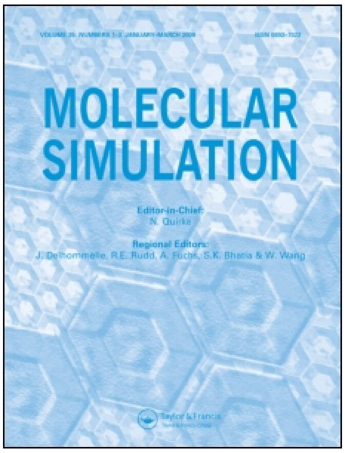

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

The Genetic Algorithm and the Conformational Search of Polypeptides and Proteins

Scott M. Le Grand[ab]; Kenneth M. Merz Jr.[a]
[a] Department of Chemistry, The Pennsylvania State University, University Park, PA [b] 219 Molecular Biology Institute, Los Angeles, CA

To cite this Article Grand, Scott M. Le and Merz Jr., Kenneth M.(1994) 'The Genetic Algorithm and the Conformational Search of Polypeptides and Proteins', Molecular Simulation, 13: 4, 299 — 320
To link to this Article: DOI: 10.1080/08927029408021995
URL: http://dx.doi.org/10.1080/08927029408021995

Molecular Simulation, 1994, Vol. 13, pp. 299-320
Reprints available directly from the publisher
Photocopying permitted by license only

THE GENETIC ALGORITHM AND THE CONFORMATIONAL SEARCH OF POLYPEPTIDES AND PROTEINS

SCOTT M. LE GRAND[1] and KENNETH M. MERZ, JR.

Department of Chemistry, The Pennsylvania State University, University Park, PA 16801

(Received November 1993, accepted April 1994)

The genetic algorithm is a technique of function optimization derived from the principles of evolutionary theory. We have adapted it to perform conformational search on polypeptides and proteins. The algorithm was first tested on several small polypeptides and the 46 amino acid protein crambin under the AMBER potential energy function. The probable global minimum conformations of the polypeptides were located 90% of the time and a non-native conformation of crambin was located that was 150 kcal/mol lower in potential energy than the minimized crystal structure conformation. Next, we used a knowledge-based potential function to predict the structures of melittin, pancreatic polypeptide, and crambin. A 2.31 Å ΔRMS conformation of melittin and a 5.33 Å ΔRMS conformation of pancreatic polypeptide were located by genetic algorithm-based conformational search under the knowledge-based potential function. Although the ΔRMS of pancreatic polypeptide was somewhat high, most of the secondary structure was correct. The secondary structure of crambin was predicted correctly, but the potential failed to promote packing interactions. Finally, we tested the packing aspects of our potential function by attempting to predict the tertiary structure of cytochrome b_{562} given correct secondary structure as a constraint. The final predicted conformation of cytochrome b_{562} was an almost completely extended continuous helix which indicated that the knowledge-based potential was useless for tertiary structure prediction. This work serves as a warning against testing potential functions designed for tertiary structure prediction on small proteins.

KEY WORDS: Genetic algorithm, conformational search, polypeptides, proteins

INTRODUCTION

In 1959, Anfinsen [1] demonstrated that the primary structure of a protein (the sequence of amino acids) can uniquely determine its tertiary structure (three dimensional conformation). This implied that there must be a consistent set of rules for deriving a protein's tertiary structure from its primary structure. The search for these rules is known as the "protein folding" problem. Despite many creative attempts, these rules have not been determined [2]. Currently, the primary structures of approximately 40,000 proteins are known. However, only a small percentage of those proteins have known tertiary structures. A solution to the protein folding problem will make 40,000 more tertiary structures available for immediate study by translating the DNA

[1] Current address: 219 Molecular Biology Institute, 405 Hilgard Avenue, UCLA, Los Angeles, CA 90024. S.M.L. is the recipient of an NSF postdoctoral research fellowship in chemistry and a Cray Computers research fellowship in computational chemistry.

sequence information in the sequence databases into three-dimensional protein structures. This translation will be indispensable for the analysis of results from the Human Genome Project, *de novo* protein design, and many other areas of biotechnological research. Finally, an in-depth study of the rules of protein folding should provide vital clues to the protein folding process. The search for these rules is therefore an important objective for theoretical molecular biology.

Many theoretical efforts aimed at solving the protein folding problem have involved the optimization of a potential energy function which approximates the thermodynamic state of a protein macromolecule. These efforts are based on the assumption that the global minimum conformations of proteins under these functions will correspond to their tertiary structures. Since the location of the global minimum conformation of a protein alone under such a potential function does not necessarily give any insight into how a protein folds, these approaches are known as protein tertiary structure prediction. A successful energy minimization-based protein tertiary structure prediction algorithm must satisfy two requirements. First, the conformational search technique it uses must be capable of locating the global minimum conformation of a protein under a specific potential energy function and second, the global minimum conformation of a given protein under the potential energy function must be close to the native structure of that protein. Researchers have developed numerous algorithms for the conformational search of small polypeptides [3, 4], hydrocarbons [5], and proteins [6–10]. Other groups have developed potential energy functions for molecular modeling [11–13] and tertiary structure prediction [14–17]. Unfortunately, a combined conformational search algorithm and potential function which satisfy both of the aforementioned requirements for successful protein tertiary structure prediction have not yet been developed [6, 7, 15, 30].

THE GENETIC ALGORITHM

The work to be described here involves the adaptation of the genetic algorithm to perform conformational search for the purpose of protein tertiary structure prediction. The genetic algorithm is an optimization technique derived from the principles of evolutionary theory [18, 19]. It has been applied to a myriad of optimization problems such as the traveling salesman problem, neural network optimization [20–24], scheduling [25], machine learning, pattern recognition, and the solution of nonlinear equations. See [19] for a review of these applications. Biological applications of the genetic algorithm include NMR refinement of small nucleotides [26, 27], the conformational search of a lattice-based proteins [28, 29], protein sequence design [29], and the conformational search of polypeptides and proteins [6–8, 30, 31].

Figure 1 illustrates a typical genetic algorithm as described by Grefenstette and Baker [32]. A genetic algorithm begins by encoding the k independent variables of an optimization problem as genes in a chromosome. For example, in conformational search, the genes would be the conformation determining dihedral angles of a molecule (Fig. 2). Next, a population of N chromosomes (hereafter known as $P(t)$) is initialized with random values for each of the genes in each chromosome (Fig. 3). After this *initialization* step, the function value of the point in parameter space represented by each chromosome x is evaluated and called the chromosome's fitness $u(x)$. In confor-

```
procedure GA
begin
    $t = 0$;
    initialize $P(t)$;
    evaluate structures in $P(t)$;
    while termination condition not satisfied do
    begin
      $t = t + 1$;
      select $M(t)$ from $P(t-1)$;
      recombine structures in $M(t)$;
      evaluate structures in $M(t)$;
     replace some or all of $P(t-1)$ with $M(t)$ to form $P(t)$
    end
end.
```

Figure 1 Flowchart of a genetic algorithm.

mational search, the fitness would be the potential energy of the conformation of the molecule represented by the chromosome (Fig. 4).

After initialization, a genetic algorithm cycles through rounds of *selection*, *recombination* and *evaluation* until termination conditions are met. During the first phase, *selection*, the mating population $M(t)$ is selected from $P(t)$. $M(t)$ consists of one or more pairs of chromosomes known as parents. There are numerous methods of selecting $M(t)$ from $P(t)$ [19, 21]. One popular method of selection is knwon as *proportionate*

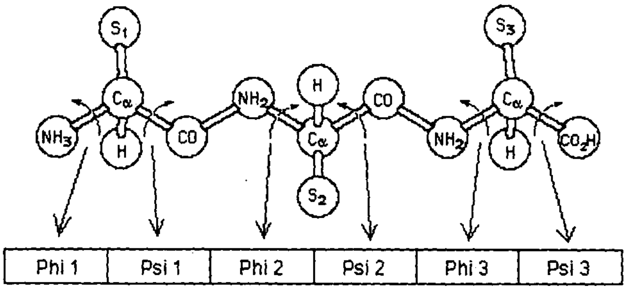

Figure 2 The encoding of a molecule's conformational variables as genes in a chromosome.

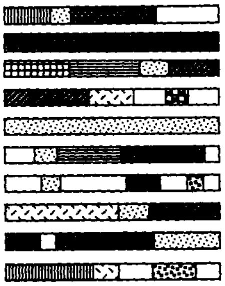

Figure 3 A randomly generated population of chromosomes.

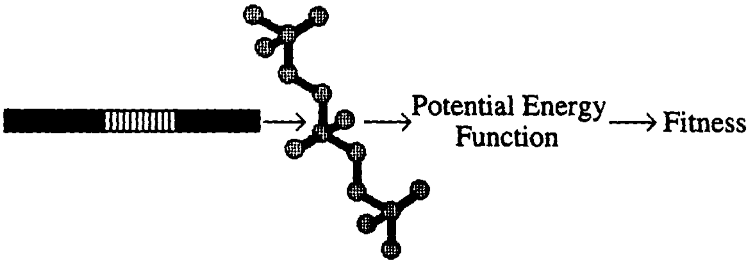

Figure 4 The evaluation of the fitness of a chromosome which represents a conformation of a molecule.

selction. In *proportionate selection*, a given chromosome x is selected for inclusion into $M(t)$ with probability $p(x)$, which is proportional to the ratio of its fitness to the mean fitness of the population $\bar{u}(t)$ (equation 1) (Fig. 5).

$$p(x) \approx \frac{u(x)}{\bar{u}(t)} \tag{1}$$

During *recombination*, the genes in the pairs of parents in $M(t)$ are mixed together to produce hybrid chromosomes (hereafter to be called *children*) via the use of operators which perform processes analogous to genetic crossover and mutation. The first phase of recombination is the use of a crossover operator to create a hybrid child from each of the pairs of parents in $M(t)$. There are many crossover operators in use [19, 33–36]. The most common crossover operator, known as simple two-point crossover, creates a child containing all the genes from the beginning of one parent's chromosome up to a cut point, and the rest of its genes from that cut point to the end of the chromosome from the second parent (Fig. 6a). A second child can be created from the genes in both parental chromosomes which are not in the first child if desired. In conformational search, simple two-point crossover creates a child which takes one section of the molecule's dihedral angles from the first parent while the complementary section of the molecule takes its dihedral angles from the second parent. A second popular crossover operator is known as two-point wraparound crossover. In two-point wraparound

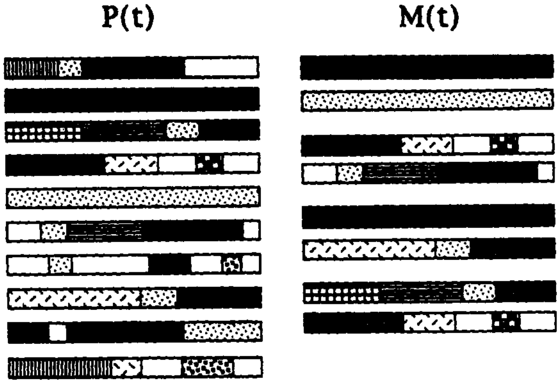

Figure 5 Selection of $M(t)$ from $P(t)$.

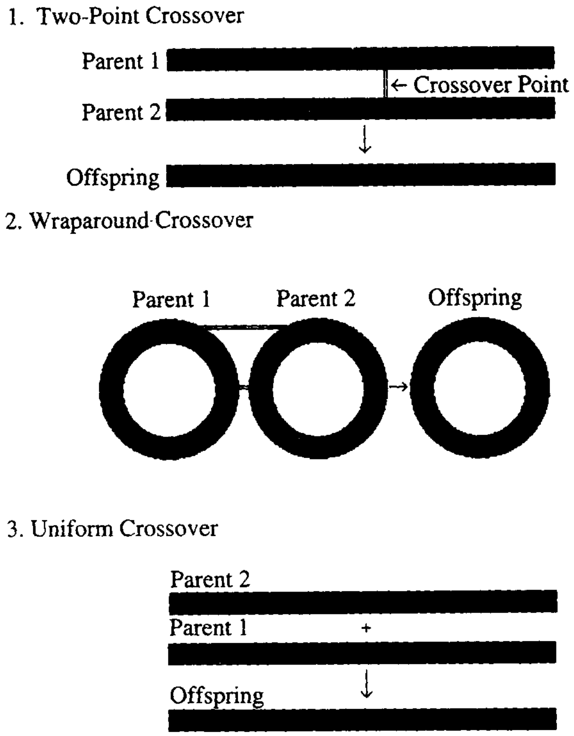

Figure 6 Three popular methods of crossover which are used in genetic algorithms.

crossover, the chromosome is treated as a ring: a child's chromosomes is created from an arc segment out of the first parent's chromosomal and the complementary arc segment from the second parent's chromosome (Fig. 6b). In conformational search, two-point wraparound crossover creates a child whose outer dihedral angles are from one parent and inner dihedral angles are from the other parent. The use of two-point wraparound crossover is thought to help transfer genes together which are on opposite ends of the chromosome which would otherwise tend to be broken apart by simple two-point crossover. A third popular crossover operator is known as uniform crossover. In uniform crossover, each gene is taken from either parent with equal probability based on the value of a random variable. In conformational search, uniform crossover would create a child whose dihedral angles were randomly selected from either parent. Like the use of two-point wraparound crossover, the use of uniform crossover also helps to solve distance-dependent crossover problems. However, it can also disrupt pairs of genes near one another that would otherwise likely to be transferred together during crossover.

The second phase of *recombination* is known as mutation. During mutation, parts of each child's chromosome are altered slightly by operators which perform processes

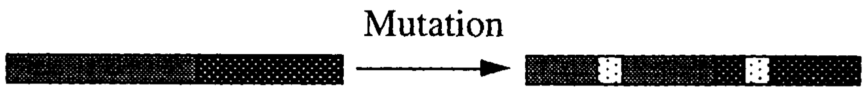

Figure 7 Illustration of the effect of mutation on a chromosome.

analogous to genetic mutation (Fig. 7). As with crossover, there are many mutation operators in use [19, 21, 37]. One such method is to give each of a child's genes a 3–5% chance of being changed to a random value based on the value of a random variable. In conformational search, the aforementioned mutation operator would randomly change the values of several dihedral angles.

During *evaluation*, the fitnesses of the chromosomes in $M(t)$ are evaluated in the same manner as after initialization. Once the fitnesses of the chromosomes in $M(t)$ have been calculated, they replace all or some of the members of $P(t)$ to form $P(t + 1)$ during *replacement*. This cyclic process of *selection*, *recombination*, *evaluation*, and *replacement* repeats until user-specified termination conditions are met.

Many variations on this basic theme are possible; there are several good introductions to the subject which cover both this basic approach [38–40] and many of these variations [19]. Holland [18] presents a rigorous analysis of the genetic algorithm. Specific modifications of the genetic algorithm to improve its performance at conformational search have been described in more detail [6–8, 16, 30].

CONFORMATIONAL SEARCH UNDER AMBER

Our first tests of the genetic algorithm's ability to perform conformational search were the conformational search of several polypeptides and the 46 amino acid protein crambin under the AMBER potential energy function [11] with a distance-dependent dielectric constant. The results demonstrated that the genetic algorithm was well-suited to perform conformational search and also demonstrated some deficiencies in the AMBER molecular mechanics potential energy function. When appropriate, the convergence of genetic algorithm-based conformational search to the global optimum conformation was demonstrated by multiple independent runs from different starting conditions on the same molecule that produced the same or very similar final

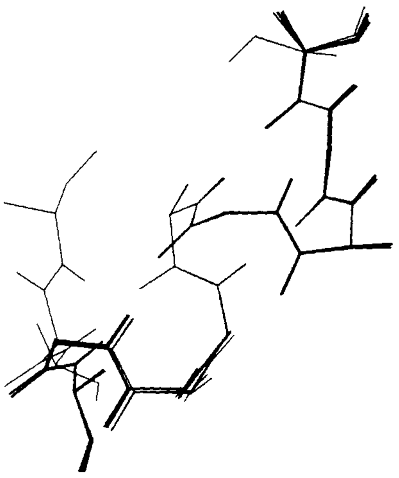

Figure 8 Superimposition of 10 final conformers of the polypeptide Ala_9 produced by genetic algorithm-based conformational search under the AMBER potential energy function.

structures. This is the only practical way to demonstrate that the global minimum conformation has been located.

Genetic algorithm-based conformational search was first performed on three small polypeptides. Each of these polypeptides had 18 degrees of freedom which made each conformational search an 18 variable optimization problem.

Figure 8 plots ten superimposed independent genetic algorithm-based conformational searches of the polypeptide A_9 under the AMBER potential energy function. The potential energies of the final conformers produced by these conformational searches ranged from -98.42 to -98.50 kcal with an average ΔRMS (Root mean square deviation) between runs to 0.20 Å. The final conformation which is obtained in all runs is an α-helix. One of the ten runs generated a final structure which is identical to the nine other final structures except that it had a slightly misaligned carboxyl terminal which raises its potential energy by approximately 0.7 kcal. Alanine rich polypeptides are known to form an α-helical structure in solution but polyalanine is insoluble [41]. These runs consumed approximately two hours of CPU time on one processor of an IRIS 4D/220 workstation (240,000 iterations).

Figure 9 plots the final structures generated by ten independent genetic algorithm-based conformational searches of the polypeptide AGAGAGAGA. Similar to the runs on A_9, we obtained almost identical results in nine out of ten runs with an average ΔRMS of 0.3 Å between the final structures generated by these nine runs. This is a slightly higher ΔRMS between final structures than with the A_9 runs, but glycine lacks a side chain group which makes it more flexible than alanine so this is not an unexpected result. The final conformation we obtained is an α-helix. In the nine mostly identical final conformers, the potential energies ranged from -94.69 to -94.78 kcal.

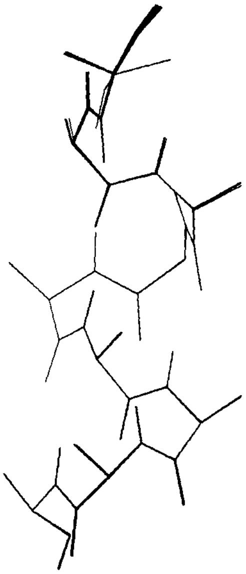

Figure 9 Superimposition of ten final conformers of the polypeptide AGAGAGAGA produced by genetic algorithm-based conformational search under the AMBER potential energy function.

In every run, most of the backbone of the final conformer is oriented as an α-helix, but in the one somewhat different final conformer, the only substantial difference from the final conformers generated by the nine other runs appears at the carboxyl end of the helix, where the final carboxyl group is improperly oriented relative to the rest of the helix. Glycine α-helices have been found as global minimum conformations of glycine-rich polypeptides by other groups [42], but have never been observed. Glycine is the most flexible amino acid, and it is probable that there are many low energy conformations which would be sampled by this polypeptide if it were synthesized and placed in solution, outweighing this helical structure by shear number [43]. Unfortunately, there is no easy way to account for this flexibility in an energy minimization-based conformational search because the objective of such algorithms is to locate a single lowest energy structure rather than sample the ensemble of low energy conformations. Similar to the runs on Ala_9, each of these runs consumed an average of two hours of CPU time on one processor of an SGI IRIS 4D/220 workstation (240,000 iterations).

The suspected global minimum conformation of [met]enkephalin (YGGFM) under the ECEPP potential energy function [13] has been located by several groups by a variety of methods [44–46, 48, 49] and once under AMBER [49]. In our first ten runs on [met]enkephalin, we consistently coverged to a structure which bore little resemblance to theirs, and which had a higher AMBER potential energy (~ -42 kcal) than the structure (~ -47.2 kcal) generated by the other groups even after extensive gradient minimization of our final structure. However, when we slightly altered several bond angles and bond distances of our amino acids near the alpha carbon which are not otherwise varied during a run, we obtained a structure more similar to that found by the other groups in eight out of ten runs (Fig. 10) all of which had lower potential energy than their structure under the AMBER potential energy function, with mostly identical backbone structure in all ten runs. In the nine successful runs, our final potential energies ranged from -48.67 to -48.94 kcal, and had an average ΔRMS of < 0.5 Å between successful runs. The conformation generated by the one completely unsuccessful run has a properly oriented backbone, but improperly oriented side chains. A second run properly oriented everything except for a methionine side chain.

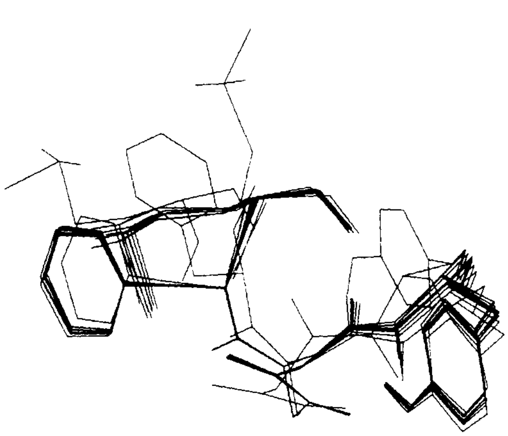

Figure 10 Superimposition of ten final conformers of the polypeptide [met]enkephalin produced by genetic alogrithm-based conformational search under the AMBER potential energy function.

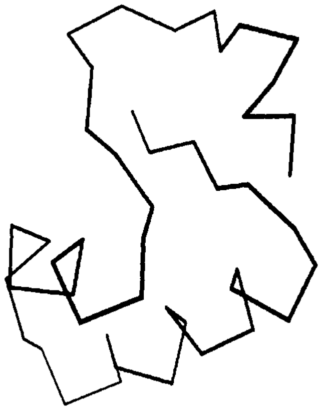

Figure 11 Crystal structure of the 46 amino acid protein crambin.

As with the two previous polypeptides, these runs consumed an average of approximately two hours of CPU time (250,000 iterations) on one processor of an IRIS 4D/220 workstation. It is difficult to compare our results with the work of others because they were either under a different force field [44–46, 48, 49] or only provided relative rather than absolute final energies [47]. However, a recent paper by Unger and Moult [31] shows a signficantly superior performance of the genetic algorithm over Monte Carlo methods when used for the conformational search of 2D lattice proteins.

Crambin (1 crn) (Fig. 11) is a 46 amino acid protein that is found in the embryonic tissue of seeds from *Crambe abyssinica*. Its structure has been solved to 0.83 Å resolution [50]. Crambin has 243 degrees of freedom which makes its conformational search a 243 variable optimization problem. Only one conformational search of crambin was performed because of the nature of the results. A single genetic algorithm-

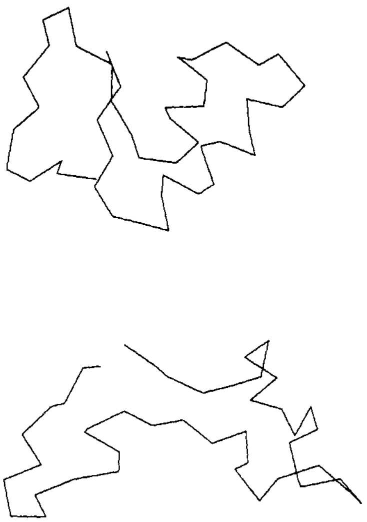

Figure 12 Comparison of crambin native structure and final conformer produced by genetic algorithm-based conformational search under the AMBER potential energy function.

based conformational search produced a final conformation which is probably not the global minimum conformation, but which has an AMBER potential energy 150 kcal/mol lower than that of the known crystal structure (Fig. 12). Unfortunately, our final conformation bears little resemblance to the known crystal structure. Brooks *et al.* [51] has stated that the minimum energy conformation of a large molecule under a potential energy function such as AMBER in the absence of solvent would be an "inside-out" protein. Large side chains would stick out of the protein, and small polar side chains would be closely paired with other such side chains in its interior. Our protein resembles an inside-out protein so our results may correspond to this situation. Our final structure appears to be mainly stabilized by electrostatic interactions and hydrogen bonds between the backbone and polar side chain groups (Fig. 12). The strength of these interactions is increased at close range because of the use of the distance-dependent dielectric constant. This indicated that potential functions used for tertiary structure prediction must account for hydration, which is believed to be a driving force for protein folding [52–54], in a more realistic fashion. Unfortunately, in this case, even including hydration energy as calculated with atomic solvation parameters [55], is not enough to make the native structure lower in potential energy [56].

CONFORMATIONAL SEARCH UNDER A KNOWLEDGE-BASED POTENTIAL

The failure of the AMBER potential energy function for the conformational search of crambin spurred the development of a new potential function based on the statistical analysis of a set of known protein structures [7, 30]. Such potential functions are generally known as knowledge-based potential functions. A knowledge-based potential function is generated by first collecting statistical data for a set of features observed in a set of known protein structures. Next, the frequencies of occurrence of each of these features are converted into free energies via inserting them into the Boltzmann equation. In the case of this potential, the frequencies of occurrence of 14,231 classes of nonbond interactions, 167 classes of atomic burial states, and 87 different classes of dihedral angles were collected and converted into free energies via the method of Sippl [15]. As with the work of Sippl, the average folded state of proteins was used as a zero energy reference state in the Boltzmann equation. The design of this potential is similar to that developed by Jones *et al.* [57] except for the use of a greater number of different atom classes and the collection of dihedral angle statistics. The results of performing genetic algorithm-based conformational search on the proteins melittin, pancreatic polypeptide, crambin, and cytochrome b_{562} were mixed and illustrated that these potential functions have their own deficiencies.

Melittin (2 mlt) is a 26 amino acid protein which is the main component of bee venom. Melittin exists as a tetramer in the crystal structure which has been refined to 2.0 Å resolution [58]. The melittin monomer is bent α-helical rod which is kinked around residues 10–13. Residues 1–10 and residues 13–26 form straight α-helices which are bent at a 120 degree angle relative to one another (Fig. 13). The tetramer is composed of two identical dimers. The two chemically identical chains in the dimer have a ΔRMS of 1:1 Å relative to one another. NMR studies of monomeric melittin bound to a lipid water interface have determined an ensemble of α-helical conformations with an

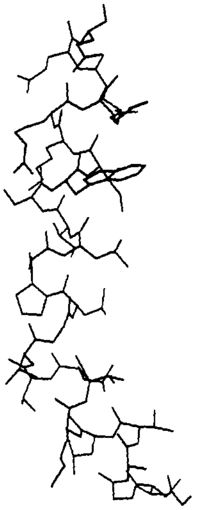

Figure 13 Crystal structure of the melittin monomer.

average ΔRMS of 1.6 Å relative to one another [59]. In a united atom representation, melittin has 97 dihedral angles which determine its conformation.

Seven genetic algorithm-based conformational searches were performed on the melittin monomer. In all seven runs, the final structure generated was two α-helical domains separated by a kink in residues 11–13 (Fig. 14). The average overall ΔRMS between independent conformational searches of the melittin monomer was 2.49 Å while the average ΔRMS from the native conformation of chain A of melittin was

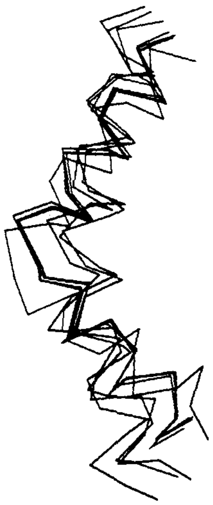

Figure 14 Superimposition of 7 final conformers of the melittin monomer produced by genetic algorithm-based conformational search under the knowledge-based potential.

3.53 Å. The high degree of similarity between final conformers generated by independent conformational searches of the melittin monomer indicated that genetic algorithm-based conformational search was probably locating final conformations under the knowledge-based potential function that were close to the global optimum conformation. The lowest ΔRMS from the native conformation of chain A of melittin was 2.31 Å while the highest ΔRMS from the native conformation of chain A of melittin was 5.90 Å. The average C_α ΔRMS (obtained by only comparing alpha carbons of each amino acid) between runs was 2.07 Å while the average C_α Δ RMS from the native conformation was 2.98 Å (Fig. 14). The lowest C_α ΔRMS from the native conformation is 1.34 Å while the highest C_α ΔRMS from the native conformation is 4.77 Å. In all cases, the final structure generated by genetic algorithm-based conformational search was lower in energy under the knowledge-based potential (−963 to −983 kcal/mol) than the native structure of chain A of melittin (−785 kcal/mol). The seven conformational searches of the melittin required from 316,000 to 442,000 energy evaluations to complete with an average run length of 375,000 energy evaluations. Although the highest ΔRMS of a final conformer, 5.90 Å, seems far from the native crystal structure, the calculated structures are correct above and below the kink region. This can be seen by comparing the ΔRMS of the isolated α-helical domains of each final conformer relative to each other and to the crystal structure of chain A of melittin. The average ΔRMS between runs of residues 1–10 is 0.234 Å for all atoms and 0.085 Å for the C_α atoms while the average ΔRMS from the crystal structure of chain A of melittin is 1.04 Å for all atoms and 0.298 Å for the C_α atoms alone (Fig. 15). The average ΔRMS between runs of residues 13–26 is 0.79 Å for all atoms and 0.30 Å for the C_α atoms while the average ΔRMS from the crystal structure of chain A of melittin is 1.52 Å for all atoms and 0.60 Å for the C_α atoms alone (Fig. 16).

Avian pancreatic polypeptide (1 ppt), which is 36 residues in length, is the smallest protein of known structure with tertiary packing interactions. This makes it a good target protein to initially test the ability of a potential function to correctly determine

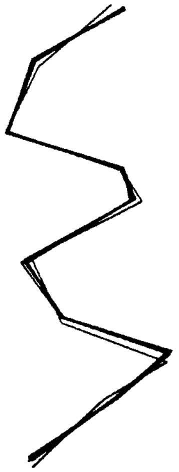

Figure 15 Superimposition of residues 1–10 of seven final conformers of the melittin monomer produced by genetic algorithm-based conformational search.

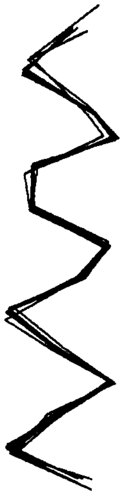

Figure 16 Superimposition of residues 13–26 of seven final conformers of the melittin monomer produced by genetic algorithm-based conformational search.

packing. Its crystal structure has been determined to 0.98 Å resolution with an *R*-factor of 0.156 [60]. The tertiary structure of avian pancreatic polypeptide is a polyproline helix formed by residues 2–8 which packs up against an α-helix which is formed by residues 14–32 (Fig. 17). CD experiments have determined that the interaction between these two helices is maintained in solution [61]. In a united atom representation, avian pancreatic polypeptide has 132 dihedral angles which determine its conformation.

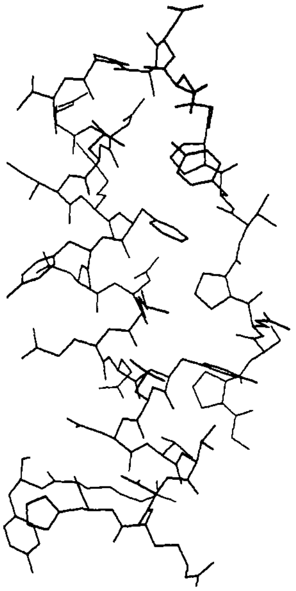

Figure 17 Crystal structure of avian pancreatic polypeptide.

Seven genetic algorithm-based conformational searches were performed on avian pancreatic polypeptide. The final structures produced by the seven conformational searches of pancreatic polypeptide were somewhat less similar to the native structure than the seven final melittin monomer conformers, but all produce the overall topology of pancreatic polypeptide with a polyproline helix formed by residues 2–8 which is packed up against an α-helix formed by residues 14–32 (Fig. 18). The average overall ΔRMS between runs is 4.58 Å while the average ΔRMS from the native conformation of avian pancreatic polypeptide is 6.77 Å. The lowest ΔRMS from the native conformation is 5.33 Å while the highest ΔRMS from the native conformation was 9.03 Å. The average C_α ΔRMS between runs is 3.79 Å while the average C_α ΔRMS from the native conformation is 6.03 Å. The lowest C_α ΔRMS from the native conformation is 4.75 Å while the highest C_α ΔRMS from the native conformation is 8.50 Å. Once again, the final structures generated by genetic algorithm-based conformational search was lower in energy under the knowledge-based potential (−1284 to −1404 kcal/mol) than the native structure of avian pancreatic polypeptide (−1107 kcal/mol). The seven conformational searches of avain pancreatic polypeptide required 356,000 to 525,000 cycles to complete with an average run length of 412,000 cycles.

Although the overall ΔRMS of every final conformer indicates that the tertiary structure is far from the native crystal structure and that there is a great deal of conformational diversity between runs, most of the secondary structure of each conformer is similar. This can be seen by comparing the ΔRMS of the isolated polyproline helix formed by residues 2–8 and the α-helix formed by residues 14–32. The average ΔRMS between runs of residues 2–8 is 0.65 Å for all atoms and 0.29 Å for the C_α atoms while the average ΔRMS from the crystal structure of avian pancreatic polypeptide is 3.13 Å for all atoms and 1.88 Å for the C_α atoms alone

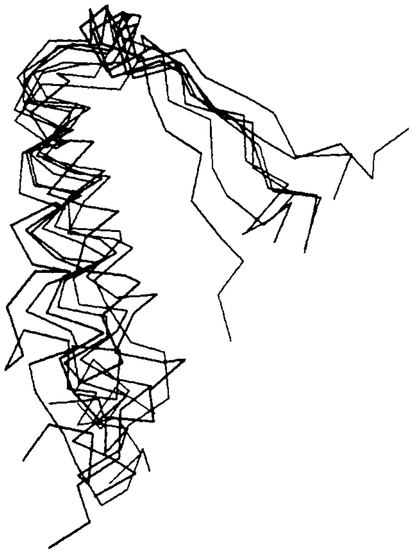

Figure 18 Superimposition of 7 final conformers of avian pancreatic polypeptide produced by genetic algorithm-based conformational search under the knowledge-based potential.

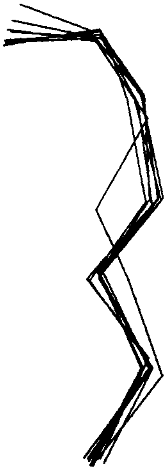

Figure 19 Superimposition of residues 2 to 8 of seven final structures of avian pancreatic polypeptide produced by genetic algorithm-based conformational search under the knowledge-based potential.

(Fig.19). Therefore, the polyproline helix has been somewhat poorly reproduced by the knowledge-based potential but is consistent between runs. The backbone conformations of all but the last 5 residues of the α-helix stretching from residues 14–32 are reasonably correctly predicted. The average ΔRMS between runs of residues 14–28 is 1.14 Å for all atoms and 0.43 Å for the C_α atoms while the average ΔRMS from the crystal structure of avian pancreatic polypeptide is 1.98 Å for all atoms

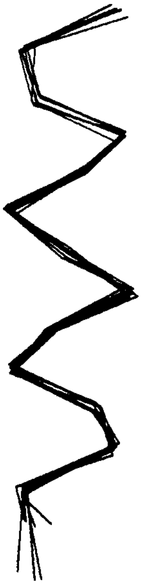

Figure 20 Superimposition of residues 14 to 32 of seven final conformers of avian pancreatic polypeptide produced by genetic algorithm-based conformational search under the knowledge-based potential.

and 0.70 Å for the C_α atoms alone (Fig. 20). The last 8 residues of pancreatic polypeptide are predicted as helical in 3 of the final conformers and as random coil in the other 4 final conformers.

The secondary structures of both melittin and avian pancreatic polypeptide were predicted well by the knowledge-based potential energy function. The tertiary structure of avian pancreatic polypeptide was crudely reproduced by the knowledge-based potential while the melittin monomer has none. Unfortunately, the results which will be presented below indicate the successful predictions of the structures of melittin and polypeptide do not imply that the tertiary structures of larger proteins can be predicted using this knowledge-based potential.

Crambin has been described previously. A single genetic algorithm-based conformational search under the knowledge-based potential produced an extended conformer of crambin (energy −1126 kcal/mol) which was 100 kcal/mol lower in energy than the native structure (energy −1026 kcal/mol) (Fig. 21). The ΔRMS between the predicted and the native structure of crambin is 9.06 Å. Although the tertiary structure of the predicted conformation is wrong, much of the helical secondary structure is predicted correctly. The native conformation of crambin has two α-helices ranging from residues 7 to 19 and from 23 to 30. The predicted conformation has a helices ranging from residues 7–17 and from 22–28. The β-sheet structure of crambin is not predicted as well. Crambin has an antiparallel β-sheet composed of residues 1–4 and 32–35. Residues 1–4 of the predicted crambin conformation are random coil while residues 32–35 are correctly predicted as β-sheet. Since the final predicted conformation of crambin was lower in potential energy than the native conformation, this confirmed that the knowledge-based potential favored the formation of correct local secondary structure but disfavored the formation of packing interactions.

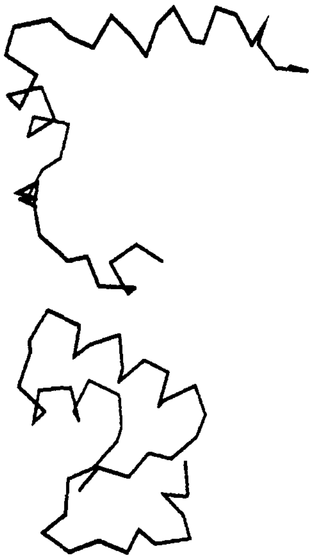

Figure 21 Comparison of native structure and final conformer of crambin produced by genetic algorithm-based conformational search under the knowledge-based potential.

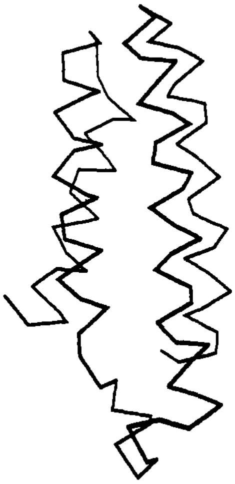

Figure 22 Crystal structure of the cytochrome b_{562} monomer.

Cytochrome b_{562} is a 106 residue 4 helix bundle (Fig. 22) which exists as a dimer. To further test the packing components of the statistical interatomic potential, we performed a genetic algorithm-based conformational search on this protein with the 4 α-helices of the protein constrained to α-helical conformations. 414 dihedral angles determine the conformation of the cytochrome b_{562} monomer. 168 of these variables are strongly constrained in the presence of the helical secondary structure constraint. This makes the secondary structure constrained prediction of the tertiary structure of cytochrome b_{562} a 246-variable optimization problem.

The final conformer produced by a single genetic algorithm-based conformational search confirmed that the knowledge-based potential was incapable of packing a protein into a native-like conformation (Fig. 23). Although it has almost perfect secondary structure, the final conformation is an almost completely extended rod rather than a packed 4 helix bundle. The potential energy of the native conformation of cytochrome b_{562} under the knowledge-based potential is -2721 kcal/mol. The potential energy of the predicted extended non-native conformer is -3800 kcal/mol. The potential energy of the dimer is -6400 kcal/mol or -3400 kcal/mol per monomer. Therefore, the knowledge-based potential favors the formation of extended structures with well-organized secondary structure but mostly handles packing as a repulsive interaction. This makes it unsuitable for tertiary structure prediction in its current form.

CONCLUSIONS

Our genetic algorithm appears to be capable of an order of magnitude improvement in speed over previous conformational search techniques as judged by its performance at the conformational search of [met]enkephalin. However, it is unwise to make general conclusions from one example. In addition, our genetic algorithm's capability to locate non-native conformers of proteins with lower energies than the native structure under our knowledge-based potentials indicates that contrary to previous opinion [15], the

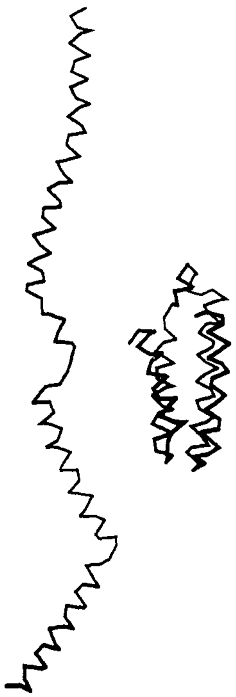

Figure 23 Comparison of native structure and a final conformer produced by a single genetic algorithm-based conformational search of the cytochrome b_{562} monomer under the knowledge-based potential.

computational complexity of conformational search is currently not the major obstacle to conformational search-based protein tertiary structure prediction but rather it is the quality of the potential functions in use. It appears to be trivial to locate non-native structures of proteins under such potential functions via unrestrained conformational search. In addition, recent results using ΔRMS as a potential function for conformational search [62] imply that conformational search techniques such as genetic algorithms are capable of locating the native conformation when given a potential function which unambiguously rates the native conformation as the unique global minimum. In this light, it would seem prudent to evaluate the quality of other potential functions in use via conformational search which appears to be a more rigorous test of potential function quality than inverse protein folding, determining which amino acid sequences are compatible with a pre-specified folding motif [15, 57, 63, 64], and model verification, the detection of errors in crystal structures [65].

Why did the knowledge-based potential function in use here fail at the task of *de novo* protein tertiary structure prediction? Based on the final conformers produced by conformational search of crambin and cytochrome b_{562}, it appears that the conformational space available to these proteins is sufficiently flexible to allow for the generation of conformations which maximize occupation of interaction states with negative energies and minimize occupation of interaction states with positive energies much more efficiently than the native structures of proteins. As long as such a conformer can be generated for a given protein, a potential function will fail when applied to unrestricted conformational search of the protein. This can be illustrated by examining the distribution of nonbond interactions within a 10 Å sphere around each atom in cytochrome b_{562} and comparing it to the same distribution generated from the incorrect extended predicted structure. The nonbond interaction distribution for cytochrome b_{562} is a continuously increasing, albeit somewhat noisy, function of interatomic separation (Fig. 24). In contrast, the nonbond distribution for our final incorret predicted structure for cytochrome b_{562} has a relatively higher fraction of close range contacts and few long range contacts (Fig. 25). Therefore, optimization of this potential function does not reproduce conformers with protein-like nonbond distributions. Why does this occur? The reference state used by Sippl [15] and in this work was the average folded state of proteins. The generation of extended conformations of crambin and of cytochrome b_{562} that are lower in energy than the native conformations of these proteins indicates that this reference state is only valid when comparing one folded conformation to another. It is clearly not valid for the comparison of an unfolded conformation to the folded conformation. The optimum conformations of proteins under this potential function are extended structures with reasonable secondary structure but little or no tertiary structure. What can be done? One way to address this problem is to recalculate the potential function using randomized conformers of the database proteins as the reference state. If this is done, then interactions which occur more frequently in random unfolded conformations than in folded conformations will be given positive, repulsive energies. Conversely, interactions which occur more frequently in folded conformations than in random unfolded conformations will be given negative, attractive energies. When the knowledge-based potential was recalculated relative to the random unfolded reference state, the incorrect final conformations of pancreatic polypeptide, crambin and cytochrome b_{562} are de- stabilized relative to the native conformation. Unfortunately, the 5.90 Å ΔRMS non-native conformation of the melittin monomer remains approximately 150 kcal/mol below the energy of the native conformation, and subsequent genetic algorithm-based conformational search of pancreatic polypeptide, crambin, and cytochrome b_{562} under the updated knowledge-based potential located compact but randomized conformations

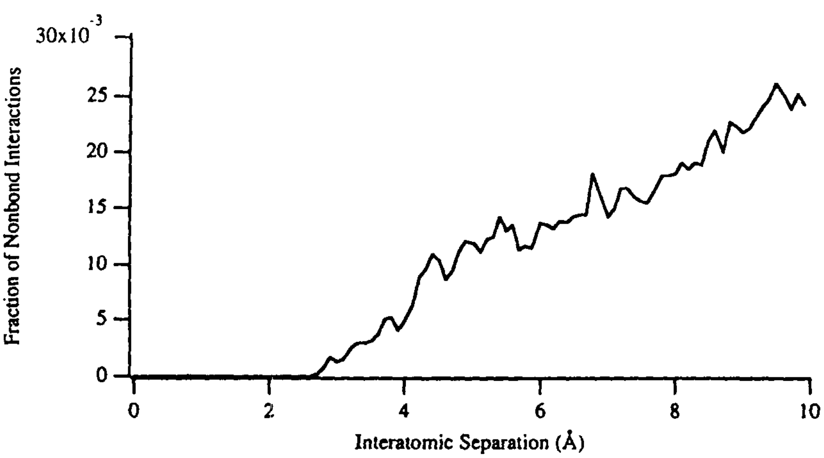

Figure 24 Nonbond interaction distribution for cytochrome b_{562}.

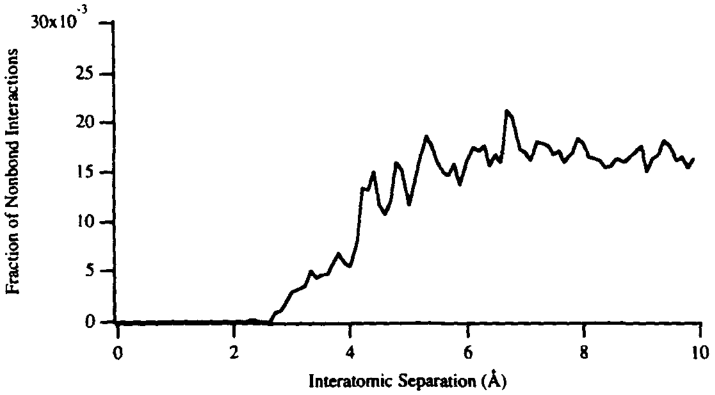

Figure 25 Nonbond interaction distribution for the incorrect predicted conformation of cytochrome b_{562}.

that were lower in energy than the native conformation but had no recognizable secondary structure. This indicates that it is difficult to strike a balance between the energetics of secondary structure and tertiary structure formation and we believe that there is much work to be done before an accurate potential function for tertiary structure prediction can be generated.

Acknowledgments

We would like to thank Drs. James Bowie, David Eisenberg, Greg Farber, and Juliette Lecomte for stimulating conversations, Dr. Davis Eisenberg, The Cornell Theory Center, and The Pittsburgh Supercomputing Center for computer time, Cray Computers and the National Science Foundation for fellowship support, and the Office of Naval Research for supporting this research (N00014-90-3-4002).

References

[1] C. Anfinsen, *The Molecular Basis of Evolution*, John Wiley & Sons, New York, 1959.

[2] G. Fasman, Development of the prediction of protein structure, in *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum Press, New York, 1989.

[3] S. Wilson and W. Cui, Applications of simulated annealing to peptides, *Biopolymers*, **29**, 225–235,

[4] G. Nemethy and H. Scheraga, Theoretical studies of protein conformation by means of energy computations, *FASEB Journal*, **4**(14), 3189–3197 (1990).

[5] F. Anet, Inflection points and chaotic behavior in searching the conformational space of cyclononane, *J. Am. Chem. Soc.*, **112**, 7172–7178 (1990).

[6] S. M. Le Grand and K. M. Merz Jr., The application of the genetic algorithm to the minimization of potential energy functions, *J. Global Opt.*, **3**, 49–66 (1993a).

[7] S. M. Le Grand and K. M. Merz Jr., The application of the genetic algorithm to protein tertiary structure prediction, to be submitted to *J. Mol. Biol.*, (1993b).

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

[9] M. E. Snow, Powerful simulated-annealing algorithm locates global minimum of protein-folding potentials from multiple starting conformations, *J. Comp. Chem.*, **13**, 597–584 (1992).

[10] J. Skolnick and A. Kolinski, Simulations of the folding of a globular protein, *Science*, **250**, 1121–1125 (1990).

[11] S. Weiner, P. Kollman, D. Nguyen and D. Case, An all atom force field for simulations of proteins and nucleic acids, *J. Comp. Chem.*, 7(2), 230–252 (1986).

[12] B. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan and M. Karplus, CHARMM: A program for macromolecular energy, minimization, and dynamics calculations, *J. Comp. Chem.*, **4**, 187–217 (1983).

[13] F. Momany, R. McGuire, A. Burgess, and H. A. Scheraga, Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids, *J. Phys. Chem.*, **79**(22), 2361–2381 (1975).

[14] C. Wilson and S. Doniach, A computer model to dynamically simulate protein folding: studies with crambin, *Proteins*, **6**, 193–209 (1989).

[15] M. J. Sippl,-Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins, *J. Mol. Biol.*, **213**, 859–883 (1990).

[16] G. Crippen, Prediction of protein folding from amino acid sequence over discrete conformational spaces, *Biochemistry*, **30**, 4232–4237 (1991).

[17] D. G. Covell and R. L. Jernigan, Conformations of folded proteins in restricted spaces, *Biochemistry*, **29**, 3287–3294 (1990).

[18] J. Holland, *Adaptation in Natural and Artifical Systems*, University of Michigan Press, Ann Arbor, MI (1975).

[19] D. Goldberg, *Genetic Alogrithms in Search, Optimization, and Machine Learning*, Addison-Wesley, San Mateo, CA (1989).

[20] D. Montana and L. Davis, Training feedforward neural networks using genetic algorithms, in *Proceedings of the Third International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Mateo, CA (1989).

[21] D. Whitley and T. Hanson, The GENITOR algorithm and selective pressure: why rank-based allocation of reproductive trials is best, in *Proceedings of the Third International Conference on Genetic Algorithms*, Morgan Kaufmann, San Mateo, CA (1989a).

[22] D. Whitley, and T. Hanson, Optimizing neural nets using faster, more accurate genetic search, in *Proceedings of the Third International Conference on Genetic Algorithms*, Morgan Kaufmann, San Mateo, CA (1989b).

[23] D. Whitley, T. Starkweather and C. Bogart, Genetic algorithms and neural networks: optimizing connections and connectivity, *Parallel Computing*, **14**, 347–361 (1990a).

[24] D. Whitley, and T. Startweather, GENITOR II: a distributed genetic algorithm, *J. Exp. Theor. Artif. Intell.*, **2**, 189–214 (1990b).

[25] G. Cleveland and S. Smith, Using genetic algorithms to schedule flow shop releases, in *Proceedings of the Third International Conference on Genetic Algorithms*, Morgan Kaufmann, San Mateo, CA (1989).

[26] C. B. Lucasius and G. Kateman, Application of genetic algorithms in chemometrics, in *Proceedings of the Third International Conference on Genetic Algorithms*, Morgan Kaufmann, San Mateo, CA (1989).

[27] C. B. Lucasius, M. J. J. Blommers, L. M. C. Buydens, and G. Kateman, A genetic algorithm for conformational analysis of DNA, in *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York (1990).

[28] M. S. Friedrichs, and P. G. Wolynes, Genetic algorithms for model bimolecular optimization problems, unpublished, 1989.

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

[30] S.M. Le Grand and K.M. Merz Jr., doctoral thesis (1993c).

[31] R. Unger and J. Moult, Genetic algorithms for protein folding simulations, *J. Mol. Biol.*, **231**, 75–81 (1993).

[32] J. Grefenstette and J. Baker, How genetic algorithms work: a critical look at implicit parallelism, in *Proceedings of the Third International Conference on Genetic Algorithms*, Morgan Kaufmann, San Mateo, CA (1989).

[33] L. Booker, Improving search in genetic algorithms, in *Genetic Algorithms and Simulated Annealing*, Morgan Kaufmann, San Mateo, CA (1987).

[34] J. Schaffer and A. Morishima, An adaptive crossover distribution mechanism for genetic algorithms, in *Genetic Algorithms and their Applications: Proceedings of the Second International Conference on Genetic Algorithms*, Lawrence Erlbaum, Hillsdale, NJ (1987).

[35] D. Sirag and P. Weisser, Toward a unified thermodynamic genetic operator, in *Genetic Algorithms and their Applications: Proceedings of the Second International Conference on Genetic Algorithms*, Lawrence Erlbaum, Hillsdale, NJ (1987).

[36] Y. Davidor, Analogous crossover, in *Proceedings of the Third International Conference on Genetic Algorithms*, Morgan Kaufmann, San Mateo, CA (1989).

[37] T. Fogarty, Varying the probability of mutation in the genetic algorithm, in *Proceedings of the Third International Conference on Genetic Algorithms*, Morgan Kaufmann, San Mateo, CA (1989).

[38] P. Wayner, Genetic algorithms, *Byte*, **16**, 361–364 (1991).

[39] C. Walbridge, Genetic algorithms: What computers can learn from Darwin, *Technology Review*, **92**, 46–53 (1989).

[40] N. Radcliffe and G. Wilson, Natural solutions give their best, *New Scientist*, **126**, 47–50 (1990).

[41] S. Marqusee, V. H. Robbins and R. L. Baldwin, Unusually stable helix formation in short alanine-based peptides, *PNAS*, **86**, 5286–5290 (1989).

[42] D. R. Ripoli, M. Vasquez and H. A. Scheraga, The electrostatically driven Monte Carlo method: Application to conformational analysis of decaglycine, *Biopolymers*, **31**, 319–330 (1991).

[43] A. Chakrabarty, J. A. Schellman and R. L. Baldwin, Large differences in the helix propensities of alanine and glycine, *Nature*, **351**, 586–588 (1991).

[44] Z. Li and H. Scheraga, Monte Carlo recursion evaluation of free energy, *J. Phys. Chem.*, **92**, 2633–2636 (1987).

[45] E. Purisima and H. Scheraga, An approach to the multiple-minima problem in protein folding by relaxing dimensionality: Tests on enkephalin, *J. Mol. Biol.*, **196**, 697–709 (1987).

[46] S. Vajda and C. Delisi, Determining minimum energy conformations of polypeptides by dynamic programming, *Biopolymers*, **29**, 1755–1772 (1990).

[47] J. K. Shin and M. S. Jhon, High directional Monte Carlo procedure coupled with the temperature heating and annealing as a method to obtain the global energy minimum structure of polypeptides and proteins, *Biopolymers*, **31**, 177–185 (1991).

[48] Y. Deng, J. Glimm and D. H. Sharp, Perspectives on parallel computing, *Daedalus*, **21**(1), 31–52 (1992).

[49] A. Nayeem, J. Vila and H. A. Scheraga, A comparative study of the simulated annealing and Monte-Carlo-with-Minimization approaches to the minimum-energy structures of polypeptides: [met]enkephalin, *J. Comp. Chem.*, **5**, 594–605 (1991).

[50] M. Teeter, Atomic resolution (0.83 Å) crystal structure of the hydrophobic protein crambin at 130 K, *J. Mol. Biol.*, **230**, 292–298 (1993).

[51] C. Brooks III, M. Karplus and B. Montgomery Pettitt, *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodyamics*, Wiley Interscience, New York (1988).

[52] R. Baldwin and D. Eisenberg, Protein stability, in *Protein Engineering*, Alan R. Liss, Inc., NY (1987).

[53] R. Baldwin, How does protein folding get started? *Trends Biochem. Sci.*, **14**, 291–294 (1989).

[54] N. Khechinashvili, Thermodynamic properties of globular proteins and the principle of stabilization of their native structure, *Biochimica et Biophysica Acta*, **1040**, 346–354 (1990).

[55] D. Eisenberg and A. McLachlan, Solvation energy in protein folding and binding, *Nature*, **319**, 199–203 (1986).

[56] S. M. Le Grand and K. M. Merz Jr., unpublished results (1993d).

[57] D. T. Jones, W. R. Taylor, J. M. Thornton, A new approach to protein fold recognition, *Nature*, **358**, 86–89 (1992).

[58] T. C. Terwilliger, L. Weissman and D. Eisenberg, The structure of melittin in the form I crystals and its implications for melittin's lytic and surface activities, *Biophys. J.*, **37**, 353–361 (1982).

[59] T. Ikura, N. Go and F. Inagaki, Refined structure of melittin bound to perdeuterated dodecylphosphocholine micelles as studied by 2D-NMR and distance geometry calculation, *Proteins*, **9**, 81–89 (1991).

[60] I. Glover, I. Haneef, J. Pitts, S. Wood, D. Moss, I. Tickle and T. Blundell, Conformational flexibility in a small globular hormone: X-ray analysis of avian pancreatic polypeptide at 0.98 Å resolution, *Biopolymers*, **22**, 293–304 (1983).

[61] J. E. Pitts, I. D. Glover, W. Strassburger and A. Wollmer, unpublished results (1982).

[62] J. Bowie and D. Eisenberg, An evolutionary approach to folding proteins from sequence information: application to small α-helical proteins, *PNAS* in press.

[63] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari and M. J. Sippl, Identification of native protein folds amongst a large number of incorrect models: The calculation of low energy conformations from potentials of mean force, *J. Mol. Biol.*, **216**, 167–180 (1990).

[64] M. J. Sippl and S. Weitckus, Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations, *Proteins*, **13**, 258–271 (1992).

[65] M. J. Sippl, Recognition of errors in three-dimensional structures of proteins, *Proteins*, **17**(4), 355–362 (1993).