1. APPENDIX: TEST DATA

The following data were taken directly from Unger & Moult [93a] as published in Proceedings of the Fifth Annual International Conference on Genetic Algorithms.

Table 4: 27 Peptide Length Test Cases

Case#	Sequence
273d.1	wbwbwbbbwwbwbw wwwwwwwwwwbbw
273d.2	wbbwwwwwwwwwwb bwwbbwwbwwbwb
273d.3	bbbbwwwwwbwwww wbbbwwwwwwwwb
273d.4	bbbwwbbbbwwwbw bwwbbwwbwwwbb
273d.5	bbbbwwwwbwbbww wbbwwwwwwwwww
273d.6	bwwwwwwbwbbbww bbwwwbwwwwbwb
273d.7	bwwbwbbwwwbwww wwbwbbwbwbwbb
273d.8	bwwwwwwwwwwwbw bwwwwwwwwbwbb
273d.9	wwwwwwwbbbwwwb wbbwwwbwwbwww
273d.10	wwwwwbbwbwbwbw bwwbbwbbwbwww

Table 5: 64 Peptide Length Test Cases

Case#	Sequence
643d.1	wwbbbbbwwwbbwwwwwbbwwwbwwwwwwbwb wwwbwwbwwbwwwwwbwwwwbbwbbwwbwwbw
643d.2	wwbwbwwbwwbbbwbbbbwwbbbwwwwbwbww wbwbwwwbwbwwwwwbwbwwbwbwwwbwwbww
643d.3	bwbbwwbbwbwwwwwbbbwbbbbwwbwwbwbb wwwbwbwwbbbwbbwbwwwwwbbbbbbbbwww
643d.4	bwwbbwwbwwbwbwwbwwwwbwwwwwwbwbwb bbwwbwbwwwbwbwwbbwwbwwbwwbwbbbwb
643d.5	bwwwbbwwbwbwwwbwwwbwbbwwwbbwbwbb wbwwbwwwbwwbwbbbwwbwwbwwbbbwbbbb
643d.6	bwwbbwbbbbwwwwwwbbwwbwwwwbbwwwbw wbwbbwbwwwwbbwwwwbwwwwwbwwwwbwbb
643d.7	wwwwbwwwbwwwbbbbwbbwwwwwbwwbwbbw bwbwwwwwbwwwwwwwwwwbbbbwwwwbbwwb
643d.8	wwwbbbwwbwbwwbwwbbwwwbwwbwwbbwbw wwbwwwwwwwbwbbbwbbbbbwwbbwwwbwwb
643d.9	bwwbwwbbbwwwwbwbwwwbwbbwbbbbbwww wbwbwbwwwwbwbwwwbbwbwwwwbwwbbwbw
643d.10	wwbwwbwwbbbwwwbwbwwbwwbwwwwwwbww bbbwwbwwbwwbwbwwwwwwbbbwwwwwbwbw

Recently, Dandekar and Argos [Dandekar & Argos 94] have shown relatively strong results using a standard GA approach and a less restrictive model than that used here. It may be of interest to study the parallels in the operation of the simplified 3-D lattice model and a more complex model built on the same concepts in order to understand how the additional degrees of freedom and complexity affect translation of the GA methods from one model to the other.

7. FURTHER INFORMATION

For those interested in other work done by the MSU GARAGe, please visit our web pages located at `http//:isl.cps.msu.edu/GA`. Both the GALLOPs code and related papers, as well as future follow-ups to this work can be found there.

8. ACKNOWLEDGMENTS

We would like to acknowledge the work of Unger and Moult upon which this research has been based.

9. REFERENCES

[1] Benner, Steven A., “Patterns of Divergence in Homologous Proteins as Indicators of Tertiary and Quartenary Structure”, Advances in Enzyme Regulation, Vol. 28, pp. 219-236, 1988.

[2] Benner, Steven A., Dietlind L. Gerloff, and Thomas F. Jenny, “Predicting Protein Crystal Structures”, Science, Vol. 265, pp. 1642-1644, 1994.

[3] Camacho, Carlos J. and D. Thirumalai, “Kinetics and Thermodynamics of Folding in Model Proteins”, Proceedings of National Academy of Sciences of the United States, Vol. 93, No. 13, pp 6369-63672, 1993.

[4] Dandekar, T. and P. Argos, ”Folding the Main Chain of Small Proteins with the Genetic Algorithm”, Journal of Molecular Biology, V. 236, pp. 844-861, 1994.

[5] Fasman, G. D. (ed.), Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Press, 1989.

[6] Fraenkel, Aviezri S., “Complexity of Protein Folding”, Bulletin of Mathematical Biology, pp. 1199-1210, 1993.

[7] Goldberg, D.E., Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wessley Publishing Company, Inc., 1989.

[8] Goodman, E., “An Introduction to GALOPPS--The ‘Genetic ALgorithm Optimized for Portability and Parallelism’ System”, Tech. Report, Michigan State University, 1994.

[9] Holland, John H., Adaptation in Natural and Artificial Systems, The University of Michigan Press, Ann Arbor, 1975.

[10] Mertz, K.M., and S.M. Le Grand, The Protein Folding Problem and Tertiary Structure Prediction, Birkhpauser, Boston, 1994.

[11] Robson, Barry, Jean Garnier, Geoffrey J. Barton, and Robert J. Russell, “Protein Structure Prediction”, Nature, Vol. 316, No. 6412, pp. 505-506, 1992.

[12] Robson, Barry and Jean Garnier, Introduction to Proteins and Protein Engineering, Elsevier, 1988.

[13] Rooman, Marianne J. and Shoshana J. Wodak, “Extracting Information on Folding from the Amino Acid Sequence: Consensus Regions with Preferred Conformation in Homologous Proteins”, Biochemistry, Vol. 31, No. 42., pp. 10239-10249, 1992.

[14] Skolnick, Jeffery and Andrzej Kolinski, “Simulations of the Folding of a Globular Protein.”, Science, Vol. 250, pp. 1621-1625, 1990.

[15] Unger, Ron and John Moult, “A Genetic Algorithm for 3D Protein Folding Simulations”, Proceedings of the Fifth Annual International Conference on Genetic Algorithms, pp. 581-588, 1993a.

[16] Unger, Ron and John Moult, “Finding the Lowest Free Energy Conformation of a Protein is an NP-Hard Problem: Proof and Implications”, Bulletin of Mathematical Biology, pp. 1183-1198, 1993b.

[17] Unger, Ron and John Moult, “Genetic Algorithms for Protein Folding Simulations”, Journal of Molecular Biology, Vol. 231, pp. 75-81, 1993c.

[18] Voet, Doland and Judith G. Voet, Biochemistry, John Wiley & Sons, New York, 1990.

[19] York, D.M., T.A. Darden, L.G. Pedersen, and M.W. Anderson, “Molecular Dynamics Simulation of HIV-1 Protease in a Crystalline Environment and in Solution”, Biochemistry, Vol. 32., No. 6, pp. 1143-1153, 1993.

Table 2: Hydrophobic Moment Ratio Results

Test Case	Moment Ratio	Test Case	Moment Ratio
273d.1	1.00	643d.1	1.20
273d.2	1.11	643d.2	1.20
273d.3	0.89	643d.3	1.15
273d.4	0.93	643d.4	1.30
273d.5	0.89	643d.5	1.20
273d.6	1.10	643d.6	1.29
273d.7	1.08	643d.7	1.13
273d.8	0.67	643d.8	1.26
273d.9	0.88	643d.9	1.22
273d.10	1.00	643d.10	1.18

Best of run results for standard GA test runs on original Unger & Moult test cases.

limit for the sequence.

Observing the ratio measurements (Table 2.), we see that most of the 27 length peptide test cases reached an average of two contacts per peptide while those of the longer 64 length peptides averaged 2.5 contacts per hydrophobe. Both data sets maintain about a one third density of hydrophobes to hydrophiles, so it appears that the length is the primary determining factor in limiting this ratio. Indeed we would expect this value to approach the theoretical limit as the sequence length (assuming a fairly even hydrophobe distribution) approaches infinity. An interesting experiment would be to observe the variance of the ratio with increasing string length and hydrophobe density.

Many of these results, especially for the 64 length sequence tests, were run for only 200-500 generations. We may potentially be able to reach lower energy configurations by running a series of test runs until the population converges. Analysis of multiple equal valued results for separate test runs show varying structure with similar local motifs (understandable in that GAs favor short defining length schema).

5. CONCLUSIONS

Our central conclusion is that a standard GA implementation can outperform a hybrid GA/Monte Carlo approach such as that offered by Unger & Moult both in terms of efficiency and in escaping local optima. In general we may conclude that biasing the search space of a GA may handicap rather than enhance its search capabilities, regardless of the positive intent of the bias.

Allowing the search to progress through illegal states appears to have the potential to increase the efficiency of the search; however, further experiments are necessary before such conclusions can be reached. Accessing a proper penalty for such states is often marked as the problem with allowing illegal phenotypes to survive. Indeed, such problems initially plagued our implementation as well - leading us to develop the preference ordering encoding in an effort to partially eliminate such difficulties. While the preference ordering is somewhat expensive, it allows the GA to search at an initially faster pace than a direct encoding as it avoids the initial “untangling” phase which appears when a direct encoding is used. The threshold under which the additional cost of the bump check for the preference encoding pays off is not known; however, such an approach may be useful in other problem spaces which contain illegal states. The preference encoding as it stands (as a full permutation) clearly will not scale into larger degrees of freedom; however, the more general concept of building in alternate or redundant secondary encodings to avoid potential illegal states should scale and presents much potential for further investigation. Indeed, the concept somewhat parallels the encoding portion of Goldberg’s messy GA, but with a fixed rather than variable length encoding. The lesson which we learned while adjusting the penalty function for illegal conformations can be summed up in the following axiom, which we offer as general advice in the fashioning of GA encodings and evaluation functions:

Tolerate but do not reward incorrect behavior.

Many of our difficulties in handling illegal states were solved by refusing to include the positive contributions of residues which were “causing” collisions.

6. FUTURE WORK

After laying a slightly more solid theoretical foundation for this work, the next step seems to be to enhance the model to more closely resemble actual 3-D protein forms. The most likely form of validation for such a model would be the spontaneous discovery of familiar secondary structures (alpha-helices, beta-sheets, etc.); however, since hydrogen bonds play a clear role in the formation of such structures, it is not clear that they will appear unless such interactions are also added to the model.

Several researchers have proposed approaches to the protein folding problem using genetic algorithms with a variety of models. [Mertz & Le Grand 94] offer an excellent up-to-date review of work in the area.

	Unger & Moult GA		Our Approach		Difference	
Seq.	#Energy eval.	Lowest Energy	Required evals.	Lowest Energy	Speedup	Relative % Evals Used
273d.1	1,227,964	-9	27,786	-9	44.19	2.3
273d.2	1,225,281	-9	81,900	-10	14.96	7.7
273d.3	1,247,208	-8	16,757	-8	74.43	1.3
273d.4	1,207,686	-15	85,447	-15	14.13	7.1
273d.5	1,118,202	-8	8,524	-8	131.18	0.8
273d.6	1,226,090	-11	44,053	-11	27.83	3.6
273d.7	1,239,519	-12	85,424	-13	14.51	6.9
273d.8	1,248,118	-4	3,603	-4	3464.09	0.03
273d.9	1,198,945	-7	10,610	-7	113.00	0.9
273d.10	1,174,297	-11	16,282	-11	72.12	1.4
Total	11,113,310		380,386		29.22	3.4
643d.1	2,119,775	-27	433,533	-27	4.88	20.5
643d.2	2,286,289	-29	167,017	-30	13.69	7.3
643d.3	1,831,102	-35	172,192	-38	17.09	5.9
643d.4	2,315,112	-34	107,143	-34	21.61	4.6
643d.5	2,040,915	-32	154,168	-36	13.23	7.6
643d.6	2,160,690	-29	454,727	-31	4.75	21.0
643d.7	2,317,862	-20	320,396	-25	7.23	13.8
643d.8	2,391,876	-29	315,036	-34	7.59	13.2
643d.9	2,121,287	-32	151,705	-33	14.58	6.9
643d.10	2,287,394	-24	191,019	-26	17.55	5.7
Total	21,872,302		2,466,936		8.86	11.3

Table 1: Result Comparison to Unger & Moult [93a]

Energy values are reported using the Unger & Moult energy function to allow direct comparison to previously published results. Results for Unger & Moult's algorithm are taken directly from [93a]. Speedup is the ratio of the number of evaluations and assumes function evaluation times for the two implementations are comparable. Reported wall clock time for the Unger & Moult 64 length sequences were around 2 hours on a SGI 4d35, wall clock times for our algorithm were under 20 minutes for 27 length sequences and under 40 minutes for 64 length sequences on a Sparc 20. Required evals. reported are the evaluation number when reported best energy value was reached. Our results reflect single representative runs only; however, all additional runs produced similar results within roughly equivalent time frames (roughly +/- 8%). All of our runs used the preference ordering encoding; however, the relative move encoding produced identical energy values, but required more evaluations (approx. +20%).

forces are the primary driving force in achieving native protein conformations. Therefore, we suggest a measure of the hydrophobic moment as a generalized measure of the potential of a conformation using this 3-D lattice model. Specifically, we suggest the ratio of non-sequential hydrophobe-hydrophobe contacts to the number of hydrophobes in the sequence as a valid measure, since its range is invariant to changes in the length of the sequence as well as the actual number of hydrophobes in the sequence.

Since each hydrophobe (discounting end points) can have at most 4 other contacts (six degrees of freedom minus 2 discounted peptides directly connected on the protein chain), and each shared contact counts effectively as +0.5 for the two hydrophobes involved, the maximum possible ratio is effectively 2.0 for reasonably long sequences; however, this is an asymptotic limit since hydrophobes on the edge of the "hydrophobic core" will have less than the ideal 4 contacts. Thus, the limit may only be reached by an infinite length protein string with hydrophobes spaced (2+2n, n $\geq$ 0) residues apart. The maximum potential value for a given sequence will be a fixed value ranging from 0 to 2.0.

3.2 GENERALIZED TEST SEQUENCES

For purposes of tuning the GA, we have created a series of test sequences for which the minimum configuration may be directly calculated. This series of "looped cubes" also has allowed us to test the effectiveness of the approach on several different types of possibly difficult sequences.

A **minimal looped cube** (**Fig. 1**) consists of 8 hydrophobic residues evenly interspersed with two hydrophilic residues each. In the minimal configuration, the hydrophilic residues form a cube with the vertices connected by hydrophilic "loops". The Unger-Moult energy value for this configuration is 12, which gives a hydrophobic moment ratio of 1.5.

The minimal looped cube may be extended simply by adding long hydrophilic tails to the ends or by increasing the interhydrophobic distance on the primary sequence to **2 n**, where **n** is any positive integer to create an **n-loop cube** (**Fig 2.**). Thus, we are able to test the effects of sequence length (long tails) versus schema defining length (longer loops). As expected, our results showed that defining length, rather than sequence length was the key factor in the speed of the GA; however, in all test cases the GA was able to correctly find a minimal configuration. Likewise, several other test cases with known minimal values were used to test and tune the GA.

4. RESULTS

After finding the best operators and parameters and fine-tuning the scoring algorithm, we were able to obtain results which compare quite favorably with those of Unger and Moult [93a]. We examined the performance of our algorithm on the same initial random sequences proposed by Unger and Moult. For two of their 27 length samples, and eight 64 length samples, the minimal energies located by our algorithm were significantly better and required only one tenth the number of energy evaluations (see Table 1.). In all cases the standard GA approach met or exceeded the energy value found through the hybrid approach. Clearly the standard GA approach seems to outperform the hybrid GA/Monte Carlo approach outlined by Unger and Moult. The global minimum configuration value for these random configurations is unknown. Certainly several of these sequences allow for multiple minimal configurations. By observing the hydrophobic moment ratio, we can measure how close we are approaching the theoretical

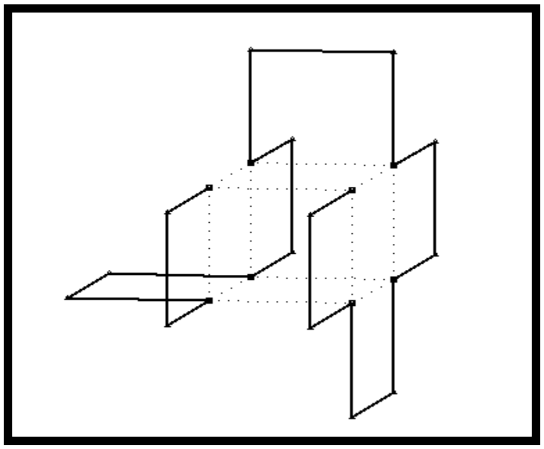

Figure 1. Sample Minimal Looped Cube

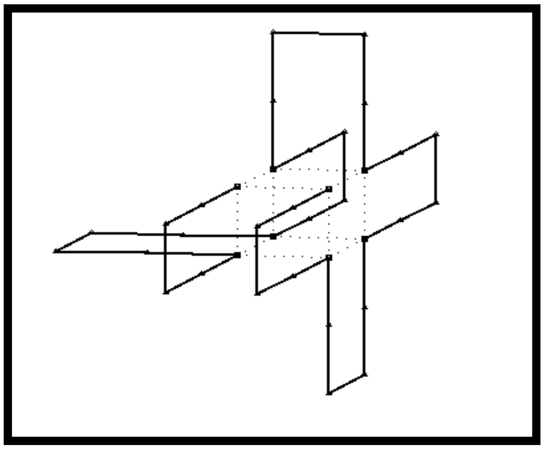

Figure 2. Sample 2-Looped Cube

relative to a fixed contact award value.

Note the following about this evaluation function:

- each contact gives a 2 point value (one from each end of the contact),
- adjacent hydrophobes in the protein's primary sequence are not discounted in the scoring,
- multiple collisions (i.e. 3 or more peptides) at a single point result in only a single penalty,
- hydrophobe collisions carry an implied additional penalty since no contacts are scored for collided hydrophobes.

This energy value computation is different than Unger and Moult [93a]; it has the following benefits:

- it allows illegal states (collisions),
- it is positively oriented (i.e. the problem is stated as a maximization rather than minimization),
- testing for sequential hydrophobes on the protein chain need not occur at each evaluation.

Essentially for cases where there is no collision along the protein chain, the equivalent Unger and Moult energy value can be obtained by multiplying ours by -0.5 then adding the number of hydrophobic contacts originally present in the protein's primary sequence. Since all test cases cited in this paper resolved to configurations without collisions, we report only the equivalent Unger-Moult energy value to allow comparison between the two works.

2.4 ALGORITHM IMPLEMENTATION

For the genetic algorithm engine, we chose the GALLOPS [Goodman 94] code developed at Michigan State University (essentially a greatly enhanced parallelizable version of SGA). For our crossover, selection, and mutation operators we have chosen to use fairly standard operators such as one and two-point crossover, tournament selection, and bit mutation.

2.5 PARAMETER TUNING

In any GA implementation, choosing the operators and tuning the parameters, such as population size, crossover rate, etc., can be a daunting task, especially when there is little or no similar experience to draw from. If the evaluation is sufficiently expensive, researchers are often forced to simply choose "reasonable" values from commonly accepted ranges, or find some way of estimating the effect of different values on the experiment. (e.g. Increased population size increases diversity thereby reducing convergence speed, etc.) One of the benefits of choosing a reduced domain model is that evaluation becomes simpler and therefore faster. Such is the case with this model; our evaluation cost is relatively small for our hardware platform. Therefore, we have been able to measure the effects of several parameters of consequence.

After much testing, the standard parameter set we chose to use for our testing was:

- **preference order** encoding
- .95 two-point crossover rate
- bit-mutation with a average rate of 1 mutation for every 1000 bits
- crowding with a crowding factor of 20 (allows for niching)
- incest reduction factor of 20 (favors crossovers of dislike parents)
- a population of 1000 genotype strings
- a collision penalty of -2 per position collided (not per collision)

Note that the net effect of most of these parameters (large population size, high replacement rate, high levels of crowding and incest reduction) is to preserve diversity. Indeed, preserving diversity, or preventing premature convergence seemed to played a major part in the eventual success of our implementation. Use of high crowding factor (replacing the *closest* individual in hamming distance out of a pool of **n** individuals *selected* for replacement) and incest reduction (choosing the individual *furthest* in hamming distance from a pool of **n** *selected* for breeding) were key in preventing premature convergence.

3. GENERALIZED SCORE & TESTING

As previously mentioned, there are several reasons why study of the simplified 3-D lattice model for protein folding seems natural. While it is not our intent to argue for reduction of this problem to a sort of "blocks world" digression, we propose a standardized conformation validity measurement which is invariant to the test sequence length and number of hydrophobes, as well as a series of test sequences with known optima. These may both ease comparison of results as well as allow the researcher to more easily tune the operation of the GA, as they did in our case.

3.1 GENERALIZED REPORTING VALUE

Since our central goal is to approach prediction of actual protein conformations, we would expect comparison of results on actual protein sequences to actual solved structures to be an ideal measure of effectiveness; however, the lattice model is such a strongly reduced model as to make such comparisons nearly impossible. Most biochemists agree that hydrophobic what resembles typical simulated annealing approaches. The net effect of rejection of lethals and the high bias of the Monte Carlo filtering, we feel, is to unduly bias the search capabilities of the GA.

2. IMPLEMENTATION

This section describes the details of our GA implementation as an approach to this problem. Our model is directly based on Unger and Moult's [93a] domain model to allow the reader to directly contrast the two approaches.

2.1 DOMAIN MODEL MODIFICATION

Unlike Unger & Moult, who choose to ignore illegal conformations, we modify the model slightly, penalizing for positional collisions, rather than simply discarding individuals having illegal states. This essentially allows the search to proceed through illegal states. While this is obviously not possible in the real world, and hence not a real world model, it is our purpose here to study the result of the given hydrophobic forces without necessarily *per se* modeling the path through which that state is reached. It may still be argued that this model will be able to reach states which are unobtainable in the real world; however, the same indictment applies to all other predictive approaches, except possibly the kinetic model.

2.2 ENCODING

Our initial approach to encoding an individual folded state for a protein was to represent a single **relative move** for each peptide in the protein. Thus, each peptide has 5 possible values (up, down, right, left, and forward) encoded in 3 bits, which disallows immediate reversal of the most recent move (i.e. backwards) in order to somewhat reduce the number of potential illegal states. Each move is accompanied by a corresponding logical shift in the relative coordinate system so that forward always repeats the last move and four duplicate turns return to the starting point. This encoding exhibits the problem that initial populations (randomly initialized) tend to have increasing numbers of collisions as the length of the protein increases; therefore, the GA appears to spend much time in untangling and pushing the peptides "outward" before good results can be obtained. In general, this representation allows a high number of collisions, which is undesirable.

In an attempt to reduce the number of illegal states searched, we enlarged the representation from 3 to 7 bits per peptide in order to encode one of the 120 permutations of the five allowable directions for each. Thus, each peptide was represented with an ordered list of preferred directions (for example: up, down, right, left, forward). Thus, if the first direction (e.g. up) were blocked, the secondary choice (e.g. down) would be tested. If the second (e.g. down) were also already occupied, the third (e.g. right) would be tested, and so on. Note that this encoding is biased toward peptides toward the front of the chain. The general form of the permutation encoding used is as follows:

```
Given: p from 0 to n!
        j = ( n - 1 )!;
        for ( i = 0; i < (n - 1); i++)
              {choice [ i ] = p / j;
                p %= j;
                j /= ( (n - 1) - i );      };
        choice [ (n - 1) ] = p;
        for (i = 0; i <= n; i++) next [ i ] = i;
        for (i = 0; i < n; i++)
              {d = -1;
              for (k = 0; k <= choice [ n ]; k++)
                   {prev = d;
                     d = next [d+1];};
              next [prev + 1] = next [d+1];
              direction [ i ] = d;};
Where n = # of items in permutation (5 in this case)
```

The cost of this decoding algorithm is $\mathbf{O(n^2)}$ time, where **n** is the number of items being permuted. While this becomes expensive for large **n**, it is not overly expensive for our purposes. Also, we may choose to use a lookup table instead, trading space for time. Use of a lookup table has the added benefit that different encoding schemes may be used interchangeably.

The permutation encoding in the algorithm above is not grey encoded, so a mutation will have somewhat random effects. A GA using this **preference order** encoding begins at a much faster pace than with the **relative move** encoding since it tends to avoid collisions by design; however, both encodings seem to converge to the same energy value over time for a given test sequence.

2.3 OBJECTIVE FUNCTION

An individual genotype is evaluated under each encoding by plotting the course encoded by the genotypic movement list in a 3-D lattice (making appropriate choices under the preference ordering encoding). After the protein is plotted, each occupied cell is tested for the presence of a hydrophobe. If one is present and no additional peptides (collisions) are detected in the cell, all "adjacent" points (adjacent by a single edge in the 90° lattice) are tested for the presence of a hydrophobe, and a single point is awarded for each contact. At the same time, each cell is tested for the presence of multiple peptides. If two or more are present, a single penalty is accessed. The penalty value was made a parameter of the evaluation which allowed us to test multiple values have been made toward applying such methods to general tertiary structure prediction [Benner 88]; however, such approaches have met with limited success, in part because of the small number of “solved” structures and the tendency for such data to be biased toward certain families of protein structures; but primarily because such a technique has difficulty taking into account the effects of the larger protein structure context which may override such local tendencies.

The primary hill climbing search technique used for locating low energy protein conformations is standard Monte-Carlo search. Under this method, the protein is randomly slightly mutated, and the likelihood of a new configuration being favored over its predecessor is directly linked to the net change in energy (positive movements are always accepted, and negative movements are randomly accepted). Typically the algorithm is slowly biased toward straight hill-climbing (accept only positive moves) and therefore takes on the features of simulated annealing.

1.2 PRECEDING WORK

This work is largely based on the work published by Unger & Moult in the area of energy function optimization of protein chains using reduced constraint models, especially that of [Unger & Moult 93a]. The following is a brief review of that work.

1.3 DOMAIN MODEL

Direct modeling of peptide chain orientation is a difficult undertaking. Unger and Moult propose a reduced protein 3-D lattice protein folding model for studying the application of GA driven optimization to configuration energies. In this model, peptides are represented as single point units without side chains. The protein structure is constrained to a 90° lattice (6 degrees of freedom at each point) with the peptides occupying intersections in the lattice. Individual amino acids may be classified as either fully hydrophobic (b) or fully hydrophilic (w) (no relative strengths). Only hydrophilic/hydrophobic forces are considered (i.e. hydrogen & covalent bonds are ignored). Cross-chain hydrophobic contacts (having two hydrophobes which are not adjacent in the primary sequence occupy adjacent points in the lattice) will be considered the primary basis for evaluation, and adjacency will be considered only in cardinal directions (i.e. immediately up, down, left, right, forward, or back from a given point).

Use of a reduced model such as the 3-D lattice is common with experimenters beginning investigations into the applicability of GA methodologies to the problem of protein structure prediction. Such reduced model approaches are natural as they allow the investigators to explore the effects of their theories without requiring the level of commitment required in producing a more detailed level of evaluation. Although the behavior of the GA with the lattice model cannot be conclusively linked to the behavior of more detailed models, such examinations also allow us to directly compare the success of encoding approaches and operators, etc. against other published results.

1.4 HYBRID GA APPROACH

The encoding proposed by Unger and Moult is a direct encoding of the direction of each peptide from the preceding peptide (5 degrees of freedom, disallowing collisions). The evaluation function solely evaluates non-sequential hydrophobe to hydrophobe contacts and is stated as a negative value (-1 per contact) with larger negative values indicating better energy conformations (thus stating the problem in terms of minimization).

The algorithm proposed by Unger and Moult begins with a population of identical unfolded configurations. Each generation begins with a series of K mutations being applied to each individual in the population, where K is equal to the length of the encoding. These mutations are filtered using a Monte Carlo acceptance algorithm which disallows lethal configurations (those with collisions), always accepts mutations resulting in better energy, and accepts increased energy mutations based upon a threshold on the energy gain which becomes stricter over time. One-point crossover with an additional random mutation at the crossover point follows, producing a single offspring for each selected pair of parents; however, lethal configurations (those with collisions) are rejected. In this situation, the crossover operation is retried for a given pair of parents until a nonlethal offspring can be located. Offspring are accepted using a second Monte Carlo filter which accepts all reduced energy confirmations and randomly accepts increased energy offspring again using a cooling threshold on the energy gain. The algorithm uses 100% replacement of all individuals in a generation through crossover except the single best, *elitist*, individual.

Test data consisted of a series of 10 randomly produced 27 length sequences and 10 randomly produced 64 length sequences. The algorithm operated on each of the 27 and 64 length sequence for roughly 1.2 million and 2.2 million function evaluations respectively using a population size of 200. Performance comparisons were given between the above algorithm and a pure Monte Carlo approach which greatly favored the former.

While the encoding and evaluation function proposed by Unger and Moult are fairly straight forward, the algorithm differs from a standard GA approach in several aspects. Most notable are the nonrandom initialization, the high level of mutation, and the Monte Carlo filtering of both the mutation and crossover results which some-

A Standard GA Approach to Native Protein Conformation Prediction

Arnold L. Patton, W. F. Punch III and E. D. Goodman
Genetic Algorithm Research and Application Group (GARAGe)
A714 Wells Hall, Computer Science Dept.
Michigan State University
E. Lansing MI, 48824
pattona@cps.msu.edu, punch@cps.msu.edu, goodman@egr.msu.edu

ABSTRACT

Finding the 3-D geometry or tertiary structure of an arbitrary protein is vital to understanding the functionality of that protein. The prediction of this structure, known as the protein folding problem, is very difficult and has been labeled one of the "grand challenge problems" for the scientific community. We report here on further work to determine tertiary structures via genetic algorithms. We build on work done first by Unger and Moult using a simplified protein model but improve on the application of GAs to this model. We show, using the same simplified model, that the genetic algorithm indeed appears effective for determining tertiary structure with far fewer computational steps than first reported.

1. INTRODUCTION

In this paper we explore the applicability of a standard GA approach to the problem of protein structure prediction. This section presents the problem of protein structure prediction in detail. For the reader with a basic understanding of biochemistry and the nature of protein folding, this section may be skipped without loss of understanding. For more information on general biochemistry, see [Voet & Voet 90]; for a more detailed treatment of protein folding issues see [Fasman 89] and [Robson & Garnier 88]. For further references see [Benner, Gerloff, and Jenny 94].

1.1 PROBLEM DESCRIPTION

Finding the natural tertiary (3-dimensional) shape of a protein in aqueous solution is of primary concern to biochemists, because much of a protein's function may be derived from its conformation. Currently, biochemists use techniques such as MRI (magnetic resonance imaging) and X-ray crystallography on protein crystals in order to "view" the conformation of a protein. These techniques are expensive, in terms of equipment, computation and time. Additionally, both of these techniques require isolation, purification, and crystallization of the target protein, which may be difficult or impossible depending upon the particular protein under study.

While research into finding faster, cheaper methods of detecting the 3-dimensional structure of proteins continues, much effort has also been expended in attempting to predict or model the resting state of a protein. The problem of ab initio prediction of the final folded conformation of a protein chain from peptide sequence information and general chemical and biochemical theory and knowledge is an extremely difficult problem. In fact, the complexity of this problem has been shown to be NP-hard [Fraenkel 93; Unger & Moult, 93b]. Currently there are three primary approaches to this problem: molecular dynamics modeling [York, Darden, Pedersen, and Andersen 93], statistical prediction [Benner 88; Robson, Benner, Garnier, Barton and Russell 92; Rooman & Wodak 92], and hill-climbing search techniques [Camacho & Thirumalai 93; Skolnick & Kolinski 90]. We will consider each approach in turn.

Molecular dynamics modeling principally consists of applying atomic level physics in a brute force attempt to model the force vectors incident on a protein chain. The basic technique is to compute the force vectors for a given static state, and then move the components in accordance with the computed forces, until sufficient change has occurred to require recomputing of the incident forces. This technique is extremely compute intensive and time consuming and because of this is only applicable for studying very small time portions a single protein folding event.

Statistical prediction is a technique whereby measurements for the predilection of an amino acid (or short sequence of amino acids) to participate in certain well identified substructures (known as secondary structures) such as alpha-helices, beta-sheets, etc. are gathered from "solved" proteins for which the tertiary structure is known. These statistics are then applied in a predictive fashion to the protein under investigation. A growing number of these approaches have been quite successful at predicting secondary structures, and recent strides