THE JOURNAL OF PHYSICAL CHEMISTRY B

 VOLUME 106, NUMBER 26, JULY 4, 2002

LETTERS

Genetic Algorithm to Design Stabilizing Surface-Charge Distributions in Proteins

Beatriz Ibarra-Molero and Jose M. Sanchez-Ruiz*

Facultad de Ciencias, Departamento de Quimica Fisica, Universidad de Granada, 18071-Granada, Spain

Received: February 21, 2002; In Final Form: April 23, 2002

Practical and technological applications of proteins are often limited by their low stability. Recent experimental and theoretical work suggests that large stability enhancements might be obtained through the design of the surface-charge distribution. However, for a typical protein, the number of different surface-charge distributions is huge, and an exhaustive search for the stabilizing distributions is clearly out of the question. We describe here a simple genetic algorithm that can search the space of the surface-charge distributions efficiently and can find many potentially stabilizing distributions under specified restrictions (such as, for instance, the fact that the active-site region is not modified or that the number of differences with the wild-type form is kept low). Also, the genetic algorithm could be employed to select small sets of interacting sites to be used in in vitro-directed evolution procedures that address the attainment of protein variants of enhanced stability.

Introduction

Practical and technological applications of proteins are very often limited by their low stability. Recent work[1−8] indicates the possibility of enhancing protein stability by introducing mutations that optimize the charge−charge interactions on the protein surface. As discussed recently,[7] this stabilization approach has, in principle, several important advantages, one of which is that mutations at surface positions are less likely to be disruptive and affect the structure and the function of the protein. Actually, experimental and theoretical work[5−12] suggests that Nature may use this approach when "designing" thermophilic proteins. However, whereas several groups have recently succeeded in achieving small stability enhancements by introducing charge-deletion or charge-reversal single mutations,[2−4] some calculations[7] suggest that Nature is able to redesign the surface-charge distribution (introducing several mutations) to make thermophilic proteins significantly more stable than their mesophilic counterparts. Clearly, achieving large stability enhancements for proteins of biotechnological interest through the design of the surface-charge distribution would also be desirable. To achieve this goal, however, we must solve several problems, some of which (but not all; see ref 7) are associated with the difficulties of the computational search in the space of the charge distributions. In connection with this, the two following points must be noted:

(1) For n surface positions (and assuming, for illustration only, two kinds of residues: positive and negative), the number of different charge distributions is given by 2^n and may become enormous, even for moderate values of n. Clearly, the exhaustive calculation of the charge−charge interaction energies for all distributions is out of the question, except, perhaps, for the smallest proteins.

(2) A system of n "sites" (surface positions) that may be in two different "states" (positive charge and negative charge) and in which there are site−site interactions (stabilizing between unlike charges and destabilizing otherwise) is reminiscent of spin glass models in the sense that we may expect a high level of frustation and a rugged energy landscape with many local minima of similar energy; that is, we may expect a significant number of different stabilizing charge distributions of similar energy (this expectation is sustained by the results reported here; see below).

* To whom correspondence should be addressed. E-mail: sanchezr@ugr.es. Tel: 34-958-243189. Fax: 34-958-272879.

10.1021/jp020483o CCC: $22.00
Published on Web 06/07/2002

The above discussion suggests that protein stabilization via design of the surface-charge distribution would greatly benefit from an algorithm that can efficiently search the associated rugged energy landscape and find stabilizing distributions under user-specified restrictions (for instance, that the active-site region is not modified or that the number of differences from the wild-type form (WT) is kept low). Here, we describe a simple genetic algorithm[13] (GA) that meets these requirements.

Calculation of the Charge−Charge Interaction Energies

During the GA runs, a large number of charge distributions are generated. Their charge−charge interaction energy was estimated using our implementation[1] of the Tanford−Kirkwood model.[14] This is admittedly a very simplistic approach; however, it has been shown to predict qualitatively for several proteins the salt-induced proton uptake/release behavior[15] and the effects of charge-deletion and charge-reversal mutations on stability.[7] For the sake of simplicity and speed of calculation, we have introduced in this work simplifications in the procedure we previously described:[1]

(1) The location of the "new" charges in the distributions generated by the GA is somewhat uncertain because there may be several sterically allowed rotamers for each ionizable side chain. For illustrative purposes of the calculations presented here, we have deemed it sufficient to place the charges in the atom with the largest accessible surface area (ASA)[16] out of all those in the side chain present in the wild-type form. This, however, is not a limitation of the approach, and the placement of charges on the basis side-chain modeling[17] could also be used.

(2) We employ "generic" positive groups (corresponding to Lys and Arg) and generic caxboxylic acid side chains (corresponding to Asp and Glu). The former are taken to be nonionizable, and, for the latter, an intrinsic pK value of 4.25 is used as the input of the Tanford−Kirkwood calculation; other ionizable groups (His, amino terminal, carboxy terminal) are treated as in ref 1. This approximation has essentially no consequences for the neutral pH calculations reported in this work.

(3) The calculation of the fractional protonations of ionizable groups is carried out using the Tanford−Roxby, mean-field procedure,[18] rather than the more rigorous but time-consuming Bashford−Karplus "reduced-set of sites" approximation.[19] See ref 19 for a comparison of the two approaches. Also, we applied Gurd's correction[20] using the average accessibility[21] of the ionizable side chains in the wild-type form to avoid time-consuming ASA calculations for all the variants generated by the GA.

Description of the Genetic Algorithm Used

Assume that we have selected n surface positions in a protein and that we consider three types of residues in those positions: generic positive (+), noncharged (o), and generic carboxylic acid that are likely to be negatively charged at neutral pH (−). Each of the 3^n possible charge distributions is determined by the nonselected ionizable residues (the same for all) and by a string that specifies the kind of residue present in the selected positions, for instance, + + o − + − o o − − +. We refer to each of those strings as a chromosome.

We set an initial population in which a certain number of chromosomes are randomly generated (i.e., the residue in each position is randomly selected between +, −, and o) and a certain number correspond to the wild-type form. A score (Z) is then assigned to each chromosome in the population; the simplest possibility is to use as a score the charge−charge interaction energy (E_{TK}) derived from the Tanford−Kirkwood calculation described above:

$$Z = E_{TK} \tag{1}$$

The best chromosome (the one with the lowest score) is passed unmodified to the next generation. Next, chromosomes are randomly selected in couples according to a probability related to their score; specifically, the probability that a given chromosome is selected for reproduction is proportional to $Z_W - Z$, where Z is its score and Z_W is the score of the worst chromosome in the population.[22] From each couple of selected chromosomes ("parents"), two "children" are generated by crossover, with a single crossover point chosen at random. Children are placed in the new population, and the process is continued until the new population is completed. The children chromosomes in the new population are then subjected to mutation according to a given mutation probability per position (P_{MUT}), that is, each position in each chromosome has a probability P_{MUT} of being selected for mutation, so a selected position may become +, −, or o with equal probability (therefore, there is a $^1/_3$ probability that the selected position is not actually mutated). Finally, Z scores are calculated, and a new cycle starts with the new population. This process is continued until the best chromosome has remained unmodified for a specified number of cycles. The relevant parameters of our genetic algorithm are the following.

Size of the Population. This can be varied within a wide range without seriously compromising the performance of the algorithm, although very large populations cause the algorithm to run slowly. We have used populations of 20 chromosomes in all GA calculations reported in this work.

Number of Copies of the WT Distribution Present in the Initial Population. We have found that starting with a certain number of WT chromosomes (usually half of the initial population) significantly increases the speed of the algorithm in runs addressed at obtaining stabilizing distributions with few differences with the WT (discussed further below).

Mutation Probability, P_{MUT}. Although a significant mutation rate is desirable, too high a value of P_{MUT} causes the genetic algorithm to approach an inefficient random search. We use $P_{MUT} = 0.05$ in all GA calculations reported in this work.

Seed for the Random Number Generator. Different seeds will generate different random number series (used in the random selection of parents, crossover points, and positions subject to mutation) and will lead to statistically different outcomes of the algorithm.

Using the Genetic Algorithm to Explore the Space of Surface-Charge Distributions

An example of a representative run with our genetic algorithm is shown in Figure 1. For illustrative purposes of this and subsequent calculations reported in this work, we have used the hen egg-white (HEW) lysozyme structure (pdb code: 1lzt) with residues of accessibility to solvent[21] higher than 0.4 taken as the surface positions. This approach gives a total of 42 surface positions and $3^{42} = 1.09 \times 10^{20}$ different charge distributions for three kinds of residues: neutral, generic positive, and generic carboxylic acid. Note in Figure 1 that a comparatively small number of GA cycles lead to a best chromosome corresponding to a charge distribution significantly more stable than that of the WT (according to our calculated Tanford−Kirkwood energy).[23]

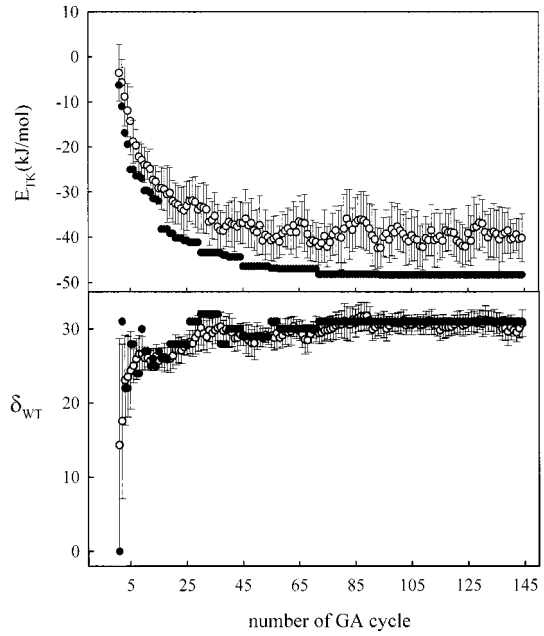

Figure 1. Representative example of a GA run using the structure of HEW lysozyme and taking as surface groups those with accessibility to solvent[21] higher than 0.4. An initial population of 20 chromosomes was set, of which 10 were randomly generated and 10 were exact copies of the WT chromosome. A mutation probability of 0.05 was used, and the run was stopped when the best chromosome had not changed for 50 cycles. The penalty energy was 0, and, therefore, the scoring function equals the Tanford−Kirkwood electrostatic energy (eq 1). The closed symbols are the Tanford−Kirkwood energy (E_{TK}) and the number of changes from WT (δ_{WT}) for the best chromosome. The open symbols are the mean values of E_{TK} and δ_{WT} for the population; the associated bars represent the corresponding standard deviations from the mean values.

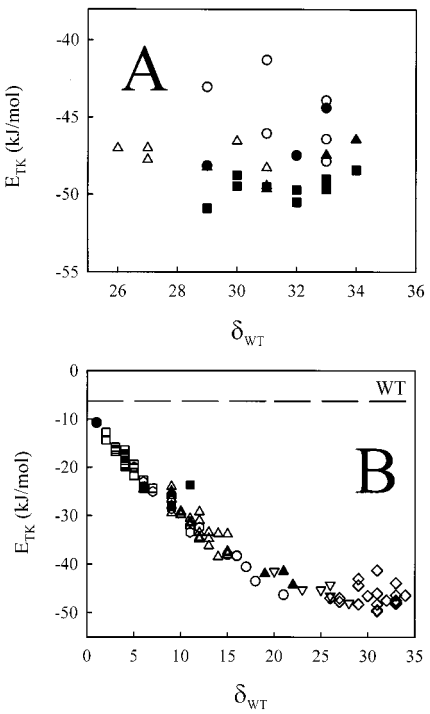

Figure 2. (A) Plot of electrostatic Tanford−Kirkwood energy (E_{TK}) vs the number of changes from WT (δ_{WT}) for the best chromosomes obtained in several GA runs, such as that shown in Figure 1. A closed symbol indicates that the chromosome is a local minimum (i.e., its scoring function cannot be improved by a single change). The type of symbol refers to the stopping criterion (the value for the number of cycles without change in the best chromosome: 20 (○, ●), 50 (△, ▲), and 500 (■). (B) Plot of E_{TK} vs δ_{WT} for the best chromosomes obtained in GA runs with different values of the penalty energy (E_P in eq 2). These runs were stopped after 20 or 50 cycles without a change in the best chromosome, and the E_P values used (kJ/mol) were 0 (◊), 0.5 (▽), 0.75 (▲), 1 (○), 1.25 (△), 1.5 (⬡), 1.75 (■), 2 (□), and 4 (●).

Figure 2A shows the charge−charge electrostatic energy and the number of differences with WT for the best distributions obtained from different GA runs using different seeds and different stopping criteria (i.e., different values for the number of cycles without a change in the best chromosome). Note that many different distributions of similar energy are obtained and that many of them are, in fact, local minima (in the sense that they cannot be improved by single changes). Although there cannot be any guarantee that the distribution with the lowest E_{TK} energy among those in Figure 2A is the global minimum, the efficiency of the GA in finding many different stabilizing distributions is clearly demonstrated.

The distributions of Figure 2A have a comparatively large number (26−34) of differences with the WT distributions. In practice, however, it may be desirable to achieve significant stabilization while keeping the number of differences with WT as low as possible (so that only a few mutations need to be introduced). The genetic algorithm can be forced to generate distributions arbitrarily close to the WT by a suitable modification of the scoring function

$$Z = E_{TK} + E_P \delta_{WT} \qquad (2)$$

where δ_{WT} is the number of differences with the WT distribution and E_P is a penalty energy per difference with WT. The results of several GA runs using different values of the penalty energy are shown in Figure 2B as a plot of E_{TK} versus the number of differences with WT. The plot suggests an approximate functional relationship between E_{TK} and δ_{WT} for the best distributions derived from the GA runs and allows us to decide the penalty energy value that must be used to explore distributions with δ_{WT} values within desired ranges. Thus, for instance, $E_P = 1.25$ kJ/mol leads to distributions with differences with WT in the 10−15 range. For illustration, we show in Table 1 the 10 best distributions of lowest energy derived from 20 GA runs with $E_P = 1.25$ kJ/mol; note that there are significant differences between these distributions, despite the fact that they have similar energy and δ_{WT} values.

Concluding Remarks

We have shown that a simple genetic algorithm can efficiently search the space of surface-charge distributions in proteins and can find many stabilizing distributions under specified restrictions. The illustrative calculations presented in this work use the structure of lysozyme and take as surface residues those with accessibility to solvent[21] higher than 0.4; however, the selection of the surface residues can be easily modified so that, for instance, active-site residues or residues involved in certain specific interactions are not included.

Our genetic algorithm could perhaps be dubbed "evolution in silico", and it was inspired by the procedure for protein stabilization through in vitro-directed evolution (Proside) that was recently developed by Schmid and co-workers.[24] Proside selects protein variants on the basis of their protease resistance, but it appears likely that a significant part of the stabilization achieved is related to charge−charge interactions.[24] Proside has been shown to work with six selected positions, whereas our genetic algorithm can efficiently handle much larger numbers of positions; however, the main difference between the two

TABLE 1: Surface Charge Distributions for the Wild Type Form of HEW Lysozyme and Several Variants Generated by Genetic Algorithm Runs with a Penalty Energy of 1.25 kJ/mol[a]

		WT	variants								
position	E_{TK} (kJ/mol)	−6.3	−38.0	−37.6	−37.3	−33.8	−33.7	−33.4	−32.1	−31.5	−31.0	−29.9
1		+										
2		o	−	−		−	−	−		−	−	−
5		+					−		−			−
7		−										
13		+										
14		+		−				−				
18		−										
19		o					+					
21		+	o		−		−					
37		o	−	−	−	−	−	−	−		−	−
39		o		+		+						
41		o		−	−	−			−	−		
44		o	+	+	+	+	+	+	+	+	+	+
45		+	−	−			−			−	−	−
47		o										
65		o	+	+	+	+	+			+	+	
68		+			−	−		−	−			
70		o										
73		+										
75		o		−		−						
77		o	−								−	
79		o			−	−						
81		o			+	+						
85		o			+	−		+	+			
86		o	−	−	−		−	−	−		−	
87		−										
89		o	+				+	+	+			+
93		o	−				−	−	−			−
97		+										
101		−										
106		o	−	−	−				−	−		
109		o	+	+	+					+	+	+
112		+									o	−
113		o	−	−	−					−	−	
114		+										
116		+				−	−	o				
118		o										
119		−					+					
121		o				+		+	+	+	+	+
122		o	+	+	+	+	−	+		+		
125		+	−	−	−	−		−		−	−	
128		+										
number of changes from WT		0	15	15	15	15	14	13	11	11	12	10

[a] Twenty runs were performed. The Table gives the distributions corresponding to the 10 chromosomes with the lowest energies among the 20 best chromosomes obtained. Only the surface positions selected for variation (corresponding to residues of solvent accessibility[21] in the WT higher than 0.4) are shown. The symbols +, −, and o correspond to the generic types of groups (positive, carboxylic acid, and neutral) described in the text. A blank space in a variant means that the residue in that position is of the same type as the residue in WT. The electrostatic Tanford−Kirkwood energies and the number of differences with WT are also given.

evolution procedures is, of course, that Proside is an experimental approach and has been shown to work in practice,[24] whereas our genetic algorithm is a computational prediction tool and, ultimately, will be no better than the electrostatic model used in the scoring function. Although the simple Tanford−Kirkwood calculation employed in this work has been shown to be reasonably successful at qualitatively predicting the effect of single charge-reversal and charge-deletion mutations on protein stability,[7] it certainly appears advisable to explore the use of more realistic electrostatic models and to take into account the electrostatic interactions in the denatured state,[25−27] as we have recently discussed.[7]

We believe that, to the extent that fast and reliable models can be used in the scoring function, our genetic algorithm may become a prediction tool that is useful in the stabilization of proteins through the design of the surface-charge distribution. In addition, the genetic algorithm could be employed to select small sets of potentially interacting positions to be used in in vitro-directed evolution procedures, an application that will probably be less demanding on the electrostatic model used.

Acknowledgment. This work was supported by Grant BIO2000-1437 from the Spanish Ministry of Science and Technology.

References and Notes

(1) Ibarra-Molero, B.; Loladze, V. V.; Makhatadze, G. I.; Sanchez-Ruiz, J. M. *Biochemistry* **1999**, *38*, 8138−8149.

(2) Loladze, V. V.; Ibarra-Molero, B.; Sanchez-Ruiz, J. M.; Makhatadze, G. I. *Biochemistry* **1999**, *38*, 16419−16423.

(3) Grimsley, G. R.; Shaw, K. L.; Fee, L. R.; Alston, R. W.; Huyghues-Despointes, B. M.; Thurlkill, R. L.; Scholtz, J. M.; Pace, C. N. *Protein Sci.* **1999**, *8*, 1843−1849.

(4) Spector, S.; Wang, M.; Carp, S. A.; Robblee, J.; Hendsch, Z. S.; Fairman, R.; Tidor, B.; Raleigh, D. P. *Biochemistry* **2000**, *39*, 872−879.
(5) Perl, D.; Mueller, U.; Heinemann, U.; Schmid, F. X. *Nat. Struct. Biol.* **2000**, *7*, 380−383.
(6) Pace, C. N. *Nat. Struct. Biol.* **2000**, *7*, 345−346.
(7) Sanchez-Ruiz, J. M.; Makhatadze, G. I. *Trends Biotechnol.* **2001**, *19*, 132−135.
(8) Perl, D.; Schmid, F. X. *J. Mol. Biol.* **2001**, *313*, 343−357.
(9) Spassov, V. Z.; Karshikoff, A. D.; Ladenstein, R. *Protein Sci.* **1994**, *3*, 1556−1569.
(10) Elcock, A. H.; McCammon, J. A. *J. Mol. Biol.* **1998**, *280*, 731−748.
(11) Xiao, L.; Honig, B. *J. Mol. Biol.* **1999**, *289*, 1435−1444.
(12) Wassenberg, D.; Welker, C.; Jaenicke, R. *J. Mol. Biol.* **1999**, *289*, 187−193.
(13) The development of the algorithm described in this work was based upon the simple and lucid description of genetic algorithms given by Marek Obitko in http://cs.felk.cvut.cz/∼xobitko/ga/.
(14) Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333−5339.
(15) Garcia-Mira, M. M.; Sanchez-Ruiz, J. M. *Biophys. J.* **2001**, *81*, 3489−3502.
(16) ASA values were calculated using a modification of the Shrake−Rupley algorithm (Shrake, A.; Rupley, J. A. *J. Mol. Biol.* **1973**, *79*, 351−372) that randomly places 2000 points on the surface of the expanded van der Waals sphere corresponding to each atom so that accurate atomic ASA values are obtained. We used a radius of 1.4 Å for the solvent probe and the Chothia set for the protein atoms (Chothia, C. *J. Mol. Biol.* **1976**, *105*, 1−12).
(17) See, for instance, Liang, S.; Grishin, N. V. *Protein Sci.* **2002**, *11*, 322−331 and references therein.
(18) Tanford, C.; Roxby, R. *Biochemistry* **1972**, *11*, 2192−2198.
(19) Bashford, D.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 9556−9561.
(20) Matthew, J. B.; Gurd, F. R. N. *Methods Enzymol.* **1986**, *130*, 413−453.
(21) The accessibility is calculated as the ratio between the side-chain ASA value in the native protein and in a Gly-X-Gly tripeptide in which the conformation of the side chain is the same as that in the native protein.
(22) Because the score is related to an energy, it might appear reasonable to use "Boltzmann selection" (i.e., the probability that a chromosome is selected for reproduction is made proportional to a Boltzmann exponential: $\exp(-Z/RT)$). This, however, causes only a few chromosomes to be systematically selected in each GA cycle, with the concomitant loss of population diversity.
(23) Our genetic algorithm program was written in Quickbasic 4.5 and made use of previously developed subroutines for the calculation of charge−charge interaction energies (see ref 1). The actual time required to complete typical GA runs, such as that shown in Figure 1, was on the order of minutes (using a PC computer with a Pentium III, 800-MHz processor). One of the reviewers of this paper has requested information about the availability of our program to other investigators. Our plan is to make our approach widely available after we have checked the practical details of its application through the comparison of the predictions with the experimental results for model proteins (work in progress).
(24) Martin, A.; Sieber, V.; Schmid, F. X. *J. Mol. Biol.* **2001**, *309*, 717−726.
(25) Pace, C. N.; Alston, R. W.; Shaw, K. L. *Protein Sci.* **2000**, *9*, 1395−1398.
(26) Elcock, A. H. *J. Mol. Biol.* **1999**, *294*, 1051−1062.
(27) Zhou, H.-X. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 3569−3574.