In Thomas Bäck, Zbigniew Michalewicz, and Hiroaki Kitano, eds., Proceedings of the Third IEEE Conference on Evolutionary Computation, Piscataway New Jersey: IEEE Service Center, 1996.

Hybrid Genetic Algorithms for Minimization of a Polypeptide Specific Energy Model

Laurence D. Merkle
Gary B. Lamont
Department of Electrical and Computer Engineering
Graduate School of Engineering
Air Force Institute of Technology
Wright-Patterson AFB, OH 45433
{lmerkle, lamont}@afit.af.mil

George H. Gates, Jr.
Ruth Pachter
Wright Laboratory
3005 P St., Ste. 1
Wright-Patterson AFB, OH 45433-7702
{gatesgh, pachterr}@ml.wpafb.af.mil

Abstract— **A hybrid genetic algorithm for polypeptide structure prediction is proposed which incorporates efficient gradient-based minimization directly in the fitness evaluation. Fitness is based on a polypeptide specific potential energy model. The algorithm includes a replacement frequency parameter which specifies the probability with which an individual is replaced by its minimized counterpart. Thus, the algorithm can implement either Baldwinian, Lamarckian, or probabilistically Lamarckian evolution.**

Experiments are described which compare the effectiveness of the genetic algorithm with and without the local minimization operator, and for various probabilities of replacement. The experiments apply the techniques to the minimization of the ECEPP/2 energy model for [Met]-Enkephalin.

Using fitness proportionate selection, the hybrid approaches obtain better energies (and better basins of attraction) than the standard genetic algorithm, and often find the global minimum. When tournament selection is used, the results are qualitatively similar, except that the hybrid approaches are prone to premature convergence.

I. Introduction

The prediction of an arbitrary polypeptide's native conformation (i.e. molecular structure) given only its amino acid sequence is beyond current capabilities, but has numerous potential applications [3]. This structure prediction problem is commonly referred to as the *protein folding problem*. Efforts to solve it nearly always assume that the native conformation corresponds to the global minimum free energy state of the system. Given this assumption, a necessary step in solving the problem is the development of efficient global energy minimization techniques. This is a difficult optimization problem because of the non-linear and multi-modal nature of the energy function. The pentapeptide [Met]-Enkephalin, for example, is estimated to have more than 10^{11} locally optimal conformations. Energy minimization is discussed in slightly more detail in Section II. Also, Vásquez et al. [28] recently reviewed the literature of polypeptide conformational energy calculations.

One class of optimization algorithms which has been applied to the energy minimization problem is that of genetic algorithms (GAs), which are described elsewhere (e.g. Goldberg [8], Holland [11], or Michalewicz [19]). The energy models to which GAs have been applied vary from lattice representations [4; 27] to simplified continuum proteins [12; 13; 24; 25], fixed backbones [22; 26], polypeptide-specific full-atom models [14; 17], and general full-atom models [6; 18; 22].

In some cases (e.g. [14; 26]), the genetic algorithm performs a search of conformations constructed from a library of frequently occurring locally optimal single residue conformations (*rotamers*). This approach may be viewed as a sequentially hybrid approach, in which efficient local optimization of single residue conformations precedes global optimization via genetic algorithm of the overall polypeptide conformation.

Similarly, McGarrah and Judson [17] use a build-up approach including step-wise local minimization to construct their initial population. Their hybrid algorithm also periodically performs local minimization, and uses the resulting energies as the fitnesses of the corresponding individuals. The individuals are never altered following the local minimization. This is in contrast to one of the algorithms studied earlier by Judson et al. [12] in which individuals are always replaced by their locally optimized structures. Unger and Moult [27] propose a hybrid, similar to the latter, in which each individual undergoes 20 steps of simulated annealing before selection is performed.

Merkle et al. propose a hybrid genetic algorithm which incorporates efficient gradient based minimization directly in the fitness evaluation, which is based on a general full-atom potential energy model [18]. The algorithm includes a *replacement frequency* parameter p_r which specifies the probability with which an individual is replaced by its minimized counterpart. Thus, the algorithm can implement either Baldwinian ($p_r = 0$) or Lamarckian ($p_r = 1$) evolution [29], or more generally probabilistically Lamarckian ($0 \leq p_r \leq 1$) evolution. Here we describe a variation on that algorithm which is based on a polypeptide specific full-atom potential (Section II). We also describe experiments comparing the effectiveness of the hybrid genetic algorithm to that of the Monte Carlo-minimization algorithm proposed by Li et al. [15]. We test versions of the genetic algorithm with and without the local minimization operator, and with various probabilities of replacement for the algorithm with the local minimization operator (Section III). Conclusions are presented in Section IV, and Section V discusses directions for future research.

II. Methodology

In this section we discuss the objective function associated with our polypeptide energy minimization application (Section II.A), as well as the encoding scheme (Section II.B). Finally, we discuss the hybridization of the genetic algorithm, which uses SUMSL to perform efficient local minimization (Section II.C).

A. Objective Function

The objective function, which we seek to minimize, is the ECEPP/2 potential [2]

$$\begin{aligned} E = & \sum_{(i,j,k,l)\in\mathcal{D}} \left(\frac{U_{0_{ijkl}}}{2}\right)(1 \pm \cos(n_{ijkl}\Theta_{ijkl})) + \\ & \sum_{(i,j)\in\mathcal{N}} \epsilon_{ij}\left[F_{ij}\left(\frac{r_0}{r_{ij}}\right)^{12} - 2.0\left(\frac{r_0}{r_{ij}}\right)^{6}\right] + \\ & \sum_{(i,j)\in\mathcal{N}} \left[332.0\left(\frac{q_i q_j}{D r_{ij}}\right)\right] + \\ & \sum_{(i,j)\in\mathcal{H}} \epsilon_{ij}\left[\left(\frac{r_0}{r_{HX}}\right)^{12} - 2.0\left(\frac{r_0}{r_{HX}}\right)^{10}\right] \end{aligned} \tag{1}$$

where the four terms represent the energy due to dihedral angle deformation, non-bonded interactions, electrostatic interactions, and hydrogen bond energy respectively. Specifically,

- $\mathcal{D}$ is the set of 4-tuples defining ω and χ dihedrals,
- $\mathcal{N}$ is the set of non-bonded atom pairs,
- $\mathcal{H}$ is the set of hydrogen bonding atom pairs,
- r_{HX} is the donor-acceptor distance,
- r_{ij} is the distance between atoms i and j,
- Θ_{ijkl} is the dihedral formed by atoms i, j, k, and l,
- q_i is the partial atomic charges of atom i,
- the $U_{0_{ijkl}}$'s, n_{ijkl}'s, F_{ij}'s, ϵ_{ij}'s, r_0's, and D are empirically determined constants.

B. Encoding Scheme

The primary determinants of a protein's 3-D structure, and thus the energetics of the system, are its independent dihedral angles [28]. Our genetic algorithm operates on individuals which encode these dihedral angles [6].

Each individual is a fixed length binary string encoding the independent dihedral angles of a polypeptide conformation. The decoding function used is the affine mapping $D : \{0,1\}^{10} \rightarrow [-\pi, \pi]$ of 10 bit subsequences to dihedral angles such that

$$D(a_1, a_2, \ldots, a_{10}) = -\pi + 2\pi \sum_{j=1}^{10} a_j 2^{-j}. \tag{2}$$

This encoding yields a precision of approximately one third of one degree.

The particular biomolecule investigated here is the pentapeptide [Met]-enkephalin. This molecule is chosen because it has been used as a test problem for many other energy minimization investigations (e.g. [14; 20]), and its minimum energy conformation with respect to the ECEPP/2 energy model is known. Twenty-four dihedral angles determine [Met]-enkephalin's structure, hence the string length is 240.

C. Hybrid Genetic Algorithm

In the context of constrained optimization problems, Orvosh and Davis [21] propose replacing infeasible individuals by their repaired counterparts with probability $p_r = 0.05$. Pseudocode for this algorithm is shown in Figure 1. This algorithm may be viewed as probabilistically

```
initialize();
for (gen = 0; gen < max_gen; gen++) {
   for (i = 0; i < pop_size; i++) {
      temp = pop[i];
      local_min(temp);
      pop[i].fitness = temp.fitness;
      if (Rand() < p_r) pop[i] = temp; }
   select();
   recombine();
   mutate(); }
```

Fig. 1. Probabilistically Lamarckian genetic algorithm pseudocode

Lamarckian. Alternatively, one may view the local minimization operator as a repair operator in the sense that it maps individuals to the "feasible region," where the nonlinear equality constraint to be satisfied is $\nabla E = 0$.

The local minimization method used in this investigation is the Secant-type Unconstrained Minimization SoLver (SUMSL) [7]. SUMSL uses a secant approximation to the Hessian, based on explicit objective function and gradient information. The ECEPP/2 software integrates SUMSL with the energy function and gradient calculations. We use a convergence tolerance of 10^{-5} kcal/mol, and a maximum number of iterations of 1000. No more than 200 iterations were ever necessary to meet the convergence criterion.

III. Results

In this section we present the results of experiments in which we empirically compare the minimum energies and associated conformations found by the algorithm described in Section II to those reported for the Monte Carlo-minimization algorithm [20]. Specifically, we present results for the standard genetic algorithm (denoted SGA), the SGA followed by local minimization of the best individual (denoted SGA+SUMSL), and proba-

bilistically Lamarckian genetic algorithms using various replacement probabilities $p_r \in \{0, 0.05, 0.10, 1.00\}$ (denoted Baldwinian, $p_r = 0.05$, $p_r = 0.10$, and Lamarckian, respectively). The experiments are performed on SPARC workstations using the 1990 version of GENESIS [9], modified to include the local minimization operator. The input parameters are as given in the typical input file shown in Figure 2. Five independent runs of each algorithm are

```
      Experiments = 1
     Total Trials = 50000
  Population Size = 50
 Structure Length = 240
   Crossover Rate = 0.65
    Mutation Rate = 0.003
   Generation Gap = 1.0
   Scaling Window = 1
  Report Interval = 1
 Structures Saved = 1
Max Gens w/o Eval = 10
    Dump Interval = 0
      Dumps Saved = 0
          Options = ce
      Random Seed = 987654321
```

Fig. 2. Typical GENESIS run time parameter file

performed, using the same set of five random seeds for each algorithm.

A. Fitness Proportionate Selection

The average minimum energies obtained in each generation are shown in Figure 3, except those for SGA+SUMSL. The results for the latter are identical to those for the SGA except in the final generation. The

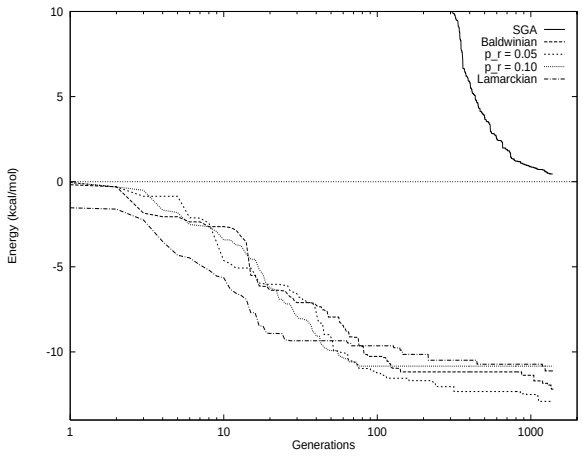

Fig. 3. Avg. min. energy vs. generation using fitness prop. selection

means and standard deviations of the final generation

Table 1. Final min. energies (kcal/mol) using fitness prop. selection

Algorithm	Mean	Std. Dev.
SGA	0.4561	3.0156
SGA+SUMSL	-1.6720	3.2647
Baldwinian	-12.188	1.2500
$p_r = 0.05$	-12.905	0.0000
$p_r = 0.10$	-10.835	1.3427
Lamarckian	-11.118	1.6326

Table 2. Final min. energies (kcal/mol) using tournament selection

Algorithm	Mean	Std. Dev.
SGA	2.6003	3.41
SGA+SUMSL	-4.0911	2.68
Baldwinian	-9.7532	0.80
$p_r = 0.05$	-10.458	1.77
$p_r = 0.10$	-11.231	1.77
Lamarckian	-10.951	1.18

minimum energies are shown in Table 1.

The final energies obtained by the SGA+SUMSL algorithm average about 2 kcal/mol lower than those obtained by the SGA. In turn, the probabilistically Lamarckian algorithms obtain final energies which average about 10 kcal/mol lower than those of the SGA+SUMSL. The latter difference is significant at the 0.005 level as determined by the Kruskal-Wallis H Test [1].[1]

Of the 20 probabilistically Lamarckian runs, 10 found the accepted global minimum (including all of the runs for $p_r = 0.05$), which has been obtained previously using Monte Carlo-minimization [15; 20]. One of the other runs identified a unique conformation with an energy within 0.0001 kcal/mol of the global minimum, having an rms deviation of 2.589 Å relative to the global minimum. None of the SGA or SGA+SUMSL runs obtained conformations with energies within 6 kcal/mol of the global minimum.

B. Tournament Selection

In each generation, most individuals have energies close to the best individual, but there are a few individuals with much larger energies. The presence of high energy individuals has the effect of reducing the selective pressure of fitness proportionate selection. Thus, we also perform the experiments using binary tournament selection, as implemented by Merkle et al [18]. The minimum energies obtained are shown in Figure 4 and in Table 2. The results are qualitatively similar to those obtained using fit-

[1] The final energies obtained by the various probabilistically Lamarckian algorithms in these experiments are not different from each other at any interesting level of statistical significance.

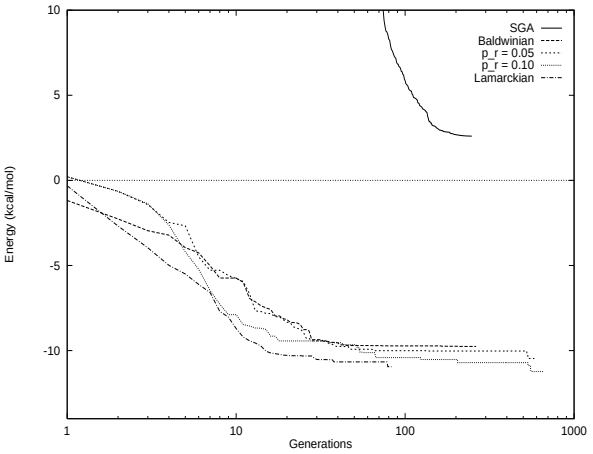

Fig. 4. Avg. min. energy vs. generation using tournament selection

Table 3. Trials prior to 99% convergence using tournament selection

Algorithm	Mean	Std. Dev.
SGA	6258	1209
Baldwinian	5865	2977
$p_r = 0.05$	5676	1654
$p_r = 0.10$	4964	2085
Lamarckian	1750	738

ness proportionate selection. The final energies obtained using SGA+SUMSL average about 7 kcal/mol lower than those obtained using SGA, and the final energies obtained using the probabilistically Lamarckian algorithms average about 6 kcal/mol lower than those obtained using SGA+SUMSL. The latter is significant at the 0.005 level. Three of the runs found the accepted global minimum.

The final energies obtained by the SGA using tournament selection average about 2 kcal/mol higher than those for the SGA using fitness proportionate selection. Also, the final energies obtained by the probabilistically Lamarckian algorithms using tournament selection average about 1.2 kcal/mol higher than those obtained using fitness proportionate selection. The latter difference is significant at the 0.01 level. The higher final energies of the tournament selection runs are primarily due to premature convergence. The means and standard deviations of the number of trials performed prior to obtaining 99% convergence are shown in Table 3. For comparison, the runs using fitness proportionate selection obtain between 76.0% and 94.7% convergence for 50000 trials.

While the final energies obtained by the SGA with tournament selection are higher than those for the SGA with fitness proportionate selection, it is interesting to note that converse holds for the SGA+SUMSL algorithm. The final energies obtained by the SGA+SUMSL using tournament selection average about 2.6 kcal/mol lower than those for the SGA+SUMSL using fitness proportionate selection. Neither difference is statistically significant for the experiments performed here.

IV. Conclusions

While Lamarckian genetic algorithms obtain good solutions for some applications (e.g. [12]), it has also been shown that Baldwinian algorithms are superior for other applications [29], while probabilistically Lamarckian approaches are superior for others [21]. All of the probabilistically Lamarckian algorithms used in this investigation obtained significantly better energies than both the SGA and the SGA followed by local minimization. This supports earlier results [18] suggesting that the local minima in the energy landscape of [Met]-Enkephalin occur somewhat regularly. For the experiments performed here, the replacement frequency was not found to have a significant effect on the final energy.

The final energies obtained by the probabilistically Lamarckian algorithms using fitness proportionate selection are significantly lower than those obtained using tournament selection, due to the premature convergence of the latter. Also, the fitness proportionate selection runs obtained the global minimum significantly more frequently than the tournament selection runs.

V. Future Directions

The premature convergence of the tournament selection experiments performed here suggests that the selective pressure of tournament selection is too high for this application. This observation, together with the low convergence rates of the fitness proportionate selection, suggests the investigation of selection operators with selective pressure intermediate to those used here. McGarrah and Judson describe such a selection operator which has been used successfully in similar applications [17].

The success of these algorithms for [Met]-Enkephalin suggests their application to larger polypeptides. Such application requires computational resources which are only available through the use of highly scalable architectures. We have previously used such architectures successfully for protein structure prediction via island model genetic algorithms [6]. We are now studying parallel designs of the hybrid algorithms presented here.

References

[1] Arnold O. Allen. *Probability, Statistics, and Queueing Theory: With Computer Science Applications.* Computer Science and Scientific Computing. Academic Press, Inc., San Diego, California, 1990.

[2] M. Jean Browman, Lucy M. Carruthers, Karen L. Kashuba, Frank A. Momany, Marcia S. Pottle, Su-

san P. Rosen, and Shirley M. Rumsey. ECEPP/2: Empirical conformational energy program for peptides. Quantum Chemistry Program Exchange QCPE Program No. 454, Indiana University, Department of Chemistry, (812)855-4784, 1983. Write-up by Gerald F. Endres. Resubmitted by H. A. Scheraga.

[3] Hue Sun Chan and Ken A. Dill. The protein folding problem. *Physics Today*, pages 24–32, February 1993.

[4] Thomas Dandekar and Patrick Argos. Potential of genetic algorithms in protein folding and protein engineering simulations. *Protein Engineering*, 5(7):637–645, 1992.

[5] Stephanie Forrest, editor. *Proceedings of the Fifth International Conference on Genetic Algorithms*, San Mateo CA, July 1993. Morgan Kaufmann Publishers, Inc.

[6] George H. Gates, Jr., Ruth Pachter, Laurence D. Merkle, and Gary B. Lamont. Parallel simple and fast messy GAs for protein structure prediction. In *Proceedings of the Intel Supercomputer Users' Group 1995 Annual North America Users Conference*, Beaverton, Oregon, 1995. Intel Supercomputer Systems Division.

[7] David M. Gay. Subroutines for unconstrained minimization using a model/trust-region approach. *ACM Transactions on Mathematical Software*, 9(4):503–524, December 1983.

[8] David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison–Wesley Publishing Company, Reading MA, 1989.

[9] John J. Grefenstette. A user's guide to Genesis. Technical report, Vanderbilt University, Nashville TN, 1986.

[10] John J. Grefenstette, editor. *Genetic Algorithms and Their Applications: Proceedings of the Second International Conference on Genetic Algorithms*, Hillsdale, New Jersey, July 1987. Lawrence Erlbaum Associates, Publishers.

[11] John H. Holland. *Adaptation in Natural and Artificial Systems.* MIT Press, Cambridge, MA, First MIT Press edition, 1992.

[12] R. S. Judson, M. E. Colvin, J. C. Meza, A. Huffer, and D. Gutierrez. Do intelligent configuration search techniques outperform random search for large molecules? *International Journal of Quantum Chemistry*, 44:277–290, 1992.

[13] Richard S. Judson. Teaching polymers to fold. *The Journal of Physical Chemistry*, 96(25):10102, 1992.

[14] Scott M. LeGrand and Kenneth M. Merz Jr. The application of the genetic algorithm to the minimization of potential energy functions. *Journal of Global Optimization*, 3:49–66, 1991.

[15] Zhenqin Li and Harold A. Scheraga. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Science USA*, 84:6611–6615, 1987.

[16] Reinhard Männer and Bernard Manderick, editors. *Parallel Problem Solving from Nature, 2*, Amsterdam, September 1992. North-Holland.

[17] D.B. McGarrah and R.S. Judson. Analysis of the genetic algorithm method of molecular conformation determination. *Journal of Computational Chemistry*, 14(11):1385–1395, 1993.

[18] Laurence D. Merkle, Robert L. Gaulke, George H. Gates, Jr., Gary B. Lamont, and Ruth Pachter. Hybrid genetic algorithms for polypeptide energy minimization. In *Applied Computing 1996: Proceedings of the 1996 Symposium on Applied Computing*, New York, 1996. The Association for Computing Machinery.

[19] Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs.* Springer, second edition, 1994.

[20] Akbar Nayeem, Jorge Vila, and Harold A. Scheraga. A comparative study of the simulated-annealing and Monte Carlo-with-minimization approaches to the minimum-energy structures of polypeptides: [Met]-Enkephalin. *Journal of Computational Chemistry*, 12(5):594–605, 1991.

[21] David Orvosh and Lawrence Davis. Shall we repair? genetic algorithms, combinatorial optimization, and feasibility constraints. In Forrest [5], page 650.

[22] Steffen Schulze-Kremer. Genetic algorithms for protein tertiary structure prediction. In Stender [23], pages 129–149.

[23] Joachim Stender, editor. *Parallel Genetic Algorithms: Theory and Applications.* IOS Press, Amsterdam, 1993.

[24] S. Sun, N. Luo, R. L. Ornstein, and R. Rein. *Biophysics Journal*, 62:104, 1992.

[25] Shaojian Sun. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Science*, 2:762–785, 1993.

[26] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rapid determination of protein side chain conformations. *Journal of Biomolecular Structure & Dynamics*, 8(6):1267, 1991.

[27] Ron Unger and John Moult. Genetic algorithms for protein folding simulations. *Journal of Molecular Biology*, 231:75–81, 1993.

[28] Maximiliano Vásquez, G. Némethy, and H. A. Scheraga. Conformational energy calculations on polypeptides and proteins. *Chemical Reviews*, 94:2183–2239, 1994.

[29] Darrell Whitley, V. Scott Gordon, and Keith Mathias. Lamarckian evolution, the Baldwin effect and function optimization. In Männer and Manderick [16], pages 6–15.