W.J.J. van den Tweel, A. Harder and R.M. Buitelaar (Eds.), *Stability and Stabilization of Enzymes*
Proceedings of an International Symposium held in Maastricht, The Netherlands, 22-25 November 1992

© 1993 Elsevier Science Publishers B.V. All rights reserved.

Genetic algorithms as a new tool to study protein stability

Thomas Dandekar and Patrick Argos

European Molecular Biology Laboratory, Postfach 10 22 09,
Meyerhofstraße 1, W-6900 Heidelberg, Germany

Abstract

Genetic algorithms use recombination and mutation of solution trials to derive near optimal solutions for problems involving complex and large state spaces. Protein engineering represents such a demanding task and we illustrate how genetic algorithms are a promising new design tool to identify optimal amino acid mutations and to study protein stability in a variety of applications. A first example illustrates optimal substitutions in the core of a protein. The next example searches for mutations in a long protein sequence which improve protein stability by several criteria which do not compromise the critical features of the starting wild type sequence. A three-dimensional lattice simulation achieves a compact and stable fold for a simple model protein composed of four beta strands. Loop length and overall hydrophobicity prove to be most important for the stability of the fold. In grid-free simulations the strength of different beta strand propensities assigned to extended residues can be simulated to test their effect on stability. The influence of hydrophobicity on helix stability is studied in grid-free simulations of four helix bundles: strong hydrophobicity in a few particular residues stabilizes more effectivly in the simulations than using hydrophobic residues distributed over larger regions of the helices.

1. INTRODUCTION

Genetic algorithms in engineering applications are only metaphors of life borrowed for the purpose of optimization [1]. In contrast, we have begun to apply genetic algorithms to analyze and improve protein stability. The following aspects of genetic algorithms promise to offer advantages over earlier approaches to model proteins in a simplified way.

(i) The genetic technique can be more realistic as it models the evolution of an adapted protein structure by mutation, recombination and selection in a natural way.

(ii) Genetic algorithms process in each generation and in parallel many structures and even more schemata [1] in search of the global minimum for a protein conformation.

(iii) The simulations should be achieved computationally faster than with the use of full molecular dynamics [2], another frequently utilized technique to minimize a protein's conformational energy.
(iv) The confomational space is searched in a unique way such that two "bad" solutions (as judged by the parameters given for the selection) which have only partial regions of optimal structure may recombine to yield a much better ("fitter") new structure.
(v) Complex parameters may be incorporated into genetic algorithms. The increase in calculation time is relatively modest as even a rough estimate of each parameter is sufficient to drive selection [1].

In the following we illustrate by simple examples for a variety of applications how these advantages can be used to simulate and identify mutations and design criteria which should enhance protein stability.

2. MATERIALS AND METHODS

Simulation programs were written in VAX-PASCAL using modified versions of the simple genetic algorithm [1]. The fittest individuals from several (0-24) selection runs ("epochs", typically runs of 120 generations, population sizes of 500 individuals) were collected for a final competition run against a random background.

In protein structure simulations it is critical to encode the structure in an efficient way.
(i) In the sequence models, amino acid sequences were encoded in each individual by six-bit tupels according to the genetic code.
(ii) In the three-dimensional grid simulations, coordinates were calculated according to the four possible directions (two-bit tupels) of the $C\alpha$ atom trace on a tetrahedral lattice.
(iii) For the grid-free simulations standard conformations were encoded by tupels and coordinates calculated from these according to their ϕ and ψ dihedral angles.

Genetic algorithms also allow discrimination amongst many different parameters known to be important in protein structural stability and folding (hydrophobicity, accessibility, charges, amino acid content, helical propensity, etc.). The optimal fitness function made from such parameters and their relative weights which efficiently drives selection to realistic and observable structures contains the crucial parameters necessary to select and fold a particular structure. In practice it is important to be as simple as possible and to introduce or alter only one parameter at a time to understand its effect. The refined fitness functions used here are sketched in the examples given. Details of the protein engineering program and the grid-bound simulation are described in [3].

3. RESULTS

3.1. Stability of core residues and protein sequences

Residues in the core of a protein are sometimes essential for the performance of its function and critical for the overall stability of the protein. They provide an example of the strategy used in genetic algorithms. Their sucessful identification by the genetic algorithm is effected by modeling evolution. A first population is composed of individuals having genomes with random nucleotides. Each chromosome is translated according to the genetic code. Three amino acids assumed in the illustrative case determine the fitness of each individual where optimal core packing is calculated and amino acids known to pack well get an additional bonus. The probability to be selected as a parent for the next generation of solution trials increases according to fitness.

Table I. Optimization of lambda-repressor core residues

Epoch 1	FFR MFY
Epoch 2	IFE IIV VIV
Epoch 3	LLA FVG FLA FFA FMA
Epoch 4	LFE FNI VLI LLV
...	
Epoch 7	VVV VVI MVV

(final epoch: MVV remains the fittest)

Three core residues, experimentally analyzed by [4] in lambda-repressor, were optimized for core packing. The simulation began with random sequences of these residues, a population of 30 individuals, and a string length of 18. The fitness function was (Aadiff x 2000 + 2000-(257-packaging)^2) where packaging is the volume of V,I,L,M or F residues in Å^3 and Aadiff is the number of V,I,L,M or F in the sequence. The value 257 represents the ideal and likely conserved core volume as observed in the known lambda-repressor core structure [4]. New fittest individuals appearing in 7 epochs, each of 11 generations, are given. MVV agrees with that known from the protein structure.

The genomes for the children in the second generation are formed by crossing over some of the parent genomes (20%) and by occasional mutations (0.01/bit) in copying the parent genome. The resulting genomes are translated again and the fitness of the encoded peptides determines who will be preferentially selected as parent for the third generation encoding fitter peptides. After several generations, (near) optimal solutions are reached. An illustration is given in Table I where amino acids are denoted by their single letter code. The optimal solution found in nature (MVV) is correctly picked by the simulation even before the final competition run (see Materials and Methods).

A particular advantage of genetic algorithms is that they can optimize several parameters in long protein sequences to improve protein stability while maintaining desired features of the native sequence which are taken into account in calculating the fitness value of a new individual. Table II illustrates an attempt to achieve in lambda-repressor a sequence with more structurally stabilizing characteristics than the wildtype. These include stronger hydrophobic packaging by increasing the overall composition of hydrophobic aminoacids (I,L,V,A,H,M,F,W,T,P) and more stable helical regions by increasing the number of helix preferring amino acids (A,L,M,E; [5]) while simultaneously conserving solvent accessibility [6], packing of seven critical core residues [7] and overall secondary structure (sometimes leading to compensatory mutations).

TableII. Engineering of the N-terminal half of lambda repressor

```
......................................................................................................
.....................................................................I..........V............E....R...
......F............................................................................L..............TK..
...........................................................H.....................................W....
................................E......I.......R.......P....................C.........................
.M.........................D.D........E......R.....D.....K........EM..................I.....K........V
......FS.......................................S....R......F............L..........LC.............TK..
...........................V...F......E........S...............I.....L..A.......V...N.......IE....RK.V
.A..N......V............N.......LK.....I.......S...............I.....L..A.......V...N..F....IE....RK.V
.A..R......V............I......NLK.....I..CE.D.S...Y...........I....MF.....I......VV.......P........S.
```

The wild type sequence is represented by dots; mutated amino acids are indicated by capital letters. Each new row represents an individual appearing in the population which is fitter than all before.

3.2. Three-dimensional grid simulation

The stability of a protein fold may also be investigated in three-dimensional simulations. The genetic algorithm starts with random conformations. Selection for fitter and fitter individuals leads in an evolving manner to a proper protein fold. This is illustrated in Figure 1 for the *ab initio* folding of a four membered beta strand bundle on a tetrahedral grid that we first investigated. The vast conformational space was searched in only 8 hours processing time on a VAX 3200 workstation. Different simulations showed that global, unspecific forces like overall hydrophobicity (modeled by the scatter of the residues of the model protein around the center of mass) were more critical for a stable fold than specific and particular (e.g. electrostatic) point interactions. Loop lengths close to or greater than those of the secondary structural elements resulted in folding instability.

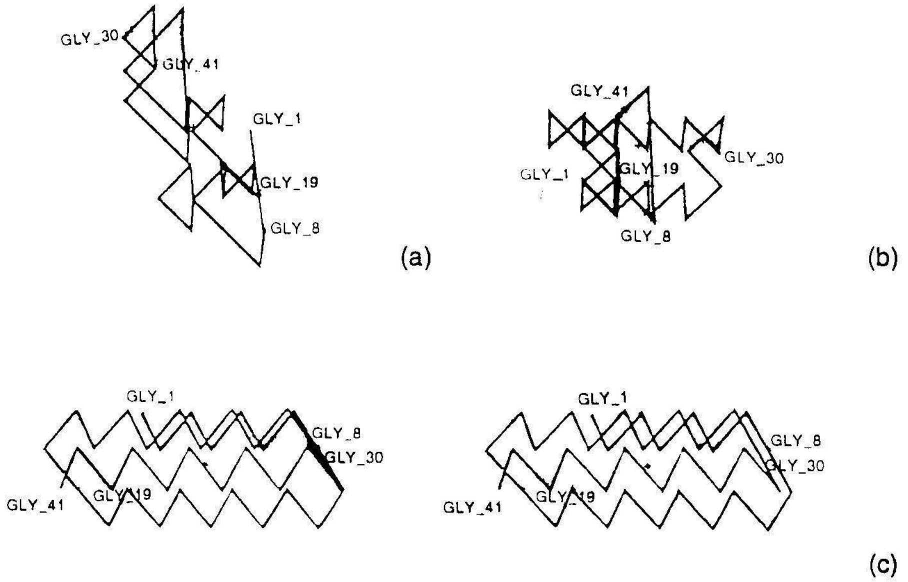

Figure 1. *Ab initio* folding of a four beta bundle model protein with eight-residue long strands and three-residue interconnecting loops: (a) random start; (b) intermediate; (c) stereo picture of the final bundle fold. Within the strands, a selection was imposed from the four possible directions on the tetrahedral grid such that a residue is in *trans* relative to the residue preceding it (zig-zag pattern). Clashes (atom overlap) in the three-dimensional structure lead to a heavy loss in fitness. The selection tried to minimize the scatter around (distance from) the center of mass for all residues, mimicking the global and attractive hydrophobicity of the protein core.

3.3. Grid-free model of a beta bundle

Simulations which use grid-free coordinates are currently under investigation. The expanded conformational space enables a more detailed simulation, avoids possible conformational biases due to grid type [8] and allows closer similarity to real structures; however, it is also a greater computational challenge due to the increased search space. The complete backbone of the proteins is modeled in the simulations, including Cα, N, C and O. The simulation starts with random chain conformations. The fitness function selects against clashes (no van der Waals overlaps between residues); for close scatter around the center of mass according to the hydrophobicity of the amino acids(loop residues are assumed to be hydrophilic); and for maximizing the number of backbone hydrogen-bonds in secondary structural regions without dictating any specific bonding pattern as found in helices or strands. In addition a high residue propensity for a given conformational state can be included in the simulation fitness such that the importance of this propensity for the stability of the overall fold can be investigated. Figure 2a illustrates the result for the folding simulation of a model protein made up of four beta-strands. The simulation shown terminates in a bundle-like conformation formed by the extended beta-strands which was only possible if a very high beta-strand propensity was present throughout the simulation for residues forming strands. In contrast the simulation terminates in a coiled and thus more compact structure if a very weak beta-strand propensity is utilized (Figure 2b; similar fitness function used but beta strand propensity changed). More intermediate cases can be modeled to test which parts of the structure remain stable if several residues are mutated from high to low beta-propensity.

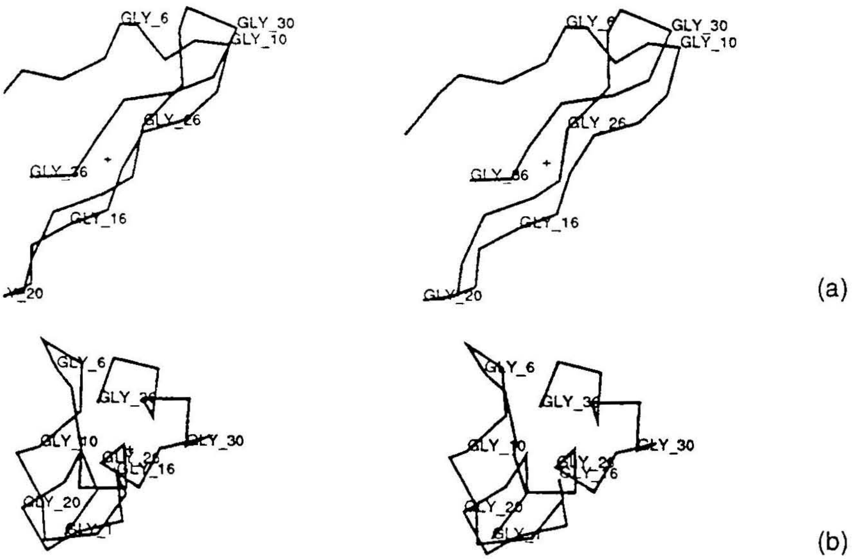

Figure 2. (a) Stereo picture of a grid-free four beta-strand folding simulation predefining high beta-strand propensity for each residue in the beta structures. (b) A folding simulation similar to (a) but each residue has only a very weak beta-strand propensity. Only the C_α trace is shown though all heavy mainchain atoms were included in the simulations.

3.4. Alpha helical structures

The stability of alpha helical structures was also studied. The fit of individual helices to those from experimental structures is good (RMS distance < 1.5 Å over the mainchain atoms N,C_α,C and O). A similar fitness function used in the grid-free beta bundle simulations was also employed here. Despite the general fitness function alpha helices were achieved without predefining for each residue a high propensity for the helical conformation; however, the hydrophobicity of the residues must be distributed according to an amphipathic wheel [9].

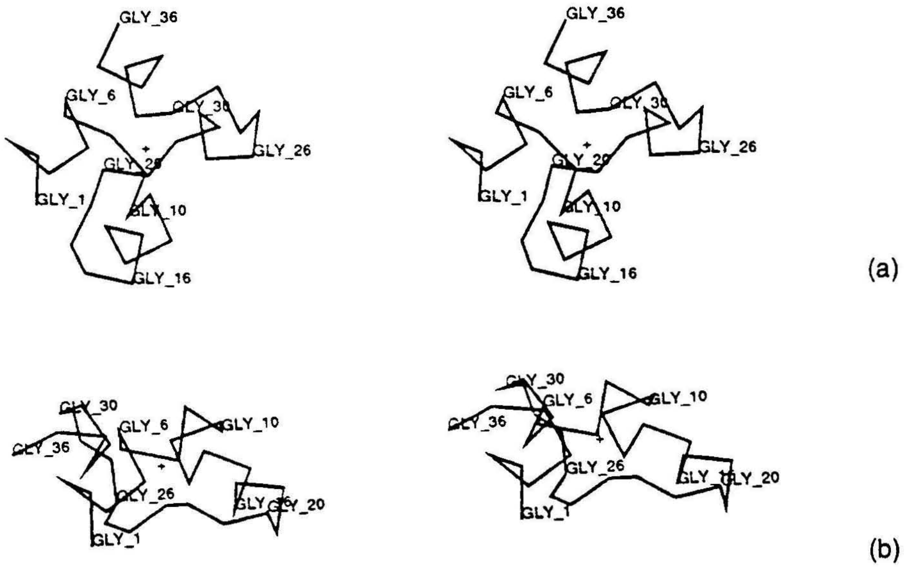

Figure 3. (a) Stereo picture of a grid-free four-helix bundle folding simulation assuming strongly hydrophobic residues to be present at the center of the hydrophobic side in a helical wheel ([10]; positions 1,5,8.. of an alpha helix) (b) A similar simulation as in (a), but assuming the entire hydrophobic half of the helical wheel to be evenly populated by hydrophobic residues; the other wheel half is always assumed to be hydrophilic. Only the C_α trace is shown, but all heavy mainchain atoms were included in the simulations.

The genetic algorithm simulation investigated two different strategies: many but weaker hydrophobic residues on the hydrophobic side of the helical wheel or only a few strong hydrophobic residues. The latter possibility reliably led to the formation of helices (Figure 3a) in the simulation while the former was far less effective (Figure 3b). These folding results are consistent with the folding experiments on apomyoglobin [11] which support the prior significance of hydrophobic interactions across secondary structural association surfaces in achieving the proper protein fold.

The simulation in Figure 3a mimicks an important native fold of many protein structures, a four helical bundle [12]. Currently an investigation of which forces most effectively stabilize this overall structure in simulations using different fitness functions representing helix / helix interactions in detail is underway.

4. DISCUSSION

The examples shown here show the potential of genetic algorithms as a new tool to study protein stability through versatile applications involving the stability of protein cores, secondary structures and loops. Potentially useful amino acid mutations and substitutions fulfilling many criteria simultaneously for engineering and stability studies can be identified by genetic algorithm applications as illustrated in the early examples given here. General principles can also be tested to improve engineering and understanding of structures like helices or loops. Experimentally solved structures may be compared with the models from the grid-free simulations which allow folding of proteins in a much less restrained space [8].

Genetic algorithms were originally borrowed from nature to solve engineering problems and were applied in artificial intelligence soon thereafter. In tackling protein folding problems one turns back again to the natural algorithm. The many degrees of freedom proteins display in folding make *ab initio* simulations by molecular dynamics computationaly expensive while the genetic approach may be able to provide with relatively small computational effort valuable suggestions for protein design.

5. REFERENCES

1. Goldberg,D. E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley Publ., Reading, Mass.
2. McCammon,J.A. and Harvey,S.C. (1988) *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, New York.
3. Dandekar,T. and Argos,P. (1992) Protein Engineering 5, in press.
4. Lim, W.A. and Sauer,R.T. (1989) Nature (London) 339, 31-36.
5. Chou,P.Y. and Fasman,G.D. (1978) *Advances in Enzymology* (Meister,A.,ed.), Vol. 47, 45-148. John Wiley & sons, New York.
6. Janin,J., Wodak,S., Levitt,M. and Maigret,B. (1978) J.Mol.Biol. 125, 357-386.
7. Lim,W.A. and Sauer,R.T. (1991) J.Mol.Biol. 219, 359-376.
8. Gregoret,L.M. and Cohen,F.E. (1991) J.Mol.Biol. 219, 109-122.
9. Schiffer, M. and Edmundson,A.B. (1967) Biophys.J. 7, 121-135.
10. Eisenberg,D., Weiss,R.M. and Terwilliger, Proc.Natl.Acad.Sci. USA 87, 6388-6392.
11. Hughson,F.M., Barnick,D. and Baldwin,R.L. (1991) Biochemistry 30, 4113-4118.
12. Argos,P., Rossmann M.G., Johnson,J.E. (1977) Biochem. Biophys. Res. Comm. 75, 83-86.