Proceedings of the Eighth SIAM Conference on Parallel Processing for Scientific Computing, Philadephia, PA: SIAM, 1997.

Case Studies in Protein Structure Prediction with Real-valued Genetic Algorithms

Charles E. Kaiser* Laurence D. Merkle*† Gary B. Lamont*
George H. Gates, Jr.‡ Ruth Pachter‡

Abstract

Accurate and reliable protein structure prediction (PSP) eludes researchers primarily because the search for the minimum energy conformer is computationally intractable. This research discusses the application of several distinct genetic algorithms (GAs) as optimum seeking techniques for PSP problems. The effectiveness and efficiency of each algorithm is studied empirically. The specific algorithmic designs studied are a *farming model* parallel hybrid simple GA, a REal-valued Genetic Algorithm with Limited Constraints (REGAL) incorporating domain knowledge, and a *island model* parallel REGAL with a novel migration operator (Para-REGAL).

1 Introduction

The prediction of an arbitrary protein's native conformation (i.e. molecular structure) given only its amino acid sequence is beyond current computational capabilities, but has numerous potential applications [3]. Efforts to solve such *protein structure prediction (PSP)* problems via *energy minimization* assume the native conformation corresponds to the global minimum free energy state of the system. The associated energy landscape is non-linear and massively multimodal. Furthermore, the basins of attraction are thought to be very "narrow," in the sense that the difference in energy between local minima is typically small compared to the height of the barriers between them. Thus, a necessary step in solving the general PSP problem is development of efficient global minimization techniques.

One probabilistic optimum seeking technique which has been applied to the PSP problem is the genetic algorithm (GA), which is described elsewhere (see, for example, Bäck [1]). The energy models to which the GA has been applied include lattice representations [4, 22], simplified continuum proteins [9, 8, 20], fixed backbones [18, 21], polypeptide-specific full-atom models [14, 15], and general full-atom models [6, 16, 18]. This research discusses the application of several variants of the GA as optimum seeking techniques for PSP problems. The effectiveness and efficiency of each algorithm is studied empirically.

Section 2 briefly discusses the general full-atom model used in this research. Experiment I, which is described in Section 3, evaluates a *farming model* parallel hybrid simple GA. Section 4 proposes a *REal-valued Genetic Algorithm with Limited Constraints (REGAL)* and describes Experiment II, which investigates the impact of incorporating domain knowledge in the constraints of the optimization problem. Experiment III, an evaluation

*Department of Electrical and Computer Engineering, Graduate School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB, OH 45433

†Center for Plasma Theory and Computation, Phillips Laboratory, Kirtland AFB, NM 87117

‡Wright Laboratory, 3005 P St., Ste. 1, Wright-Patterson AFB, OH 45433-7702

of exogenous parameters for REGAL, is discussed elsewhere [11].[1] Finally, Section 5 proposes an *island model* parallel REGAL with a novel probabilistic migration operator (Para-REGAL), and describes Experiment IV.

2 Protein Structure Prediction (PSP)

This section briefly discusses the representation scheme used in this research to describe macromolecular geometries. It also defines the objective function used by the energy minimization approaches to PSP problems described in Sections 3, 4, and 5.

Geometric descriptions of macromolecular structure typically use either Cartesian (rectilinear) or *internal* coordinates. The latter system specifies chosen *bond lengths*, *bond angles*, and *dihedral angles* sufficient to uniquely define the geometry of the molecule. Protein bond lengths and angles are substantially less flexible than dihedral angles. Consequently, many optimum seeking techniques for PSP problems, including those used in this research, reduce the dimensionality of the problem by fixing the bond lengths and angles at their "equilibrium" values during at least some phase of the search.

The objective function used in this research is based on the CHARMM [2] energy function. Because bond lengths and bond angles are fixed, only the three terms representing the energy due to dihedral angle deformation, non-bonded interactions, and 1-4 interactions are included:

$$
\begin{aligned}
&(1)E = \\
&\quad \sum_{(i,j)\in\mathcal{B}} K_{r_{ij}}(r_{ij}-r_{eq})^2 + \sum_{(i,j,k)\in\mathcal{A}} K_{\Theta_{ijk}}(\Theta_{ijk}-\Theta_{eq})^2 + \\
&\quad \sum_{(i,j,k,l)\in\mathcal{D}} K_{\Phi_{ijkl}}[1+\cos(n_{ijkl}\Phi_{ijkl}-\gamma_{ijkl})] + \\
&(2)\quad \sum_{(i,j)\in\mathcal{N}} \left[\left(\frac{A_{ij}}{r_{ij}}\right)^{12} - \left(\frac{B_{ij}}{r_{ij}}\right)^{6} + \frac{q_i q_j}{4\pi\varepsilon r_{ij}}\right] + \frac{1}{2}\sum_{(i,j)\in\mathcal{N}'} \left[\left(\frac{A_{ij}}{r_{ij}}\right)^{12} - \left(\frac{B_{ij}}{r_{ij}}\right)^{6} + \frac{q_i q_j}{4\pi\varepsilon r_{ij}}\right] .
\end{aligned}
$$

In Equation 2, $\mathcal{D}$ is the set of atom 4-tuples defining dihedral angles, $\mathcal{N}$ is the set of non-bonded atom pairs, $\mathcal{N}'$ is the set of 1-4 interaction pairs, r_{ij} is the distance between atoms i and j, Φ_{ijkl} is the dihedral angle formed by atoms i, j, k, and l, q_i is the partial atomic charge of atom i, the $K_{\Phi_{ijkl}}$'s, γ_{ijkl}'s, A_{ij}'s, B_{ij}'s, and ε are empirically determined constants.

3 Experiment I: Parallel and Distributed Hybrid GAs

Limited PSP studies (e.g. [16]) compare the GENESIS genetic algorithm implementation [7] to a *probabilistically Lamarckian* hybrid GA incorporating conjugate gradient local minimization. Pseudocode for the hybrid algorithm is shown in Figure 1, in which `p_m` is the *probability of minimization*, and `p_r` is the *probability of replacement*. In these studies, the hybrid is found to be significantly more effective in the energy minimization of [Met]-enkephalin due to the somewhat regular occurrence of local minima within the energy landscape. The hybrid is also less efficient due to the function evaluations required for local minimization.

The combinations of p_m, p_r and selection operator which result in the lowest energies are shown in Table 1. The *Baldwinian* algorithm (FPBald) has $p_m = 1.0$ and $p_r = 0.0$,

[1] Experiments I, II, III, and IV are named for consistency with [10].

```
initialize();
for (gen=0 ; gen < max_gen; gen++){
   for (i=0 ; i < pop_size ; i++) {
      temp = pop[i];
      if (Rand() < p_m) 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*

TABLE 1

Hybrid genetic algorithm test cases and final [Met]-enkephalin energies (kcal/mol). μ denotes population size.

Algorithm	Selection	p_m	p_r	$\mu = 20$	$\mu = 50$	$\mu = 100$
FPBald	FP	1.0	0.0	-21.23	-2.62	9.50
FPLam	FP	1.0	1.0	-30.84	-22.26	-15.99
TSLam	TS	1.0	1.0	-28.73	-28.16	-27.68
FPSGA	FP	0.0	0.0	-13.51	23.55	35.21
TSSGA	TS	0.0	0.0	3.43	-19.45	-1.43

while the *Lamarckian* algorithms (FPLam and TSLam) have $p_m = p_r = 1.0$. The remaining algorithms (FPSGA and TSSGA) have $p_m = 0.0$. Algorithms FPBald, FPLam, and FPSGA use fitness proportionate selection, while TSLam and TSSGA use tournament selection.

By definition, a *farming model* parallel hybrid GA (PHGA) has the same search trajectory as the sequential hybrid GA, hence the same effectiveness, and potentially better efficiency. The objective of Experiment I is to empirically characterize the efficiency of the PHGA in terms of overhead, speed-up, and scalability.

Each individual is a fixed length binary string encoding the independent dihedral angles of a [Met]-enkephalin conformation. The decoding function used is $D(a_1, a_2, \ldots, a_{10}) = -\pi + 2\pi \sum_{j=1}^{10} a_j 2^{-j}$.

Experiments for the cases shown in Table 1 are performed using $P \in \{1, 2, 6, 12, 18, 24\}$ processors of an Intel Paragon, with population sizes $\mu \in \{20, 50, 100\}$, and a maximum of $t_{\max} \in \{500, 1000, 1500, 2000\}$ function evaluations. Results are averaged over three executions to account for variable loading of the communications network.

The right hand side of Table 1 presents the final energies after 2000 evaluations (independent of P). The best results for fitness proportionate selection are obtained using $\mu = 20$. In contrast, the lowest energies for tournament selection are obtained with $\mu = 50$. Results for each of the algorithms with $\mu = 20$ are shown in Figure 2. The TSSGA converges prematurely, terminating after 1300 evaluations.

Speedup measures the relative benefit of solving a problem in parallel [13]. Speedup for each PHGA is plotted for $\mu = 20$ in Figure 3. *Efficiency* is the fraction of time for which

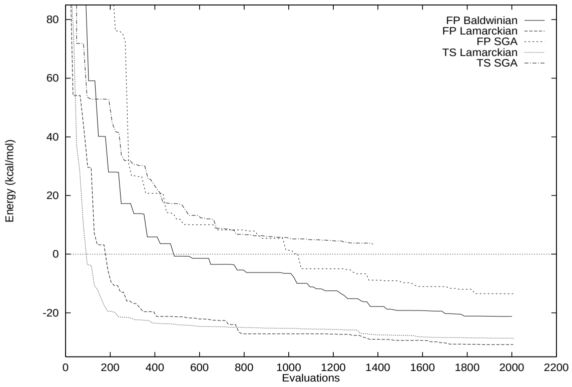

FIG. 2. *Energy of Best Individual, Population Size of 20*

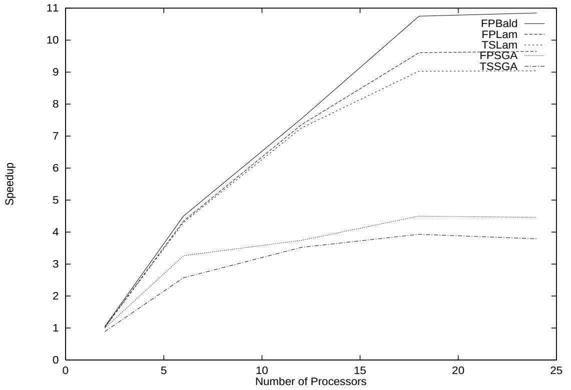

FIG. 3. *Speedup, Population Size 20*

each processor is usefully employed [13]. Efficiency for each PHGA is plotted for $\mu = 20$ in Figure 4. The PHGAs studied in this research are most efficient when the number of

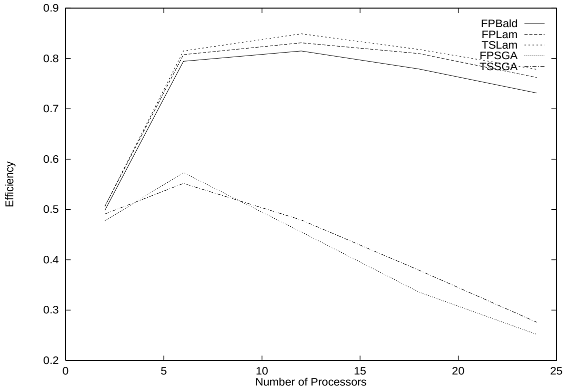

FIG. 4. *Efficiency, Population Size 20*

processors is roughly half the population size. With fewer processors, the idle time of the master processor is significant, while with more processors the idle time of the slave processors becomes substantial.

4 Experiment II: Evaluation of Constraints for PSP

Much is known about macromolecular structure, and about protein structure in particular. Previous applications of GAs to PSP problems neglect much of this knowledge. The objective of Experiment II is to evaluate the impact on effectiveness and efficiency of the incorporation of domain specific knowledge in the GA. Computational experiments are performed for two molecules: [Met]-enkephalin and a 14 residue model of Polyalanine.

The experiment is performed using the GENOCOP III real-valued GA software [17]. GENOCOP maintains a reference population consisting entirely of feasible individuals, as well as the usual search population (see Figure 5). The GENOCOP III real-valued GA software is integrated with domain specific constraints to create a software package called REGAL (REal-valued GA with Limited constraints). Three constraint sets, *none*, *loose*, and *tight*, are developed for each molecule which reflect to varying degrees information obtained from Ramachandran plots [12].

```
t := 0;
initialize_search_pop();
initialize_reference_pop();
while (not termination-condition) do {
   t := t + 1;
   act_on_search_pop();
   if (t mod k = 0) then act_on_reference_pop();
}
```

FIG. 5. *GENOCOP III structure*

Five experiments are performed for each molecule and each constraint set using different random number seeds. Final energies after 10^4 evaluations are compared to those obtained with the GENESIS-based hybrid GA. The experiments are conducted on a variety of platforms, including a 368 node Intel Paragon, 100 and 200 MHz Silicon Graphics workstations, as well as SUN Sparc (2, 5, and 20) and Ultra workstations. Of course, execution times vary with system loading for all platforms.

The hybrid GA is typically more effective than the REGAL approach in minimizing [Met]-enkephalin (Table 2). The overall best energy is obtained by a REGAL experiment, using no constraints and Lamarckian minimization. Tight constraints together with local

TABLE 2
GENESIS and REGAL test cases and final minimum [Met]-enkephalin energies (kcal/mol)

Algorithm	Constraints	Mean	Std. Dev.	RMSD
GENESIS	N/A	-22.58	1.57	4.51
Baldwinian GENESIS	N/A	-22.57	1.62	3.96
Lamarckian GENESIS	N/A	-28.35	1.29	3.33
REGAL	None	-24.92	2.99	4.57
REGAL	Loose	-22.01	2.69	4.25
REGAL	Tight	-23.55	1.69	3.23
Lamarckian REGAL	None	-26.38	2.69	4.40
Lamarckian REGAL	Loose	-24.95	4.23	4.26
Lamarckian REGAL	Tight	-17.71	0.50	5.05

minimization result in poor effectiveness, due to reduced exploration of the search space. Average execution time for Lamarckian GENESIS is approximately 13 hours, while that for REGAL is approximately 2 hours.

In contrast to the situation with [Met]-enkephalin, the REGAL algorithm is typically more effective than the hybrid GA for minimization of the larger molecule Polyalanine (Table 3). Significant improvement is observed when a more conservative *step size* is chosen in the bracketing phase of the conjugate gradient minimization routine. The resulting conformations do not appear to form the expected α-helix secondary structures. With adequate domain knowledge, in the form of tight constraints, REGAL effectively minimizes the energy of the larger molecule. After 150,000 evaluations, the energy value is almost that of the optimal conformation without relaxation of bond lengths and bond angles. The resulting conformations appear to form the expected α-helix secondary structures. Local minimization is not effective when used in conjunction with constraints, due to reduced exploration.

TABLE 3
GENESIS and REGAL final Polyalanine minimum energies (kcal/mol)

Algorithm	Constraints	Mean	Std. Dev.	RMSD
GENESIS	N/A	-93.25	10.85	9.67
Baldwinian GENESIS	N/A	-103.73	16.5	7.36
Lamarckian GENESIS	N/A	-140.60	5.39	12.74
Lamarckian GENESIS w/ conservative step size	N/A	-308.51	8.26	5.03
REGAL	Loose	-336.65	4.50	1.87
REGAL	Tight	-337.64	4.40	0.98
REGAL	Tight w/ relaxed terminals	-338.30	4.24	1.42
REGAL 150K evals	Tight w/ relaxed terminals	-351.76	0.57	1.40
Lamarckian REGAL	Loose	-309.00	8.19	2.70
Lamarckian REGAL	Tight	-316.47	0.0	1.17

Average execution time for Lamarckian GENESIS is approximately 120 hours, while that for REGAL is approximately 4 hours. These data suggest that the REGAL approach scales better than Lamarckian GENESIS.

5 Experiment IV: Evaluation of Para-REGAL

The results of Experiment II indicate that REGAL is an effective optimum seeking technique in applications to PSP problems. In this section, an *island model* parallel REGAL with a novel migration operator is proposed (Para-REGAL).

The search and reference populations are each partitioned into subpopulations. The search subpopulation and reference subpopulation assigned to a particular processor are together called an *island*, and evolve independently of the other islands, with possibly distinct parameters and domain constraints. When a new feasible solution is added to a reference subpopulation, it *emigrates* with probability P_m. The solution *immigrates* to each other island with independent probability P_{cm}. It is included in that island's reference population if it is feasible with respect to the constraints of that island, and the search subpopulation otherwise.

The objective of Experiment IV is to empirically characterize the effectiveness of

Para-REGAL. The experiments are performed with [Met]-enkephalin as the test molecule, $P_m, P_{cm} \in \{0.0, 0.33, 0.66, 1.00\}$ and processor counts $P \in \{4, 16, 32\}$. In each case, one fourth each of the processors use no domain constraints, a reduced set of loose constraints, the full set of *loose* constraints, and the *tight* constraints.

TABLE 4

Para-REGAL test cases and best final minimum energies (kcal/mol)

	P_{cm}			
P_m	0.00	0.33	0.66	1.00
0.00	-22.38	-22.38	-22.38	-22.38
0.33	-21.62	-29.10	-24.33	-25.55
0.66	-21.92	-23.90	-25.79	-26.03
1.00	-23.35	-25.16	-25.45	-22.67

Several of the experiments result is lower energies than those obtained by the best non-Lamarckian sequential REGAL experiments. The lowest final energy is never obtained by a tightly constrained island. This is a consequence of the fact that a solution migrating to a more loosely constrained island is more likely to be included in the reference subpopulation than a solution migrating to a more tightly constrained island.

6 Conclusions

These case studies show beneficial results from the proposed approaches. The hybrid GA is a highly effective optimum seeking technique in applications to PSP problems using the CHARMM energy model. In Experiment I, the farming model parallel hybrid GA is more efficient, considerably reducing wall clock time. As the ratio of processors to population size increases beyond approximately one half, idle time increases.

In Experiment II, results using a real-valued GA implementation demonstrate the feasibility of using domain knowledge to limit the GA's search. Lower energies result when tight domain constraints are used. The constraint sets developed in this research are relatively loose. A biochemist studying a particular molecule may choose to develop tighter constraints.

The results of Experiment IV, involving Para-REGAL, indicate a synergistic relationship between tightly and loosely constrained subpopulations. Better trajectories are observed in subpopulations with looser constraints.

References

[1] T. Bäck, *Evolutionary Algorithms in Theory and Practice*, Oxford University Press, New York, 1996.

[2] B. R. Brooks et al., *Charmm: A program for macromolecular energy, minimization, and dynamic calculations*, Jounal of Computational Chemistry, 4 (1983), pp. 187–217.

[3] H. S. Chan and K. A. Dill, *The protein folding problem*, Physics Today, (1993), pp. 24–32.

[4] T. Dandekar and P. Argos, *Potential of genetic algorithms in protein folding and protein engineering simulations*, Protein Engineering, 5 (1992), pp. 637–645.

[5] D. FOGEL, ed., *Proceedings of the Second IEEE Conference on Evolutionary Computation*, Piscataway NJ, 1995, IEEE Service Center.

[6] G. H. Gates, Jr., R. Pachter, L. D. Merkle, and G. 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] J. J. Grefenstette, *A user's guide to Genesis*, tech. rep., Vanderbilt University, Nashville TN, 1986.

[8] R. S. Judson, *Teaching polymers to fold*, The Journal of Physical Chemistry, 96 (1992), p. 10102.

[9] 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 (1992), pp. 277–290.

[10] C. E. Kaiser, Jr., *Refined genetic algorithms for polypeptide structure prediction*, Master's thesis, Graduate School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH 45433, Dec. 96.

[11] C. E. Kaiser, Jr., G. B. Lamont, L. D. Merkle, G. H. Gates, Jr., and R. Pachter, *Exogenous parameter selection in a real-valued genetic algorithm*, in Proceedings of the Fourth IEEE Conference on Evolutionary Computation, 1997. In review.

[12] C. E. Kaiser, Jr., L. D. Merkle, G. H. Gates, Jr., G. B. Lamont, and R. Pachter, *Real-valued and hybrid genetic algorithms for polypeptide structure prediction*, in Applied Computing 1997: Proceedings of the 1997 Symposium on Applied Computing, New York, 1997, The Association for Computing Machinery. To appear.

[13] V. Kumar, A. Grama, A. Gupta, and G. Karypis, *Introduction to Parallel Computing: Design and Analysis of Algorithms*, The Benjamin/Cummings Publishing Company, Inc., Redwood City, CA, 1994.

[14] S. M. LeGrand and K. M. Merz Jr., *The application of the genetic algorithm to the minimization of potential energy functions*, Journal of Global Optimization, 3 (1991), pp. 49–66.

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

[16] L. D. Merkle, R. L. Gaulke, G. H. Gates, Jr., G. B. Lamont, and R. 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, pp. 396–400.

[17] Z. Michalewicz and G. Nazhiyath, *Genocop III: a co-evolutionary algorithm for numerical optimization*, in Fogel [5], pp. 647–651.

[18] S. Schulze-Kremer, *Genetic algorithms for protein tertiary structure prediction*, in Stender [19], pp. 129–149.

[19] J. Stender, ed., *Parallel Genetic Algorithms: Theory and Applications*, IOS Press, Amsterdam, 1993.

[20] S. Sun, *Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms*, Protein Science, 2 (1993), pp. 762–785.

[21] 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 (1991), p. 1267.

[22] R. Unger and J. Moult, *Genetic algorithms for protein folding simulations*, Journal of Molecular Biology, 231 (1993), pp. 75–81.