POLYPEPTIDE STRUCTURE PREDICTION: REAL-VALUED VERSUS BINARY HYBRID GENETIC ALGORITHMS

Charles E. Kaiser, Jr., Gary B. Lamont, Laurence D. Merkle
Department of Electrical and Computer Engineering
Graduate School of Engineering
Air Force Institute of Technology
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

Key Words: Genetic Algorithms, Hybrid Genetic Algorithms, Polypeptide Structure Prediction, Protein Folding, Real-Valued Genetic Algorithms.

Abstract

Energy minimization efforts to predict polypeptide structures assume their native conformation corresponds to the global minimum free energy state. Given this assumption, the problem becomes that of developing efficient global optimization techniques applicable to polypeptide energy models. This general structure prediction objective is also known as the protein folding problem. Our prediction algorithms, based on general full-atom potential energy models, are expanded to incorporate domain knowledge into the search process. Specifically, we evaluate the effectiveness of a real-valued genetic algorithm exploiting domain knowledge about certain dihedral angle values inorder to limit the search space. We contrast this approach with our hybrid binary genetic algorithms. Various experiments apply these techniques to minimization of the potential energy for the specific proteins [Met]-Enkephalin and Polyalanine using the CHARMM energy model.

1997 ACM 0-89791-850-9 97 0002

1 Introduction

Given only the amino acid sequence for an arbitrary polypeptide, the prediction of it's native conformation (i.e., molecular structure) is beyond current computational capabilities.[1] This structure prediction problem is commonly referred to as the *protein folding problem* and it's solution would be a watershed event in biochemistry with numerous potential applications [3]. 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 nonlinear 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 2. Also, Vásquez et al. [31] has reviewed the literature of polypeptide conformational energy calculations. For detailed insight into the protein folding problem consult [4, 14, 24, 27].

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. Bäck [1], Goldberg [8], Holland [9], Michalewicz [20]). The energy models to which GAs have been applied vary from lattice representations [5, 30] to simplified continuum proteins [11, 28], fixed backbones [26, 29], polypeptide-specific full-atom mod-

[1] Even for molecules of modest size, the size of the conformation search space quickly exceeds the size of the universe

els [15, 17], and general full-atom models [7, 19].

In some cases (e.g. [15, 29]), 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. [10] in which individuals are always replaced by their locally optimized structures. Unger and Moult [30] propose a hybrid, similar to the latter, in which each individual undergoes 20 steps of simulated annealing before selection is performed. Simulated annealing has also been applied to a variety of protein energy models [22].

We have proposed [19] hybrid genetic algorithm variations which incorporate efficient gradient based minimization directly in the fitness evaluation, which is based on a general full-atom potential energy model. 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 [32], or more generally probabilistically Lamarckian ($0 \leq p_r \leq 1$) evolution. This approach has resulted in energy values smaller than those found in the current literature, although the associated polypeptide conformations are somewhat different than those achieved by other researchers due to symmetry such as the symmetric positioning of the end residues.

Here we introduce the **REGAL** (REal-valued Genetic Algorithm, Limited by constraints) approach to polypeptide structure prediction. Using the *Evolution Program* concept of Michalewicz [20], a genetic algorithm is transformed into a *stronger* algorithm by incorporating "natural" data structures (usually real-valued) that capture problem specific domain knowledge, thus limiting the algorithm to a specific problem, but enhancing its effectiveness. Such an approach is consistent with Kauffman's NK model [13] in which the use of real-valued alleles tends to provide for easier GA population movements towards global optimum vs binary-encoded alleles. General comparisons of real-valued and binary GAs have been discussed elsewhere[6, 20, 33]

We describe experiments comparing the effectiveness of the real-valued genetic REGAL algorithm to that of our previously developed hybrid GAs [19, 18]. We test this approach on two molecular structures (Section 3). Conclusions are presented in Section 4, and Section 5 discusses directions for future research. The following section presents the methodology.

2 Methodology

In this section we discuss the objective function associated with our polypeptide energy minimization application (Section 2.1) as well as binary and real-valued encoding scheme Section 2.2). We discuss the implementation of the real-valued GA in Section 2.3 and briefly review our minimization technique in Section 2.4.

2.1 Objective Function

The conformational state of a molecule, defined by the physical locations of the component atoms, is represented by the *internal coordinate system*. The location of atom, i is defined in terms of *bond length* (i, j), *bond angle* (i, j, k), and *dihedral angle* (i, j, k, l) with respect to neighboring atoms. Bond lengths and angles are assumed fixed leaving the dihedral angles $(\phi_i, \psi_i, \omega_i, \chi_{i_n})$ as independent variables. See Figure 1

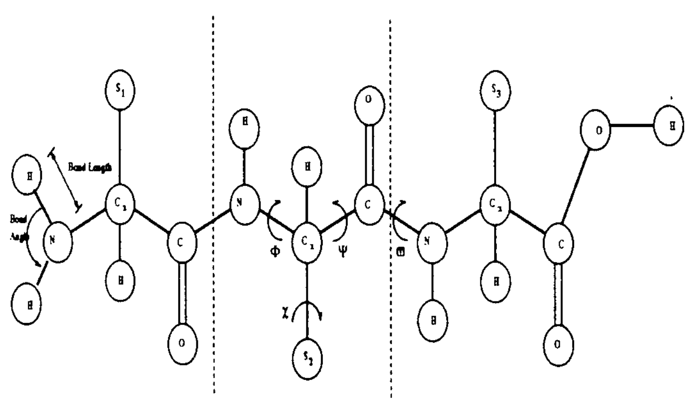

Figure 1: Tri-peptide showing internal coordinate system references

The objective function, which we seek to minimize, is based on the CHARMM [2] energy function

$$E = \sum_{(i,j)\in\mathcal{B}} K_{r_{ij}}(r_{ij} - r_{eq})^2 + \sum_{(i,j,k)\in\mathcal{A}} K_{\Theta_{ijk}}(\Theta_{ijk} - \Theta_{eq})^2 + \sum_{(i,j,k,l)\in\mathcal{D}} K_{\Phi_{ijkl}}[1 + \cos(n_{ijkl}\Phi_{ijkl} - \gamma_{ijkl})] +$$

$$\sum_{(i,j)\in\mathcal{N}}\left[\left(\frac{A_{ij}}{r_{ij}}\right)^{12}-\left(\frac{B_{ij}}{r_{ij}}\right)^{6}+\frac{q_iq_j}{4\pi\varepsilon r_{ij}}\right]+$$
$$\frac{1}{2}\sum_{(i,j)\in\mathcal{N}'}\left[\left(\frac{A_{ij}}{r_{ij}}\right)^{12}-\left(\frac{B_{ij}}{r_{ij}}\right)^{6}+\frac{q_iq_j}{4\pi\varepsilon r_{ij}}\right] \tag{1}$$

where the five terms (which we denote $E_{\mathcal{B}}$, $E_{\mathcal{A}}$, $E_{\mathcal{D}}$, $E_{\mathcal{N}}$, $E_{\mathcal{N}'}$) represent the energy due to bond stretching, bond angle deformation, dihedral angle deformation, non-bonded interactions, and 1-4 interactions, respectively. Specifically,

- $\mathcal{B}$ is the set of bonded atom pairs,
- $\mathcal{A}$ is the set of atom triples defining bond angles,
- $\mathcal{D}$ 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,
- Θ_{ijk} is the angle formed by atoms i, j, and k,
- Φ_{ijkl} is the dihedral angle formed by atoms i, j, k, and l,
- q_i is the partial atomic charges of atom i,
- the $K_{r_{ij}}$'s, r_{eq}'s, $K_{\Theta_{ijk}}$'s, Θ_{eq}'s, $K_{\Phi_{ijkl}}$'s, γ_{ijkl}'s, A_{ij}'s, B_{ij}'s, and ε are empirically determined constants (taken from the QUANTA parameter files).

The CHARMM energy model seems to be the most complex of those available for polypeptide prediction including AMBER and ECEPP/2 which we have employed elsewhere [18].

The primary determinants of a protein's 3-D structure, and thus the energetics of the system, are its independent dihedral angles [31]. Our genetic algorithm operates on individuals which encode these dihedral angles [7]. In Equation 1, E is expressed as a function of both the internal coordinates (bond lengths r_{ij} for $(i,j)\in\mathcal{B}$, bond angles Θ_{ijk}, and dihedral angles Φ_{ijkl}) and the inter-atomic distances r_{ij} for $(i,j)\in\mathcal{N}\cup\mathcal{N}'$. Thus, in order to calculate E (and hence the fitness) for the conformation encoded by an individual, it is necessary to calculate its Cartesian coordinates from its internal coordinates.

2.2 Encoding Scheme

Two specific biomolecules are investigated based on variety of reasons. The first is the pentapeptide [Met]-enkephalin[2]. This molecule is chosen because it has been used as a test problem for many other energy minimization investigations (e.g. [15, 23]), and its minimum energy conformation is known (with respect to the ECEPP/2 energy model)[16]. A minimum energy for [Met]-enkephalin with respect to CHARMM was published[22] but there is a question about this minimum since the ECEPP optimal was used as the starting conformation. The second is a 14 residue model of Polyalanine. That is, it is a homogeneous molecule made up of 14 residues of the amino acid alanine. An amino acid becomes a residue when a water H_2O molecule is freed during the formation of the peptide bond. Its native structure is an α-helix.

[Met]-enkephalin has 75 atoms. In it, 24 dihedral angles are treated as independent, the rest are either fixed, or treated as dependent. Polyalanine has 143 atoms. In it, 56 dihedral angles are treated as independent.

In the binary representation, 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

$$D(a_1,a_2,\ldots,a_{10})=-\pi+2\pi\sum_{j=1}^{10}a_j2^{-j}. \tag{2}$$

This encoding yields a precision of approximately one third of one degree. Recall that 24 dihedral angles determine [Met]-enkephalin's structure, hence its string length is 240. Likewise, 56 dihedral angles determine Polyalanine's structure, hence its string length is 560.

In the real-valued representation each individual is vector of real variables,

$$\vec{x}=(x_1,\ldots,x_n)\in\boldsymbol{R}^n. \tag{3}$$

for independent dihedral angles $-\pi\le x_i\le\pi$.

2.3 Real-valued GA Design

Our real-valued GA for PSP,[3] REGAL, integrates our previously developed molecular and energy models with the Genocop-III algorithm developed by Michalewicz and Nazhiyath[21]. Genocop-III is a co-evolutionary algorithm implementation for numerical optimization. It deals with the problem of infeasible candidate solutions in constrained problems by repairing, rather than penalizing. This is done by maintaining two populations, a *Search* population, P_s, whose members are feasible for linear constraints, and a *Reference* population, P_r, whose members are feasible for all constraints. A unique domain can be defined for each variable, else it defaults to $\boldsymbol{R}$. Also, any number of linear inequalities, nonlinear

[2]the five amino acids in [Met]-enkephalin are in order tyrosine, glycine, glycine, phenylalanine, methionine [27]

[3]Polypeptide Structure Prediction

equalities, nonlinear inequalities may be defined. Figure 2 reflects the general GA structure of the Genocop III stochastic search algorithm. Figure 3 presents the critical population *evaluate* operator which includes a repair function.

```
procedure Genocop III
begin
   t ← 0, "t is number of generations"
   initialize P_s(t)
   initialize P_r(t)
   evaluate P_s(t)
   evaluate P_r(t)
   while (not termination-condition) do
   begin
      t ← t + 1
      select P_s(t) from P_s(t − 1)
      alter P_s(t)
      evaluate P_s(t)
      if t mod k = 0 then
      begin
         alter P_r(t)
         select P_r(t) from P_r(t − 1)
         evaluate P_r(t)
      end
   end
end
```

Figure 2: The structure of Genocop III

```
procedure evaluate P_s(t)
begin
  for each s ∈ P_s(t) do
   if s ∈ F "feasibility set"
   then evaluate s (as f(s)) else
   begin
    select r ∈ P_r(t)
    generate z ∈ F
    evaluate s (as f(z))
    if f(r) > f(z) then replace r by z in P_r
    replace s by z in P_s with probability p_r
   end
end
```

Figure 3: Evaluation of population P_s in Genocop III

2.3.1 Domain Knowledge

While no general algorithmic solutions to the *protein folding problem* exist today in spite of more than 30 years effort, a considerable body of knowledge has been amassed. A few examples follow:

- ω angles assume either a native state *cis* or *trans*[4] orientation; i.e., a unique isomerization conformation for each residue [24]
- χ_1 angles are usually -60, 60, 180 degrees $\pm$ some deviation. One can also use data from *rotamer* libraries; i.e., libraries of known side-chain structures
- Certain values for ϕ and ψ angle pair are frequently or rarely observed. These constraints can be visualized with a *Ramachandran* plot [27]

Assuming bond lengths and bond angles are held constant, the search space for the fixed geometry model is $[-pi, pi]^n$ where n in the number of independent dihedral (or torsional) angles. Knowledge about the problem space can be used to constrain this search space. Most constraints can be expressed as nonlinear inequalities in one of the following generalized forms as developed by Kaiser[12]:

$$0 \le \cos(\theta - \frac{\theta_{min} + \theta_{max}}{2}) - \cos(\frac{\theta_{min} - \theta_{max}}{2}) \quad (4)$$

are the constraints for the $\{\phi, \psi, \omega\}$ angles, and

$$0 \le \cos(3\theta - \frac{\theta_{min} + \theta_{max}}{2}) - \cos(\frac{\theta_{min} - \theta_{max}}{2}) \quad (5)$$

are the constraints for the $\{\chi_1\}$ angles.

2.3.2 Constraint Sets

Our focus here is on development of techniques usable by biochemistry researcher for structure prediction. The constraint sets developed in this research demonstrate the feasibility of the approach. In general they are conservation, that is, less restrictive. A biochemistry researcher studying a particular molecule may choose to develop more aggressive constraints. This flexibility is inherent in our design.

The "loose" constraints for [Met]-enkephalin (Table 1) were developed by examining Ramachandran plots of observed values of *phi* and *psi* angle for the residues alanine and glycine [4]. Of the twenty amino acids, proline and glycine have unique $\phi\psi$ distributions. The other residues are similar to alanine. The "tight" constraints (Table 2) consider the above data and infer additional insights from "homologous" molecules.

Values for the Polyalanine constraints (Table 3) were developed in a similar way. It was known *a priori* that this molecule forms an α-helix secondary structure. Thus a plot from Stryer's text [27] that specifies the $\phi\psi$ region for an α-helix was used. A similar process to that above was used for the "tight" constraints (Table 4). After consulting with biochemistry domain

[4] $cis \equiv 0$, $trans \equiv \pm\pi$ *or* $\pm 180°$

Table 1: Loose constraints for [Met]-enkephalin

Dihedral	Midpoint	Radius
$\Phi_{Non-glycine}$	-120	90
$\Phi_{Glycine}$	180	135
Ψ	60	150
Ω	180	20
χ_1	$-60 \mid 60 \mid 180$	30

Table 2: Tight constraints for [Met]-enkephalin

Dihedral	Midpoint	Radius
$\Phi_{Non-glycine}$	-120	60
$\Phi_{Glycine}$	130	70
Ψ	150	140
Ω	180	12.5
χ_1	$-60 \mid 60 \mid 180$	7.5

experts, a third set of constraints "tight, relaxed terminals" were defined. These are based on the knowledge that the dihedral angles for the terminals residues will not be consistent with the non-terminal angles even in a very regular secondary structure like an α-helix. They are the same as "tight" without constraints on residues 1 and 14.

Table 3: Tight constraints for Polyalanine

Dihedral	Midpoint	Radius
Φ	-67.5	22.5
Ψ	-30	30
Ω	180	20
χ_1	$-60 \mid 60 \mid 180$	30

2.3.3 Input Parameter File

Genocop-III introduces only a few new variables. However, there are now two populations and ten operators to control. Thus, there many times more permutations to the parameter mix. While initial results are encouraging, we plan additional study in this area to develop "better" parameter values[12].

2.4 Hybrid Binary GA Design

The CHARMM objective function defined by Equation 1 is such that all of its second partial derivatives exist and are continuous almost everywhere. That is, for each derivative the set of discontinuities is finite in this case. We have considered [19] three local minimization techniques which exploit to varying degrees this smoothness property along with the ready availability of software. The three deterministic local search approaches

Table 4: Tight constraints for Polyalanine

Dihedral	Midpoint	Radius
Φ	-60	15
Ψ	-45	15
Ω	180	5
χ_1	$-60 \mid 60 \mid 180$	5

considered were the first derivative method, the critical point method and the exact second derivative method.

We selected a readily available implementation of the first derivative method known as the conjugate gradient technique [25]. This method was chosen since it is less computational expensive than the others for complete minimization execution per individual. Yet, it retains the minimization benefits for the hybrid GA approach. Here we modify the bracketing procedure used in the line minimizations. The standard bracketing procedure assumes that the domain of each of the independent variables is the set of all real numbers, whereas our independent variables assume values only in the interval $[-\pi, \pi]$. Consequently, the intervals produced by the standard procedure typically are not limited to the basin of attraction in which the encoded conformation lies. Our method heuristically corrects this problem by choosing an interval over which no dihedral angle varies by more than $\frac{\pi}{6}$. Neglecting non-bonded interactions, this guarantees that the bracketed interval is contained in the conformation's basin of attraction, and that it contains the local minimum along the direction of minimization. Details of the CHARMM analytical gradient derivation and our associated hybrid GA are presented in [19].

The local minimization discussed above is implemented in the hybrid GA such that there are runtime parameters for *probability of minimization* and *probability of replacement*. With both set to 0.0, the hybrid performs as a simple GA (SGA). If *probability of minimization* = 1.0 $\wedge$ *probability of replacementt* = 1.0, then Lamarckian minimization is performed. If *probability of minimization* = 1.0 $\wedge$ *probability of replacement* = 0.0, then Baldwinian minimization is performed. Results for all three techniques are presented in the next section.

3 Results and Comparison

In this section we present the results of experiments on two separate molecular models, [Met]-enkephalin and a 14 residue model of Polyalanine.

3.1 [Met]-enkephalin

In general the hybrid GA is no more effective than the REGAL approach in minimizing [Met]-enkephalin, ex-

Table 5: Final minimum energies (kcal/mol) for [Met]-enkephalin using binary GA with FP selection

Algorithm	Mean	SD	RMSD Best
SGA	-22.58	1.57	4.51
Baldwinian	-22.57	1.62	3.96
Lamarckian	-28.35	1.29	3.33

Table 6: Final minimum energies (kcal/mol) for [Met]-enkephalin using REGAL

Algorithm	Mean	SD	RMSD Best
No constraints	-24.92	2.99	4.57
No constraints w/local min	-26.38	2.69	4.40
Loose constraints	-22.01	2.69	4.25
Loose constraints w/local min	-24.95	4.23	4.26
Tight constraints	-23.55	1.69	3.23
Tight constraints w/local min	-17.71	0.50	5.05

cept when using Lamarckian minimization with a best average of -28.35 kcal/mol (Table 5). However, the best over value, -30.32 kcal/mol, was a REGAL technique, no constraints with Lamarckian minimization. This single example demonstrates a potential for local minimization incorporated with REGAL. But in general, tighter constrains appear to interfere with local minimization. That is, a local minima is found during the initial evaluation from which the experiment is unable to escape. We suspect that as the ratio of feasible space $\mathcal{F}$ to search space $\mathcal{S}$ gets smaller, the operators are unable to generate a more fit "feasible" candidate and escape the local minima.

It is interesting to note that conformers have been identified with values less than the accepted optimal conformation (CHARMM equivalent of the ECEPP/2 conformation of Li and Scheraga [16]). We have suspected the optimal conformation for ECEPP/2 and CHARMM are different–this was confirmed during the 1996 American Chemical Society National Meeting.

3.2 Polyalanine

The effectiveness of the Binary GA (even with minimization) did not hold for the larger molecule Polyalanine (Table 7). Significant improvement were observed when the *step* size in the conjugate gradient minimization was properly sized for for the larger molecule (another example of using "domain knowledge"). When examined visually, these conformations did not appear to be forming the expected α-helix secondary structures.

With adequate domain knowledge, in the form of tight constraints, REGAL performs well on the larger

Table 7: Final minimum energies (kcal/mol) for Polyalanine using binary GA with FP selection

Algorithm	Mean	SD	RMSD Best
SGA	-93.25	10.85	9.67
Baldwinian	-103.73	16.5	7.36
Lamarckian	-140.60	5.39	12.74
Lamarckian corrected	-308.51	8.26	5.03

Table 8: Final minimum energies (kcal/mol) for Polyalanine using REGAL

Algorithm	Mean	SD	RMSD Best
No constraints	-273.08	13.81	6.25
Loose constraints	-336.65	4.50	1.87
Loose constraints w/local min	-309.00	8.19	2.70
Tight constraints	-337.64	4.40	0.98
Tight constraints w/local min	-316.47	0.0^5	1.17
Tight constraints w/ relaxed terminals	-338.30	4.24	1.42
Tight, relaxed 150K evals	-351.76	0.57	1.40

molecule (Table 8). When allowed to reach 150,000 evaluations, the energy value is almost that of the optimal conformation with relaxation of bond lengths and bond angles. When examined visually, these conformations definitely formed the expected α-helix secondary structures.

Again, local minimization was not effective when used in conjunction with constraints. This time, the difference between the results is more substantial.

3.3 Efficiency

The experiments in this paper were conducted on a variety of platforms. They include 368 node Paragon supercomputer, 100 and 200 mhz Silicon Graphics workstation, SUN Sparc workstations (2, 5, and 20), and SUN Ultra Sparc workstations. The bulk of the effort was accomplished in a common user lab of 46 networked Sparc20 workstations. As is to be expected, run times (wall clock) varied with system loading. However, a few general observations can be made:

- Met-enkephalin
  - Lamarckian Binary GA$\approx$ 13.3 hours
  - REGAL $\approx$ 2 hours
- Polyalanine
  - Lamarckian Binary GA $\approx$ 120 hours

- REGAL $\approx$ 4 hours

While results prove nothing, initial data suggest the REGAL approach scales better than the binary GA with local minimization. While the above times might seem excessive, it takes years to identify protein conformations using experimental methods such as x-ray crystallography.

4 Conclusions

The binary-valued Lamarckian GA algorithms obtained better energies than the simple GA and Baldwinian approach for the minimization of the CHARMM potential for [Met]-Enkephalin and Polyalanine using fitness proportionate selection. The effectiveness of the Lamarckian GA suggests that the low-energy local minima in the energy landscape of [Met]-Enkephalin may occur somewhat regularly within the conformation space. If this is the case for [Met]-Enkephalin, this approach may hold for larger polypeptides as well, however, this was not the case for the larger dimension Polyalanine. The REGAL approach achieved considerably better results. Thus, the use of our real-valued REGAL method may be appropiate for higher dimensional polypeptides since good minimum energy values for both [Met]-enkkephalin and Polyalanine were obtained. Moreover, the local minima for complex high-dimensional proteins may not appear regularly in the energy landscape indicating more difficult computations for more complex proteins. Note that the associated conformations of the two proteins reflected the general structural results of other researchers as indicated by reference.

Of course, determination of appropriate linear and non-linear constraints associated with polypeptide structure is critical to achieving low-energy conformations as shown in our REGAL experiments. In addition, replacement frequencies must be appropriate to the level of selective pressure in order to insure the presence of enough locally optimal individuals to prevent premature convergence. Thus, in the application of GAs to the specific protein folding problem (polypeptide structure prediction), the quest of at least some general GAs for solving a class of proteins continues. The ongoing results of our efforts tend to indicate that a REGAL approach may solve some restricted class of protein folding problems.

5 Future Directions

The results and conclusions of this effort indicated that real-valued GAs for solving the polypeptide structure problem have excellent potential. Also, the appropriate use of linear and nonliner constraints has considerable impact on population evolution and deserves to be furter investigated. Moreover, the appropriate use of real-valued GA operators has a very large impact on population and also deserves follow-on investigations. Comparing experimental energy data using statistical analysis is still to be accomplished.

The success of using binary encoded and real-valued GAs for the two proteins suggests their application to more complex protein folding problems. In applying GAs to more and more complex proteins, the use of constraints may be the only way of obtaining acceptable solutions due to the exponentially increasing number of local and golbal optimal. Such applications require additional computational platforms as found in highly scalable architertures. We have previously[7] used our own GA island and farming algorithms in solving the polypeptide structure problem for the two polypeptides and are now mapping the real-valued GA software to such platforms for structure prediction of complex polypeptides.

References

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

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

[3] H. Chan and K. Dill. The protein folding problem. *Physics Today*, pages 24–32, February 1993.

[4] Thomas E. Creighton, editor. *Protein Folding.* W. H. Freeman, New York, 1992.

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

[6] Lawrence Davis, editor. *Handbook of Genetic Algorithms.* Pitman, London, 1987.

[7] 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.

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

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

[10] 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.

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

[12] Charles E. Kaiser, Jr. Refined genetic algorithms for polypeptide structure prediction. Ms thesis, afit/gce/eng/96d-13, Graduate School of Engineering, Air Force Institute of Technology (AETC), Wright-Patterson AFB OH, December 96. In Process.

[13] Stuart A. Kauffman. *The Origins of Order, Self-Organization and Selection in Evolution.* Oxford Unbiversity Press, New York, 93.

[14] Jr. Kenneth M. Merz and Scot M. LeGrand. *The Protein Folding Problem and Tertiary Structure Prediction.* Birhäuser, Boston, 1994.

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

[16] 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.

[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, George H. Gates, Jr., Gary B. Lamont, and Ruth Pachter. Hybrid genetic algorithms for minimization of a polypeptide specific energy model. In *Proteedings of the Third IEEE Conference on Evolutionary Computation*, 1996.

[19] 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.

[20] Zbigniew Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs.* Springer-Verlag, New York, 2nd edition, 1994.

[21] Zbigniew Michalewicz and Girish Nazhiyath. Genocop III: a co-evolutionary algorithm for numerical optimization. volume 2, pages 647–651, 1995.

[22] Thierry Montcalm, Weili Cui, Hong Zhao, Frank Guarnieri, and Stephen R. Wilson. Simulated annealing of met-enkephalin: low-energy states and their relevance to membrane-bound, solution and solid-state conformations. *Molecular Structure (Theochem)*, 308:37–51, 1994.

[23] 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.

[24] Roger H. Pain, editor. *Mechanisms of Protein Folding.* IRL Press, New York, 1994.

[25] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, New York, second edition, 1992.

[26] Steffen Schulze-Kremer. Genetic algorithms for protein tertiary structure prediction. pages 129–149.

[27] Lubert Stryer. *Biochemistry.* W.H. Freeman, New York, 4rd edition, 1995.

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

[29] 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.

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

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

[32] Darrell Whitley, V. Scott Gordon, and Keith Mathias. Lamarckian evolution, the Baldwin effect and function optimization. In *PPSN III*, pages 6–15.

[33] Alden H. Wright. Genetic algorithms for real parameter optimization. In Gregory J. E. Rawlins, editor, *Foundation of Genetic Algotithms*, pages 205–218, San Mateo, CA, 1991. Morgan Kaufmann.