 2] \cdot EXP(-R_{ij}S) \quad (6)$$

Moreover, for a given rotamer, its energy of the interaction with other atoms is invariable, and the calculation of this energy item is the main source of the time-consuming computation, increasing steadily and linearly with the expanding of the rotamer number. So it is possible to be memorized to reduce the computational time.

Sequence energy

Although the calculation of the conformational energy can find the optimal conformations for a given sequence, it is not suitable to compare the energies between different sequences, which is necessary in *de novo* sequence design or mutational studies for a given protein backbone structure. An additional energy item (i.e., sequence energy) is required to reflect the intrinsic principles of the different sequences. Here, the sequence energy Es, reflecting the secondary structural tendency of the amino acids, is introduced. It is calculated by the logarithmic function of Chou-Fasman factors (Z_{ip}, i represents the amino acid, p represents the secondary structure)[42–45] of amino acids to certain secondary structures (see Eq. 7).

$$Es = -RT \ln(\sum_i Z_{ip}). \quad (7)$$

The total energy in our program includes both conformational energy and sequence energy.

Data Selection

The structures of the proteins/protein–protein complexes containing the groups near rebuilt side-chain atoms, such as metal ions, SO_4^{2-}, PO_4^{3-}, H_2O, nucleotides, etc., are not selected because of the lack of parameters for these groups. To simulate the conformations of side-chains precisely, the structures having breaks and structure-undetermined residues are omitted. In addition, because there are no items in the force field to describe the formation of the disulfide bond accurately, as a simplified step, all the cystines or cysteines will not be rebuilt in the calculations. For single proteins, the side-chains of all the residues were rebuilt; for the protein–protein complexes, only the side-chains that locate at the interface (i.e., the atoms are related to the interactions between the two proteins within 5.0 Å) are built up simultaneously. Finally, a total of 12 proteins with calculated residue numbers distributed from 36 to 321 and 17 protein–protein complexes with numbers from 31 to 77 were chosen for repacking the side-chains (see Table I). The resolutions of the structures are better than 2.5 Å except 1hrt (2.8 Å), 3hfm (3.0 Å), 1igc (2.6 Å), and 2kai (2.5 Å).

RESULTS

Test of the Parameters Used in the GGA Method

We compared the standard GA with the GGA method to test the parameters of the GGA method. In the process of the GAs, the RMSDs were used as the pseudo energy scoring function. Five proteins and five protein–protein complexes are arbitrarily selected as testing examples (see the Table II). To compare the results using the same criterion, the calculated conformations in each example is about 1000 times of its predicted residue number.

At first, the calculations are executed at fixed generation amounts (FGA) from the beginning to the end with different values of the generation amounts 200, 300, 400, 500, and 600. Under these conditions, all calculations seem to converge. However, further comparison between these results revealed that under smaller generation amounts some examples converged at a local minimum energy conformation with a fast speed, such as 1crn, 4pti, 1cgj, and 1cse for 200, 1cgi for 300, and 2ovo and 4pti for 400,

TABLE I. The Proteins and Protein–Protein Complexes Used in the Side-Chain Packing Calculations

Protein PDB ID	Real size	Calculated residue number	Resolution (Å)	Percent of buried residues
1crn[p,a]	46	40	1.50	23.1
1ifc[p]	131	131	1.19	42.5
1igd[p]	61	61	1.10	20.4
1lz1[p]	130	122	1.50	45.3
1ppt[p]	36	36	1.37	13.8
1ubq[p]	76	76	1.80	40.0
2ovo[p]	56	50	1.50	25.6
2rhe[p]	114	112	1.60	32.9
4pti[p]	58	52	1.50	25.0
1arb[p]	263	257	1.20	60.7
3app[p]	323	321	1.80	55.5
6lyz[p]	129	121	2.00	42.1
1cgj[c,b]	301	48	2.30	75.0
1fle[c]	287	42	1.90	62.9
1hrt[c]	360	77	2.80	69.1
1igc[c]	493	31	2.60	55.6
1jhl[c]	353	31	2.40	65.4
1vfb[c]	352	39	1.80	61.3
2kai[c]	288	36	2.50	81.5
3hfm[c]	558	42	3.00	76.3
3sgb[c]	235	37	1.80	73.1
1cse[c]	337	46	1.20	81.8
1tec[c]	342	48	2.20	78.4
1tpa[c]	281	39	1.90	66.7
2ptc[c]	281	39	1.90	63.3
2sec[c]	338	45	1.80	81.8
2sni[c]	339	47	2.10	69.4
2tgp[c]	281	38	1.90	69.0
4sgb[c]	236	37	2.10	75.0
4tpi[c]	283	40	2.20	58.6

[a]p refers to the structure of a single protein.
[b]c refers to the structure of a protein–protein complex.

whereas under larger generation amounts (500 or more), all computations converged at the LEC but with more expensive calculations (see Table II), just as we described in Methods and Materials.

Based on the knowledge from the FGA results, we repeated the calculations with the GGA method under two sets of parameters, in which total candidate number and initial candidate number is 500, 200 or 500, and 250, respectively. The results have shown that the set of parameters 500–250 is enough to get the LEC with fast speed, and it is noted that the computational demanding is similar to that under the corresponding low FGA from 200 to 300, when producing similar results (see Table II). All of the calculations have shown that the GGA method greatly expedites the converging speed of the generation algorithm and its prediction has similar precision with the general GA with FGAs.

Validating the Rationality of the Rotamer Library and DGA Method

The accuracy and completeness of the rotamer library used and the rationality of our DGA method have been analyzed. The RMSD between the calculated and the crystal structures is still used as the energy scoring function to evaluate the validity of the method. The same five proteins and five protein–protein complexes in the above section were selected as the testing examples and the results from the standard GA and our DGA are compared based on the similar value of the generation amounts.

It is shown that all the RMSD values from DGA are obviously lower than that from general GA and most of them are <0.5 Å except the protein complex 1fle (see Fig. 3). The accuracy analysis of side-chain torsion angles χ_1, $\chi_1 + \chi 2$ (cutoff value is $\pm 40°$) also illustrates the same tendency. Almost all the results from DGA are higher than that from general GA except 1cse, and most of them are $>95\%$ except 2ovo and 1fle (see Table III). The lower RMSD values and higher accuracy of side-chain torsion angles have obviously proved that the rotamer library used for the current DGA method is accurate and sufficient enough. At the same time, they also verify that the DGA method is superior to the general GA method in that it improves the precision, whereas keeping the same computational speed will be more suitable for upbuilding protein side-chains.

The Calculations Based on the Real Energy Function

The real energy function has been used as the energy scoring function here. The 12 proteins and 17 protein–protein complexes were calculated separately, and the lowest calculated energy conformations were compared with the crystal structures (see Tables IV V). Furthermore, the buried residues were defined as those residues having relative accessibility $<25\%$ (calculated using the program Naccess[46]). For the 12 proteins, the average value of side-chain RMSD (excluding Pro, Ala, Gly residues, and including C_β atom) is 1.123 Å for buried residues, and 1.652 Å for all residues. The average percent of accurately predicted torsion angles χ_1, $\chi_1 + \chi_2$ are respectively 88.9%, 77.4% for buried residues and 75.6%, 62.8% for all residues. For the 17 protein complexes, the above values are 1.321 Å, 1.501 Å, 82.6%, 64.4%, 77.9%, and 61.5%, respectively. In addition, the results show that the prediction accuracy is tightly related to the resolution of the crystal structures. For the four protein–protein complexes (i.e., 1igc, 1hrt, 2kai, and 3hfm) with the resolution worse than 2.5 Å, the accuracy of side-chain torsion angles χ_1, $\chi_1 + \chi_2$ are obviously lower than others, especially the $\chi_1 + \chi_2$ torsion angles. After excluding those four structures, the average values for protein–protein complexes are 1.204 Å, 1.347 Å, 87.5%, 68.8%, 82.6%, and 66.4%, respectively; the average values for both protein and protein–protein complex are 1.165 Å, 1.493 Å, 88.2%, 72.9%, 79.2%, and 64.7%. This shows that our calculations have a similar performance and quality for the side-chain prediction of single proteins and protein–protein complexes, especially the side-chain prediction for the core of single proteins and the interface of protein–protein complexes are more precise.

TABLE II. Comparison Between FGA and GGA Methods Using General Genetic Algorithm With Pseudo RMSD Energy Function†

	200 Fixed		300 Fixed		400 Fixed		500 Fixed		600 Fixed		500–200 Growing		500–250 Growing	
Protein	RMS D (Å)	N′/N[a] (%)	RMS D (Å)	N′/N (%)	RMS D (Å)	N′/N (%)	RMS D (Å)	N′/N (%)	RMS D (Å)	N′/N (%)	RMS D (Å)	N′/N (%)	RMS D (Å)	N′/N (%)
1crn[p,b]	***0.582***	14.4	0.479	22.2	0.479	26.3	0.479	35.2	0.479	42.1	***0.582***	14.8	0.479	31.4
1ctf[p]	0.528	34.5	0.528	31.4	0.528	37.9	0.528	50.4	0.528	71.7	0.528	27.8	0.528	27.9
1ppt[p]	0.781	20.7	0.781	33.6	0.781	41.6	0.781	66.2	0.781	54.2	0.781	21.6	0.781	27.8
2ovo[p]	0.676	21.5	0.676	34.4	***0.678***	38.8	0.676	46.8	0.676	47.3	0.676	22.4	0.676	27.9
4pti[p]	***0.562***	32.5	0.561	46.1	***0.562***	37.5	0.561	60.9	0.561	55.2	0.561	43.2	0.561	27.9
1cgi[c]	0.816	22.6	***0.817***	32.3	0.816	44.6	0.816	44.1	0.816	52.7	0.816	15.8	0.816	28.9
1cgj[c]	***0.789***	24.3	0.788	28.3	0.788	33.3	0.788	46.6	0.788	64.0	0.788	25.2	0.788	27.6
1cho[c]	0.563	23.4	0.563	29.5	0.563	41.3	0.563	49.4	0.563	56.5	***0.695***	23.5	0.563	32.8
1cse[c]	***0.596***	20.5	0.594	28.8	0.594	37.7	0.594	51.6	0.594	51.3	***0.596***	28.9	0.594	24.8
1fle[c]	0.861	29.3	0.861	30.8	0.861	39.5	0.861	55.8	0.861	58.2	0.861	27.6	0.861	30.0

†The bold examples are represented as those converged at a local minimum energy conformation.
[a]N′/N refers to the percentage of calculated conformation numbers when approaching to a convergence.
[b]For other details refer to Table I.

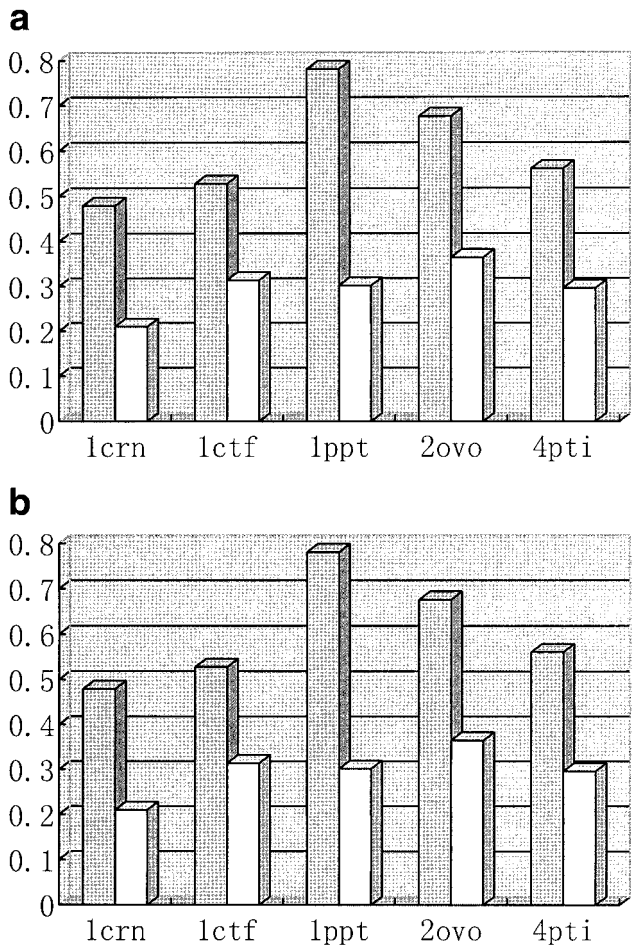

Fig. 3. Comparison of RMSDs between general genetic algorithm and disturbing genetic algorithm with pseudo RMSD energy function. The results with dark color are from general GA, and those with light color are from DGA. **(a)** Single protein, **(b)** the interface of protein–protein complex.

TABLE III. Comparison of the Accuracy of Calculated Side-Chain Torsion Angles With Crystal Structures Between General Genetic Algorithm and Disturbing Genetic Algorithm

	General GA		Disturbed GA	
Protein	χ_1 (%)	$\chi_1 + \chi_2$ (%)	χ_1 (%)	$\chi_1 + \chi_2$ (%)
1crn	100.0	96.9	100.0	100.0
1ctf	95.7	95.7	100.0	97.8
1ppt	100.0	93.1	100.0	96.6
2ovo	95.6	86.7	95.6	93.3
4pti	97.6	90.5	97.6	95.2
1cgi	95.7	89.4	100.0	100.0
1cgj	97.7	90.9	97.6	97.6
1cho	100.0	97.2	100.0	97.1
1cse	100.0	97.0	100.0	100.0
1fle	91.7	91.7	94.7	92.1

Comparison With Other Methods

The above 25 protein examples (excluding the four protein complexes with a low resolution) are also calculated by one of the successful side-chain programs—SCWRL[29–31] and compared with our results using RMSD comparison and torsion-angle comparison in detail (see Fig. 4). It is obvious that the two methods have different performances for the prediction of different regions of proteins. When comparing the predictions of the 12 single proteins, our DGA method has similar capability with SCWRL for the buried residues, and performs somewhat worse than SCWRL for all the residues. When comparing the predictions of the 13 protein–protein complexes, our DGA method obviously works better than SCWRL for the buried residues and has comparable quality to SCWL for all the residues. The average RMSD value and accuracy of torsion angles also validates the same conclusion (see Table VI). Moreover, we have calculated the average values for the total 25 cases, though the numbers of single proteins and protein–protein complexes would influence the results somewhat. For the buried residues, the average

TABLE IV. The DGA Results of Single Proteins With Real Energy Function

	RMSD		χ_1 (%)		$\chi_1 + \chi_2$ (%)		
Protein	Buried	All	Buried	All	Buried	All	Resolution (Å)
1arb	1.434	1.522	82.9	79.8	71.2	67.2	1.20
1crn	0.956	0.771	100.0	96.2	100.0	88.5	1.50
1ifc	1.149	1.691	89.6	69.0	60.4	50.4	1.19
1igd	0.905	1.772	80.0	71.4	80.0	63.3	1.10
1lz1	1.366	1.804	86.0	72.6	79.1	63.2	1.50
1ppt	0.872	1.934	100.0	79.3	75.0	65.5	1.37
1ubq	0.836	1.843	92.3	67.7	84.6	52.3	1.80
2ovo	0.622	1.402	100.0	76.9	90.0	64.1	1.50
2rhe	1.137	1.597	77.8	69.5	74.1	61.0	1.60
3app	1.241	1.314	81.6	76.7	69.1	63.7	1.80
4pti	1.670	2.259	88.9	72.2	77.8	52.8	1.50
6lyz	1.294	1.909	87.5	75.8	67.5	62.1	2.00
Average	1.123	1.652	88.9	75.6	77.4	62.8	—

TABLE V. The DGA Results of Interfaces in Protein–Protein Complex With Real Energy Function

	RMSD		χ_1 (%)		$\chi_1 + \chi_2$ (%)		
Protein	Buried	All	Buried	All	Buried	All	Resolution (Å)
1cgj	1.599	1.742	73.3	67.5	60.0	55.0	2.30
1cse	1.367	1.344	88.9	81.8	63.0	57.6	1.20
1hrt	1.700	2.070	57.4	52.9	46.8	41.2	2.80
1igc	1.210	1.686	73.3	66.7	60.0	48.1	2.60
1jh1	0.828	1.152	100.0	88.5	88.2	80.8	2.40
1tec	1.741	1.852	72.4	70.3	62.1	56.8	2.20
1tpa	1.429	1.697	95.0	86.7	75.0	73.3	1.90
1vfb	1.163	1.167	89.5	90.3	68.4	71.0	1.80
2kai	2.181	2.358	63.6	63.0	40.9	40.7	2.50
2ptc	0.984	1.117	89.5	83.3	68.4	70.0	1.90
2sec	1.499	1.720	81.5	75.8	70.4	63.6	1.80
2sni	0.980	1.329	92.0	88.9	76.0	69.4	2.10
2tgp	0.770	0.877	95.0	89.7	75.0	72.4	1.90
3hfm	1.716	1.896	72.4	68.4	51.7	52.6	3.00
3sgb	1.164	1.364	89.5	80.8	73.7	69.2	1.80
4sgb	1.059	1.062	83.3	87.5	55.6	58.3	2.10
4tpi	1.067	1.091	88.2	82.8	58.8	65.5	2.20
Average	1.321	1.501	82.6	77.9	64.4	61.5	—

RMSD of side-chain atoms was 1.165 Å in DGA, which was lower than 1.329 Å in SCWRL, the accuracy of torsion angles χ_1 and $\chi_1 + \chi_2$ were 88.2% and 72.9% in DGA, which were higher than 87.3% and 71.3% in SCWRL. Both of the above results illustrated that our DGA method has higher precision than SCWRL. For all residues, the above values were 1.493 Å, 79.2%, and 64.7% in DGA and 1.506 Å, 82.1%, and 66.5% in SCWRL, respectively. The inconsistency between comparisons of RMSD and torsion angle might come from the defects and characters of these two comparison methods. However, it has still shown that DGA and SCWRL have equivalent power in rebuilding protein side-chains. Compared with SCWRL, our DGA is relatively worse in the prediction for surface residues. However, for the buried core of the single protein in which formation was mainly driven by hydrophobic effects, DGA can reach as high precision with RMSD 1.123 Å, χ_1 88.9%, and $\chi_1 + \chi_2$ 77.4% as SCWRL with 1.073 Å, 92.1%, and 78.7%. Especially for the buried part of the interface in the protein–protein complex in which formation may involve more strong electrostatistic interactions and hydrophobic effects, DGA works more robustly with RMSD 1.204 Å, χ_1 87.5%, and $\chi_1 + \chi_2$ 68.8% than SCWRL with 1.565Å, 83.0%, and 64.4%.

Additional analysis on each of the amino acids has also been conducted. Both DGA and SCWRL have similar tendencies for the same residues. The comparison of RMSD has revealed that for single protein, DGA has equal predictive quality to SCWRL on each of the residues, except the polar residues SER, ASP, GLN, and LYS in the buried region and TRP on the surface; the total RMSDs are slightly higher than SCWRL (see Fig. 5).

The result matches partly with the foregoing analysis that DGA considered the electrostatic interaction extravagantly. As for the interface of protein–protein complex, DGA showed its great advantages that it almost predicted better than SCWRL for each of the buried residues, especially for hydrophobic residues such as MET, PHE,

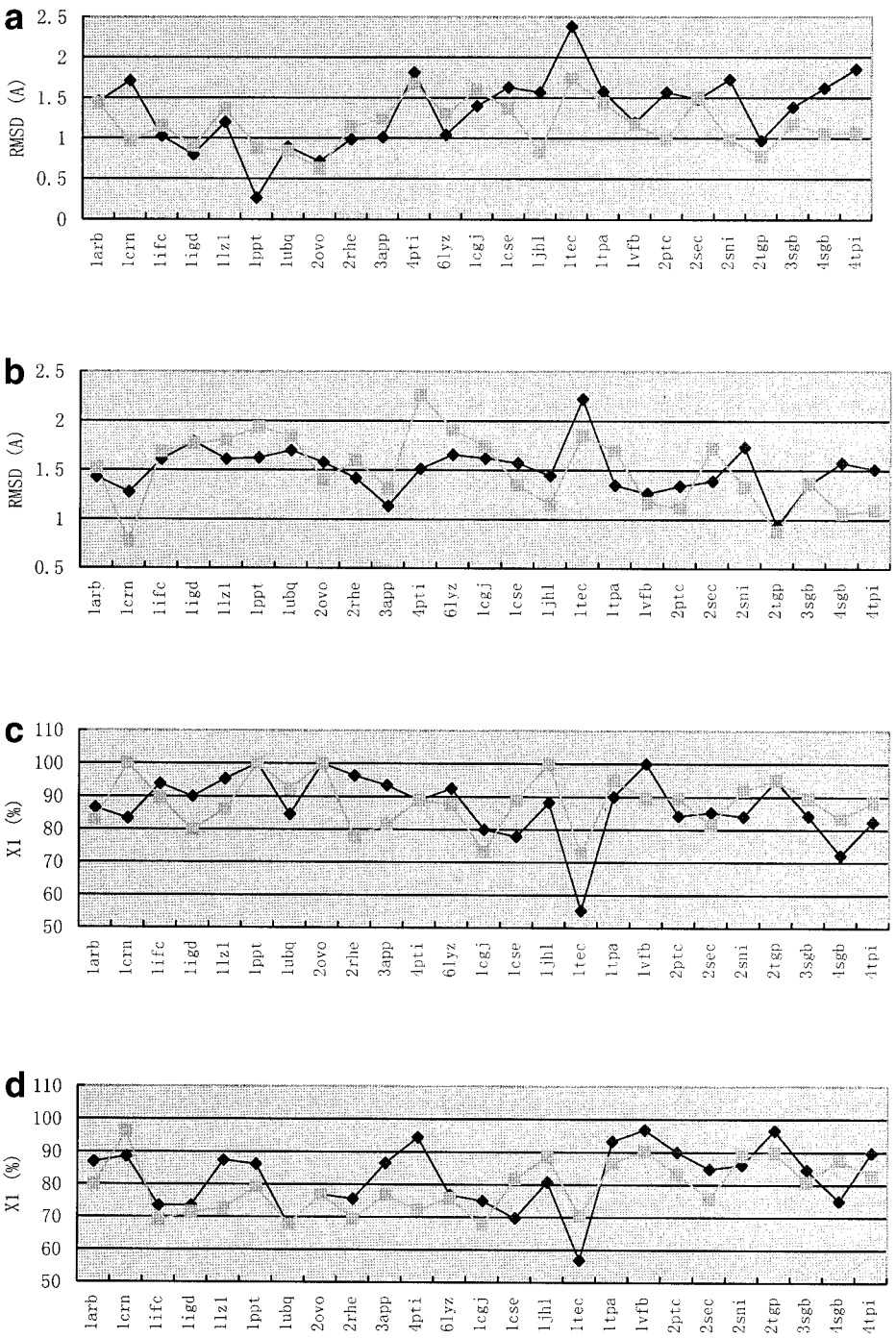

Fig. 4. Detailed comparison of 25 cases between DGA and SCWRL program. **(a)** The RMSD comparison of buried residues, **(b)** the RMSD comparison of all residues, **(c)** the comparison of accuracy of torsion angle χ_1 of buried residues, **(d)** the comparison of accuracy of torsion angle χ_1 of all residues, **(e)** the comparison of accuracy of torsion angle $\chi_1 + \chi_2$ of buried residues, and **(f)** the comparison of accuracy of torsion angle $\chi_1 + \chi_2$ of all residues. The results of DGA are figured in light color, the results of SCWRL are figured in dark color.

HIS, and TRP. As for those unburied residues, it still kept comparable high precision to SCWRL and worked better for the aromatic residues PHE and HIS (see Fig. 6).

These results also imply that the hydrophobic interaction is dominant in the formation of the protein core, whereas both the hydrophobic effect and electrostatic effect cooperate to favor the formation of the protein–protein interface.

Moreover, the comprehensive comparison of all residues has shown that, in the RMSD comparison, DGA has an equal performance with SCWRL, and it performed better for buried residues, especially for hydrophobic residues

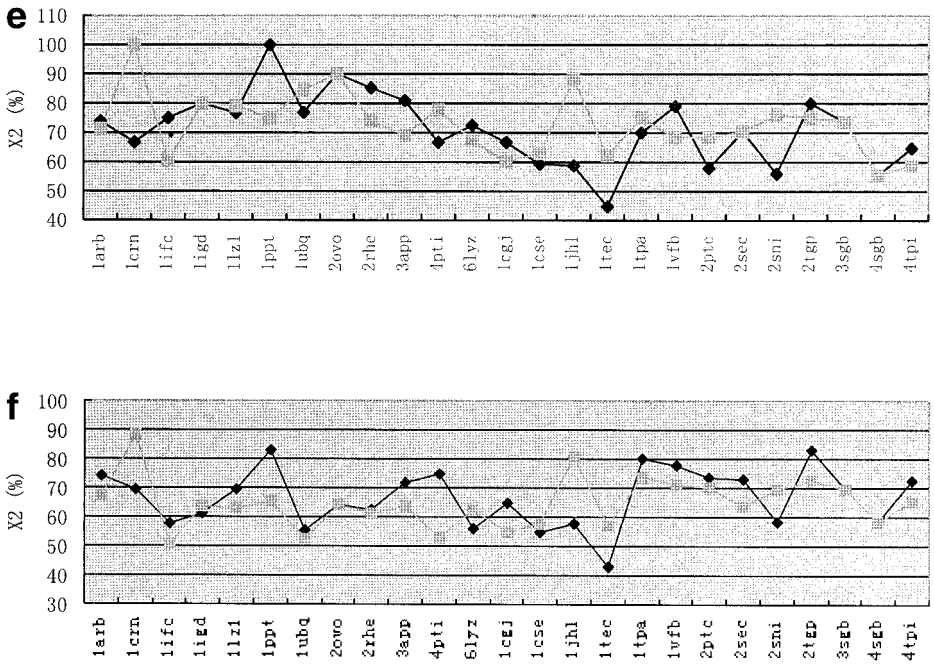

Figure 4. (Continued.)

MET, PHE, HIS, and TPR, and worse for unburied TYR and TRP; in the torsion angle comparison, the calculation of χ_1 shows that DGA calculates the conformation of the side-chain more precisely than SCWRL for hydrophobic residues except SER, GLU, and GLN, and the calculation of χ_2 shows that the precision of our prediction seems to be low for both aromatic residues and long-chain residues (see the Fig. 7), which is the common problem for all current side-chain methods.

We also compared our results with Kono and Doi[20] and Tuffery et al.[16] coarsely. Because of the different criterions adopted, the comparison was not very precise. For buried residues of single proteins, the RMSD value and accuracy of χ_1 torsion angle were 1.0 Å and 87.5% in Kono and Doi's work, whereas they were 1.17 Å and 88.2% in ours. The precision in the DGA method was considerable to theirs. For the whole set of proteins, the average RMSDs of buried residues and all residues were 1.17 Å and 1.49 Å in our work, and 1.23 Å, 1.66 Å in the work of Tuffery et al.; the DGA played better. In general, our DGA method at least had comparable precision to other successful methods for side-chain rebuilding with a higher reliability in side-chain modeling for the core of single proteins and the interface of protein–protein complexes.

Applications in Protein Design: Determination and Sequence Screening of Binding Sites

We applied our side-chain repacking to protein sequence design. One well-studied protein complex, the complex of Barnase-Barstar, has been chosen as the tested example to explore the ability of our DGA method in recognizing and redesigning the binding sites at the protein–protein interface. The sequence mutation experiment had proved that K27 R59 R83 R87 H102 in Barnase and Y29 D35 W38 D39 T42 W44 E76 E80 in Barstar were key binding residues at the interface and directly relative to the protein association.[47,48]

Our DGA method is applied to the random sequence search upon the binding sites and the predictions are compared with the experimental data from a single mutation. First, we conducted the random sequence search for the binding sites in one chain of the complex when fixing the other. When fixing the sites of the Barstar and mutating the sites of the Barnase, the calculated sequence with the lowest energy is R27 (K, the native residue) R59 R83 K87(R) H102. It is interesting to note that the side-chain orientations and the conformations of all five residues are similar to the native states, and the residues in positions 27 and 87 are replaced by the residues whose nature is very similar to the native one.

The calculations have also implied that the residues in the above five positions are rather conservative and have an important role in the protein binding process. And our conclusion is strengthened by the single mutation experiment, in which the differences of variations of Gibbs free energy $\Delta\Delta G$ in these five positions are greater than 5.0 kcal/mol (see Table VII). However, when fixing the sites of the Barnase and mutating the sites of the Barstar, the calculated sequence is F29(Y) N35(D) W38 D39 R42(T) W44 D76(E) R80(E), which is analogous to the native one (see Table VII). All the aromatic residues are accurately predicted and the side-chain orientations and conformations of these residues are similar to those native ones in crystal structures. Compared with the experimental results, it is found that the residues with the $\Delta\Delta G$ (to alanine) of >3.0 kcal/mol are kept to be the same or homologous ones, especially for Y29 mutated to F29, whose mutation data has revealed that it is a relatively

TABLE VI. Comparison of Average Results Between DGA and SCWRL

	Protein		Protein complex		Total	
	DGA	SCWRL	DGA	SCWRL	DGA	SCWRL
RMSD/Å (Buried)	1.123	1.073	1.204	1.565	1.165	1.329
RMSD/Å (All)	1.652	1.527	1.347	1.486	1.493	1.506
χ_1/% (Buried)	88.9	92.1	87.5	83.0	88.2	87.3
χ_1/% (All)	75.6	81.2	82.6	83.0	79.2	82.1
$\chi_1 + \chi_2$/% (Buried)	77.4	78.7	68.8	64.4	72.9	71.3
$\chi_1 + \chi_2$/% (All)	62.8	66.6	66.4	66.5	64.7	66.5

more active substitution, though the ΔΔGs is only −0.1 kcal/mol. Whereas for those whose ΔΔG is <3.0 kcal/mol, the predicted residues are relatively much different from the native ones, such as E80 mutated to R80 whose charge is reversed. It might imply that those residues are not as important as others, and the calculations may be affected by the larger solvent accessible surface area.[46]

Second, the sequence redesign is conducted in the binding sites on both sides of the complex simultaneously. The sequence with the lowest energy has showed that most of the positions originally occupied by the aromatic residues (they are directly related to the hydrophobic interactions) are still located by the native residues or those of similar nature (see Table VII).

The predictions for the polar residues favoring the salt-bridge or other strong electrostatic interactions varied much from the native ones. However, the calculations are acceptable and reasonable—some polar residue pairs reversed their positions, for instance, R59 in the Barnase and E76 in the Barstar are mutated to D59 and R76, which still keeps the strong salt-bridge at the interface.

DISCUSSION

The Two Advantages of the Current Method

Our method has two obvious features different from other rotamer library-based side-chain methods. One is the DGA, and the other is the GGA method.

DGA

The introduction of the rotamer library has shown much success in the prediction of side-chain conformation. However, the library is generated from limited protein crystal structures, and the used methods for statistical analysis may not be very effective. The rotamer library may inevitably have some defects; for instance, it cannot provide enough conformations that access all possible conformational space, but only limited conformations covering partial conformational space. However, although the rotamers of the library are usually close to the native state, there are still obvious deviations observed in some rotamers, which may result in the wrong prediction. Most of the rotamer library-based methods are not able to overcome this, so they have to enlarge the library and perform a further minimization to optimize the conformations.

To accurately predict the side-chain conformation that is deviate from the rotamer library, a more comprehensive searching step should be used to screen the whole conformational space, such as the theoretical searching modeling methods (i.e., *ab initio* modeling method). In our method, the disturbing mutation is incorporated into the mutation process of GA. Thus, the conformation search improves its ability to access much wider conformation space. Through the introduction of disturbing mutation, this DGA is able to search the neighboring region around a certain rigid rotamer more precisely. And this kind of local disturbing algorithm makes the searching cover the whole conformational space, and has a great probability of approaching the native states of the side-chains. Therefore, the DGA search will compensate the incompleteness and inaccuracy of the rotamer-based method using the general search algorithm.

GGA method

Another improvement in our method is the rational determination of the amount of candidates in each generation to speed up the calculation and obtain a reasonable result. Our calculations have shown that for the calculations of FGAs, a small candidate number will lead to fast convergence whereas hardly avoid falling into local minimum energy conformations; a large one may reduce the possibility of trapping into the local energy pitfalls, but will slow down the convergent speed. The growing generation amount idea has been adopted here: in the beginning of the computational cycles, the smaller generation amount is selected, and then the amount is gradually increased with the process of the computation until it reaches the largest generation amount (see Fig. 2). The GGA idea simulates the real situation of natural evolution, in which the generation becomes larger and larger and the species grows more and more with the evolutional process. Our calculations have indicated that it can improve the searching speed and at the same time have the prediction of similar precision with the general GA.

The rotamer library, widely used in protein modeling and protein design, often makes the predicted side-chain conformations deviate from the real states because of the limitation of conformational accuracy and lack of completeness of the library. Through the introduction of the DGA and the GGA, our method combines the advantages of both the database-based GA method and the *ab initio* modeling method, and successfully overcomes the defects of the rotamer library. It is expected that the method will have the extensive application in the study of protein sequence design, functional protein rebuilding, and protein–protein interaction.

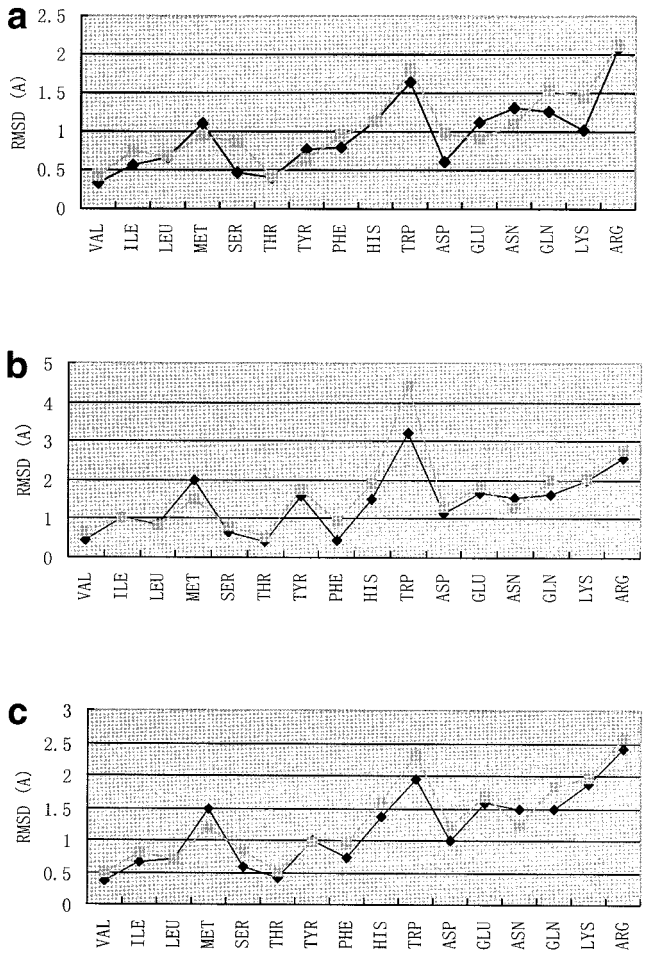

Fig. 5. The RMSD comparison of each residues in single protein between DGA and SCWRL. **(a)** Compared results for buried residues, **(b)** compared results for surface residues, and **(c)** compared results for all residues. The results of DGA are figured in light color, the results of SCWRL are figured in dark color.

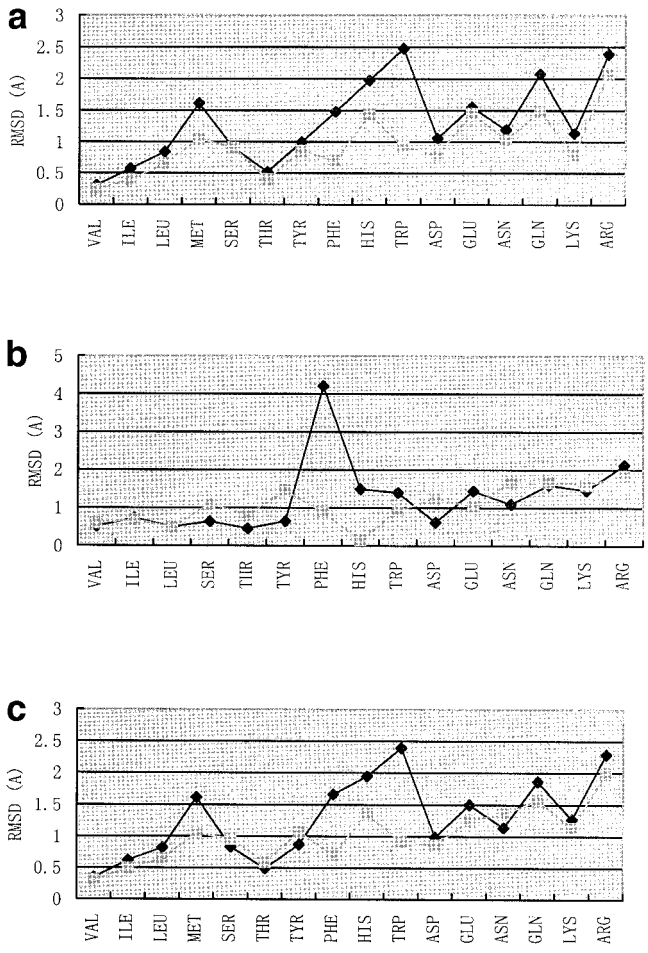

Fig. 6. The RMSD comparison of each residue in the interface of the protein–protein complex between DGA and SCWRL. **(a)** Compared results for buried residues, **(b)** compared results for surface residues, and **(c)** compared results for all residues. The results of DGA are figured in light color, the results of SCWRL are figured in dark color.

The Backbone-Depended Rotamer Library or Not?

Recently, the backbone-depended rotamer library has been widely applied because of its high precision,[1,21–24] in our DGA method; it can also be introduced to make our rotamer library more complete and more accurate and overcome the randomicity of the searching process to screen the conformational space more quickly and reasonably. However, because the rotamer library is based on experimental data, it requires the high accuracy and reliability of the backbone structures. In our calculation, the structures with the lower resolutions (>2.5 Å, such as 1hrt, 1igc, 3hfm, and 2kai) were predicted relatively poorly, indicating that the accuracy of backbone on side-chain conformations strongly influences the prediction of the conformation side-chain. And our conclusion is strengthened by the study of Tuffery et al.[17] Therefore, for the general side-chain methods based on the backbone-depended rotamer library, in which conformations of rotamers tightly depend on the backbones from known structures, it would be very difficult to find a suitable rotamer when the dependent backbone structures are different from the given ones; thus, the precision of the prediction would be destroyed. However, in our DGA method, the adopted conformational searching step—disturbing mutation—could tolerate and endure the variations of protein main chain. Moreover, it can not only consider the variations of the backbone properly, but also find the new or modified suitable conformations according to the environment of protein bulk. In fact, the DGA method has two obvious advantages. First, it maintains the most virtues of the GA method based on the rotamer library to search the conformational space quickly. Second, it also has advantages from the *ab initio* modeling method; it can subtly access the local region and provide new proper conformations which do not exist in the rotamer library. Therefore, it overcomes the incompleteness and inaccuracy of the rotamer library to make the deviation smaller; especially, it can rebuild side-chain conformation successfully, although there may be deviations with the backbone structure.

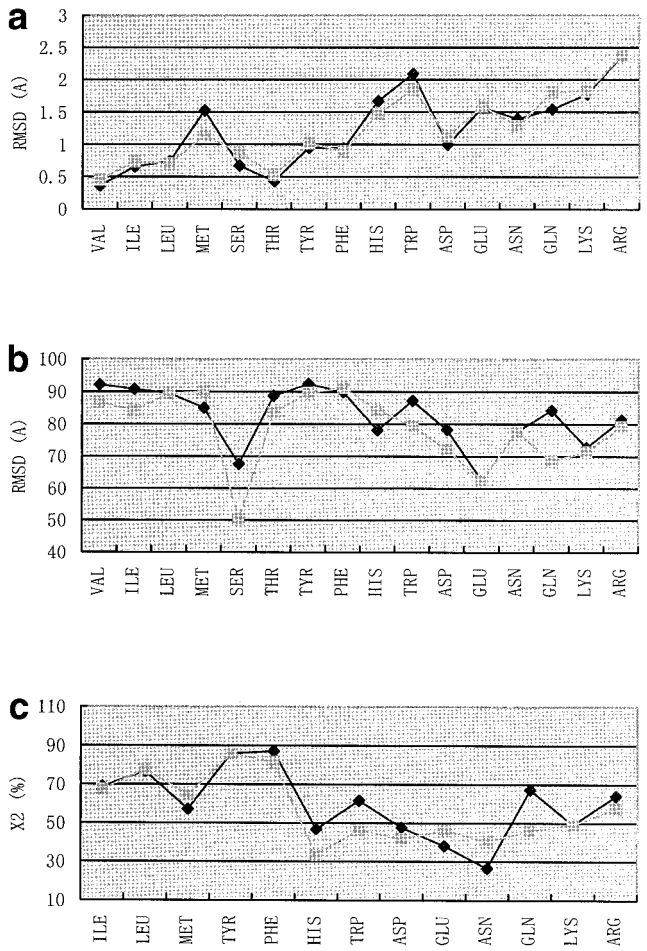

Fig. 7. The full comparison of each residue between DGA and SCWRL. **(a)** RMSD comparison, **(b)** comparison of accuracy of χ_1 torsion angles, and **(c)** comparison of accuracy of $\chi_1 + \chi_1$ torsion angles. The results of DGA are figured in light color, the results of SCWRL are figured in dark color.

Energy Function Suitable for the Description of the Protein Interior and the Protein–Protein Interface

Side-chain packing seems to be rather esoteric in character because of the size and the flexibility of proteins. Now it is a common understanding that van der Waals energy has an important role in side-chain packing. Many methods had only adopted van der Waals energy or van der Waals-like contribution in the energy evaluation.[19–24,28–34] At the same time, some groups have investigated the effects of other interactions in this process. Kono and Doi[20] had never incorporated the hydrogen bond item into the energy calculation, whereas Tuffery et al.,[15–17] Wilson et al.,[5] Lee and Subbiah[10,11] and Mendes et al.[25] had introduced the electrostatic interaction in, and all of them obtained reasonable results. Recently Mendes has explicitly proved that the electrostatic effect could improve the precision of side-chain prediction for all practical polar or charged residues significantly.

To describe the conformational energy of the protein system accurately, especially in the interface of the protein–protein complex, the formation of which involves more electrostatic interactions, three types of energies are considered: van der Waals energy, electrostatic energy, and torsional energy, which reflect the main nonbonded interactions and the stability of the rotamer. The results have indicated that the conformational energy is able to screen out the low-energy conformation close to the crystal structure, especially for the side-chain prediction of the protein interior and the protein–protein interface.

In the typical force field,[36–38] the energy function can mainly describe the physical energy changes, such as the energy difference between different conformational states, but cannot deal with the chemical energy changes, such as residue mutation, formation or break of the new disulfide bond, etc. Especially in protein design, the different sequences could not be compared because there was no common energy standard. In this work, the Chou-Fasman propensities for protein sequence (i.e., sequence energy) have been incorporated into the energy function to reflect the secondary structural tendency of different amino acids on a given protein backbone structure. So the sequence energy can describe the energy difference between different sequences to some extent. The introduction of the Chou-Fasman propensities will promote the convenience for comparing different sequences and make the predicted sequence more favorable to the main-chain backbone, especially in sequence mutation or designing new protein sequences. All of the results have shown the reasonability of the energy function for the description of the protein interior and the protein–protein interface. At the same time, it is noted that more accurate energy function, which could take the chemical changes into account, should be developed. For example, when redesigning the interface of the protein–protein complex, the proper binding energy should be introduced. Moreover, there were so many complicated interactions in protein, the energy should include more effects to upbuild the protein side-chains accurately.

The RMSD Value Acts as the Justifying Function

Since Tuffery et al.[16] compared the methods for protein side-chain rebuilding in 1993, and proved that GA was a successful method in this working field, different methods with different comparing criterions had been developed. For example, the volume overlap method had been exploited in the work of Maeyer et al.[19] However, the most popular methods were still the comparison using the RMSD value of the atom coordinates or the accuracy of the torsion angles. Considering that the RMSD has great advantages to indicate the difference between the calculated and the crystal structures, we also chose the RMSD function and the results have shown the power and robustness of the justifying function.

Comparison With the SCWRL Method

In our method, the high accuracy of torsion angles and low RMSD value indicates that our method is able to accurately predict the side-chain conformations. It is especially suitable to repack side-chain conformations for the protein core and the interface of the protein–protein complex, the regions where the atoms are tightly packed.

TABLE VII. The Test Example of the Barnase-Barstar Case: Designed Sequence Compared with the Native and the Experimental Mutation Data

	Barnase					Barstar							
Position	27	59	83	87	102	29	35	38	39	42	44	76	80
SCS (Å^2)[a]	8.7	21.6	5.3	0.1	0.7	29.9	2.7	10.5	2.4	13.1	21.2	39.7	70.1
$\Delta\Delta G$ (kcal · mol^{-1})[b]	5.4 A	5.2 A	5.4 Q	5.5 A	6.1 A	−0.1 F 3.4 A	4.5 A	1.6 F	7.7 A	1.8 A	0.0 F	1.4 A	0.5 A
NS[c]	K	R	R	R	H	Y	D	W	D	T	W	E	E
FDS[d]	R	R	R	K	H	F	N	W	D	R	W	D	R
UDS[e]	R	D	H	K	E	W	K	W	R	R	W	R	E

[a]Solvent accessible surface area.
[b]Difference from variation of Gibbs free energy in mutation experiment.
[c]Native sequence.
[d]Designed sequence by fixing one part in the Barnase-Barstar complex.
[e]Designed sequence of binding sites in the whole interface of the Barnase-Barstar complex.

Compared with the energy function in SCWRL which only counts the steric effects of the simplified non-bonded collision, the strong electrostatic interactions from the energy function exploited in DGA would spoil the proper conformations of surface residues. Generally, the surface residues are often solvated by water molecules, which will greatly weaken the electrostatic effects between polar atoms. So our DGA calculation is relatively worse than the SCWRL program in the prediction of surface residues. However, in the side-chain prediction for the packed region where the side-chain conformation was mainly driven by hydrophobic effects or strong electrostatistic interactions and hydrophobic effects, such as the protein core and the buried protein–protein interface, the DGA method has shown its accuracy and rationality in the side-chain prediction.

Applications in Protein Design

In general, our DGA method is able to detect the conservation and the importance of the binding sites. The predicted sequence may be helpful for redesigning new sequence for binding sites in protein complex for protein engineering studies. In addition, through rebuilding the side-chains on the main-chain backbone of certain protein cores or known protein folding motifs, it will reveal the relationship between the sequence and the structure, which may be helpful for research on protein evolution.

CONCLUSION

By comparing with general GA, our DGA method has shown obvious advantages in rebuilding side-chain conformations. The disturbing mutation process overcomes the limitations brought by the rotamer library and has the merits of both the knowledge-based method and the *ab initio* modeling method. However, the GGA method can simulate the natural evolution process and makes the sampling procedure more rational. The comparison with the reported successful side-chain method (such as SCWRL) shows that the DGA method has comparable precision and is more suitable for side-chain modeling in the protein core and the protein–protein interface. A practical sequence search for the interface of Barnase-Barstar complex proves that our DGA method not only can determine the conservation and importance of the sites, but also has a prospective future in protein mutation and new protein sequence design. It is expected that this method will have wide applications in protein modeling, protein design, and protein–protein interaction studies.

REFERENCES

1. Su A, Mayo SL. Coupling backbone flexibility and amino acid sequence selection in protein design. Protein Sci 1997;6:1701–1707.
2. Honda N, Komeiji Y, Uebayasi M, Yamato I. Computational design of a substrate specificity mutant of a protein. Proteins 1996;26:459–464.
3. Tidor B. Helix-capping interaction in lambda Cro protein: a free-energy simulation analysis. Proteins 1994;19:310–323.
4. Summers NL, Carlson WD, Karplus M. Analysis of side-chain orientations in homologous proteins. J Mol Biol 1987;196:175–198.
5. Wilson C, Gregoret LM, Agrad DA. Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. J Mol Biol 1993;229:996–1006.
6. Blundell TL, Sibanda BL, Sternberg MJE, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 1987;326:347–352.
7. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744.
8. Warme PK, Momany FA, Rumball SV, Tuttle RW, Sheraga HA. Computation of structures of homologous proteins: α-lactalbumin from lysozyme. Biochemistry 1974;13:768–782.
9. de la Paz P, Sutton MJ, Darsley NJ, Rees AR. Modelling of the combining sites of three anti-lysozyme monoclonal antibodies and of the complex between one of the antibodies and its epitope. EMBO J 1986;5:415–425.
10. Lee C, Subbiah S. Prediction of protein side-chain conformation by packing optimization. J Mol Biol 1991;217:373–388.
11. Lee C. Predicting protein mutant energetics by self-consistent ensemble optimization. J Mol Biol 1994;236:918–939.
12. Janin J, Wodak S, Levitt M, Maigret B. Conformation of amino acid side-chains in proteins. J Mol Biol 1978;125:357–386.
13. Bhat TN, Sasisekheran V, Vijayan M. An analysis of sidechain conformations in proteins. Int J Pept Protein Res 1979;13:170–184.
14. Ponder JW, Richards FM. Tertiary template for protein use of packing criterion in the enumeration of allowed sequences for different structural classes. J Mol Biol 1987;193:775–791.
15. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side chain conformations. J Biomol Struct Dyn 1991;8:1267–1289.
16. Tuffery P, Etchebest C, Hazout S, Lavery R. A Critical comparison of search algorithm applied to the optimization of protein side-chain conformations. J Comput Chem 1993;14:790–798.
17. Tuffery P, Etchebest C, Hazout S. Prediction of protein side chain conformations: a study on the influence of backbone accuracy on

conformation stability in the rotamer space. Protein Eng 1997;10: 361–372.
18. Schrauber H, Eisenhaber F, Argos P. Rotamers: to be or not to be? An analysis of amino acid side-chain conformation in globular proteins. J Mol Biol 1993;230:592–612.
19. De Maeyer M, Desmet J, Lasters I. All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of side-chains by dead-end elimination. Fold Des 1997;2:53–66.
20. Kono H, Doi J. A new method for side-chain conformation prediction using a Hopfield network and reproduced rotamers. J Comput Chem 1996;17:1667–1683.
21. Gordon DB, Mayo SL. Branch-and-terminate: a combinatorial optimization algorithm for protein design. Structure Fold Des 1999;7:1089–1098.
22. Dunbrack RL Jr, Karplus M. Backbone dependent rotamer library for protein application to side-chain prediction. J Mol Biol 1993;230: 543–574.
23. Dunbrack RL Jr, Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein side-chains. Nat Struct Biol 1994;1:334–340.
24. Dunbrack RL Jr, Cohen FE. Coupling backbone flexibility and amino acid sequence selection in protein design. Protein Sci 1997;6:1661–1681.
25. Mendes J, Baptista AM, Carrondo MA, Soares CM. Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. Proteins 1999;37:530–543.
26. Holm L, Sander C. Database algorithm for generation protein backbone and side-chain co-ordinates from a Cα trace: application to model building and detection of co-ordinate errors. J Mol Biol 1991;218:183–194.
27. Holm L, Sander C. Fast and simple Monte Carlo algorithm for side-chain optimization in proteins: application to model building by homology. Proteins 1992;14:213–223.
28. Desmet J, Maeyer MD, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. Nature 1992;356:539–542.
29. Lasters I, De Maeyer M, Desmet J. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side-chains. Protein Eng 1995;8:815–822.
30. Goldstein RF. Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys J 1994;66:1335–1340.
31. Tanimura R, Kidera A, Nakamura H. Determination of protein side-chain packing. Protein Sci 1994;3:2358–2365.
32. Lasters I, Desmet J. The fuzzy-end elimination theorem: correctly implementing the side-chain placement algorithm based on the dead-end elimination theorem. Protein Eng 1993;6:717–722.
33. Keller DA, Shibata M, Marcus E, Ornstein RL, Rein R. Finding the global minimum: a fuzzy-end elimination implementation. Protein Eng 1985;8:893–904.
34. Eyrich VA, Standley DM, Felts AK, Friesner RA. Protein tertiary structure prediction using a branch and bound algorithm. Proteins 1999;35:41–57.
35. Sybyl 6.4. Menu manual. St. Louis: Tripos Inc; 1997. 99 p.
36. Weiner SJ, Kollman PA, Case DA, et al. A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 1984;106:765–784.
37. Weiner SJ, Kollman PA, Nguyen DT, Case DA. An all atom force field for simulations of proteins and nucleic acids. J Comput Chem 1985;7(2):230–252.
38. Cornell WD, Cieplak P, Bayly CI, et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
39. Lewis PAW, Goodman AS, Miller JM. A pseudo-random number generation for the system 360. IBM Sys J 1969;2:136–146.
40. Arora N, Jayaram B. Strength of hydrogen bonds in α helices. J Comput Chem 1997;18:1245–1252.
41. Hingerty BE, Ritchie RH, Ferrell TL, Turner JE. Dielectric effects in biopolymers: the theory of ionic saturation revisited. Biopolymers 1985;24:427–439.
42. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet and random coil regions calculated from proteins. Biochemistry 1974;13:211–222.
43. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry 1974;13:222–244.
44. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 1978;47:45–148.
45. Chou PY, Fasman GD. Empirical predictions of protein conformation. Annu Rev Biochem 1978;47:251–276.
46. Hubbard SJ, Campbell SF, Thornton JM. Molecular recognition: conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. J Mol Biol 1991;220:507–530.
47. Schreiber G, Fersht AR. Energetics of protein–protein interactions: analysis of the Barnase-Barstar interface by single mutations and double mutant cycles. J Mol Biol 1995;248:478–486.
48. Vaughan CK, Buckle AM, Fersht AR. Structural response to mutation at a protein–protein interface. J Mol Biol 1999;286:1487–1506.