A Parallel Genetic Algorithm for Polypeptide Three Dimensional Structure Prediction. A Transputer Implementation

Carlos Adriel Del Carpio

Laboratory for Informatics & AI in Molecular and Biological Sciences, Department of Knowledge-based Information Engineering, Toyohashi University of Technology, Tempaku, Toyohashi 441, Japan

Received September 6, 1995⊗

We propose a hybrid genetic algorithm to carry out a mapping of the conformational space of polypeptides, and at the same time we suggest a fitness function that incorporates quantitatively many factors that experimentally are known to influence the process of protein folding. Some of these factors are the hydrophobicity of the molecule, the disulfide bond among cysteine sulfur atoms and the compactness of the molecule. The steric energy of the molecule is calculated using molecular mechanics force fields. To account for the packing of the side chains of the polypeptide, the conventional GA is hybridized, endowing it with a local optimization function to perform the task. Moreover, due to the huge number of conformers the system processes, and because of the nature of the fitness function it manipulates, it has been parallelized, and its implementation on a network of transputers is also discussed.

1. INTRODUCTION

Since Anfinsen[1] demonstrated in 1959, that the three dimensional (3D) structure of a protein is uniquely determined by the composition of its primary structure, many methods have been developed in order to predict three dimensional structure of proteins from their primary sequence.

Among these methods, we can distinguish mainly two categories: (i) in the first category are methods which start with the assignment of secondary structure tendencies to the sequence and build the 3D structure from this knowledge, and (ii) methods which optimize the geometry of the polypeptide by minimization of a potential energy function that corresponds to the thermodynamical state of the polypeptide.

The first type of approach relies on the prediction of the secondary structure by extracting information from crystallographic data bases. Although the accuracy of the predictions has been increased steadily in recent years, the prediction of the tertiary structure is still hampered by the inaccuracy of these approaches.[2]

The second type of approach assumes that the conformation of a protein corresponds to the global minimum of the potential energy surface. The fact that there is a very large number of local minima makes the simulation process very difficult, and the unambiguous determination of the conformation of the macromolecule becomes a challenging task.

Because of the difficulties in exploring the whole conformational space of a polypeptide, the structure predicted by this type of approaches depends on the initial conformation of the molecule, i.e., the starting point on the potential energy surface.

The implementation of this kind of approach requires (i) a potential energy function which takes into account all the factors involved in the folding process and (ii) a conformational search method capable of exhaustively exploring the conformational space in an efficient manner.

In recent years several conformational search methods were developed for locating global minima of multivariable functions and within them for searching the protein conformational space in particular.

A robust conformational search method that seems appropriate for this task is that based on the use of the genetic algorithm (GA). Since its first usage, the method has been applied at solving many problems in several fields of science and technology, such as pattern recognition, machine learning, robot control, and neural network optimization.[3]

In fact, some attempts have been carried out to predict characteristics of polypeptides using conventional GAs, and within them 3D structural features of proteins. For example, Le Grand et al.[4] used a variant of a GA for minimization of AMBER potential function which was used in their work as the fitness function in predicting tertiary structures of peptides from primary sequences, Dandekar et al.[5] uses GAs for zinc finger sequence identifications, while Lucasius et al.[3] use them to perform a conformational analysis of DNA.

In the present work we propose a hybrid GA for the prediction of the three dimensional structure of proteins. Our algorithm uses a newly developed potential energy function, which is a sum of terms accounting for the internal forces between atoms of the protein (i.e., the molecular mechanics steric energy term), and terms specific to the folding process of the polypeptide (e.g., disulfide bonding term, compactness of the molecule, the solvent effect). The latter are terms derived by fitting computed 3D structures of proteins with 3D structures of proteins extracted from the crystallographic data recorded in the Brookhaven Protein Data Bank (PDB). Our solution to the huge amount of computational time required in (i) the calculation of the terms of the potential energy function introduced here and (ii) the mapping of the whole conformational space is a parallel hybrid GA which we have developed for a message-passage architecture of a network of processors, and its implementation on a network of transputers is presented here.

⊗ Abstract published in *Advance ACS Abstracts,* February 15, 1996.

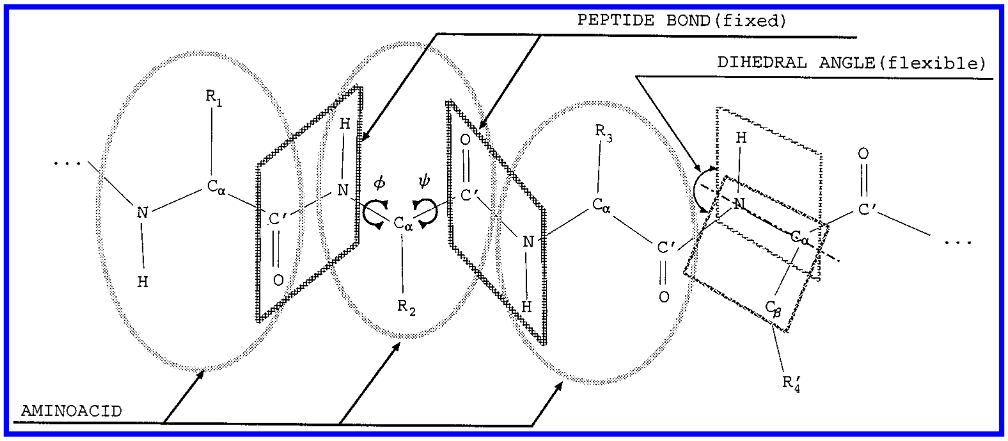

Figure 1. The polypeptide backbone and the ϕ and ψ angles.

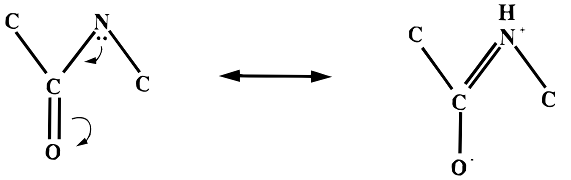

Figure 2. The trans conformation of the amide bond.

2. THE PROTEIN FOLDING PROBLEM AND THE GENETIC ALGORITHM

2.1. Characteristics of Protein Tertiary Structure. Proteins, together with the RNA and DNA, constitute the central entities of the mechanism of life at the molecular level. While DNA and RNA are in charge of supplying and transmitting genetic information, proteins are accomplishing functions that range from signal transmission, ligand receptor, and immunological functions to functions involving the regulation of the expression and transmission of genetic information by binding to DNA. They also play an important role as catalyzing enzymes for chemical reactions in biosystems.

Studies on proteins have as their main goal the elucidation of the structure and function of these polypeptides. In fact the relationship between the structure and the function of a protein is so important that studies on the latter are meaningless without a thoroughly understanding of structural details of the biomacromolecule. However, the determination of the tertiary structure of a protein is still a challenging problem in protein engineering. The difficulty stems from the large number of the degrees of freedom of the rotations around the peptide bonds of the molecule.

The protein is a polymer constituted by a long sequence of 20 different kind of α-amino acids. These amino acids, linked by peptide bonds and arranged in a head-to-tail fashion, form a polymer, whose tertiary structure characteristics depend on the particular conformation of the chain (i.e., rotation angles around the single bonds in the polypeptide chain as shown in Figure 1). However, the fact that the amide bond has a planar trans conformation reduces the number of dihedral angles which affect the folding of the protein to essentially two. The planar conformation of the peptide bond is a consequence of the resonance form that imparts double-bond character to the amide group as shown in Figure 2. As a result of this fact, the amide bond is shortened by about 0.1 Å, and the amide group adopts a planar trans which makes the distances between all α-carbons to be fixed at 3.8 Å. Consequently, the conformation of the biopolymer chain is determined by the rotations about the $N-C^{\alpha}$ and $C^{\alpha}-C'$ bonds.

2.2. The Protein Folding Process. In as much as stated before, the prediction of the tertiary structure of proteins is bound to the determination of the dihedral angles corresponding to $N-C^{\alpha}$ and $C^{\alpha}-C'$ bonds. However, folding of a protein occurs by coiling portions of the molecule into helice or pleated sheet structures. Other regions are characterized by loops, bends, or kinks. These protein segments are joined together by various kinds of bonding and nonbonding interactions between the amino acid residues, such as hydrogen bonds, van der Waals interactions, and bridges between sulfur atoms in cysteine amino acids. Moreover, the findings of Weissman and Kim[6] that nonnative interactions stabilize substantially some folding intermediates, complicate further the understanding of the folding process and makes the prediction of the conformation of proteins from its primary structure even a more difficult task. However, as the same authors pointed out, the fact that there is a scarce population of such nonnative intermediates and that the same interactions that stabilize the final folded structure also guide the protein in attaining its final structure leaves room for further efforts in elucidating the 3D structures of polypeptides.

From this premise, we present in this study a potential energy function that, besides the terms expressing the steric potential energy of the molecule, includes factors thought to affect largely the protein folding process. Among these factors we have specially considered the hydrogen bond and the disulfide bridge, a factor that accounts for the compactness of the molecule and an empirical factor expressing the solvent effect due to hydrophobic interactions between regions of the protein. These factors are calibrated by fitting computed protein structures to protein structures extracted from the Brookhaven Protein Data Bank. The details of the calibration of this potential energy function will be given elsewhere.[7]

2.3. The Genetic Algorithm. A genetic algorithm is an optimization technique that imitates some of the processes observed in natural evolution.[8] In a way similar to biological life, these algorithms solve the problem of finding better "chromosomes" by manipulating the material in chromosomes blindly. The chromosomes manipulated in GAs are usually strings of binary digits—1's and 0's, although this representation may vary from one problem to another. The chromosomes encoded in this way, know nothing about the type of the problem being solved. A mechanism to evaluate the goodness of each chromosome is the only information they are supplied with. With operations that mimic natural reproduction, selection, crossover, and mutation, these

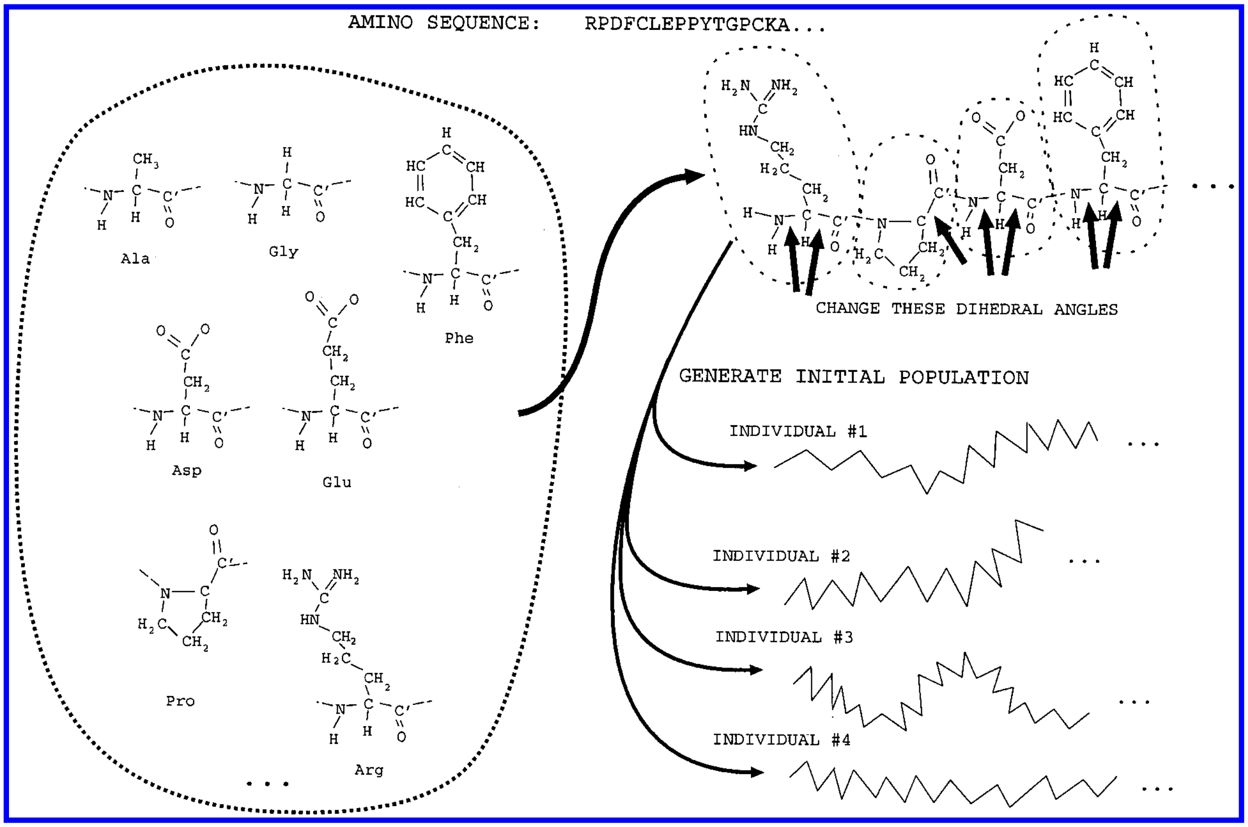
AMINO SEQUENCE: RPDFCLEPPYTGPCKA...
Ala
Gly
Phe
Asp
Glu
Pro
Arg
...
CHANGE THESE DIHEDRAL ANGLES
GENERATE INITIAL POPULATION
INDIVIDUAL #1
INDIVIDUAL #2
INDIVIDUAL #3
INDIVIDUAL #4

Figure 3. Initial population of chromosomes (conformers for the primary sequence).

algorithms are able to display complicated behavior and solve extremely difficult problems. A conventional GA can be described by the following steps:

1. Generate a random initial population of chromosomes.
2. Examine each individual in the population.
3. Select the best chromosomes of the current population, crossover, and produce mutation to create new chromosomes.
4. Build the next population from the new chromosomes and eliminate unfitted chromosomes of the earlier generation.
5. Examine the new chromosomes and construct the next generation of chromosomes.
6. Stop if the halt condition is achieved otherwise go to step three.

A successful run of this procedure will finish with a population of chromosomes whose fitness scores are improved as compared to those of the set of random chromosomes of the initial population. The probability is very high that the solution to a particular problem is represented by the best evolved chromosome within the population.

The use of the algorithm in the prediction of protein tertiary structure is straightforward, since a chromosome would represent a conformer expressed as a set of dihedral angles that encodes the conformation of the backbone of the molecule.

3. REPRESENTATION OF CHROMOSOMES AND STEPS IN THE GENETIC ALGORITHM

3.1. Chromosome Representation. The straightforward coding of the chromosomes as sets of dihedral angles, and the robustness of the GA as a searching engine for the huge conformational space of the biomolecule are two of the advantages of the method over other searching-optimization schemes.

In the conventional GA, the initial population of chromosomes is created by random generation of strings of 0's and 1's. In the present work, the initial population of chromosomes, which represent protein conformers, is created by generating strands of linked mono acid structures following the order of the given primary sequence (Figure 3). Each amino acid structure is an energetically optimized structure extracted from a file containing the 20 optimized amino acid molecules. The diversity of the initial population is achieved by rotating and translating segments of the strands at the sites of the ϕ and ψ dihedral angles.

In theory, a chain molecule can adopt an essentially infinite variety of backbone conformations, each corresponding to a unique set of values for the various backbone rotation angles. However, many of these hypothetical conformations can be excluded due to unfavorable steric overlaps. Ramachandran et al.[9] performed several studies on the sterically allowed values of ϕ_i and ψ_i and proposed the steric contour diagrams which are maps in ψ and ϕ. Regions in this map are then displayed to inscribe the ϕ and ψ coordinates in which no unfavorable steric contact occur. We have also carried out such an analysis, but for each kind of amino acid residue, using the data for 100 different proteins recorded in the Brookhaven Protein Data Bank. Figure 4 shows the steric contour map for alanine obtained in this way.

The contour maps show regions of energetically allowed and forbidden pairs of dihedral angles ϕ and ψ. After clustering all points (ϕ, ψ) belonging to a certain kind of amino acid residue, the convex hull surrounding each cluster is computed. The convex hulls enclosing pairs of angles in the plane are the regions of allowed dihedral angles ϕ and ψ (Figure 5). A class can be represented by the coordinates of the vertices of the simplest convex hull and the centroid of each class. These sets of values can be easily manipulated to generate pairs of sterically allowed torsional angles. Thus the computation reduces to the calculation of the segment lying between the points of intersection of two opposite sides

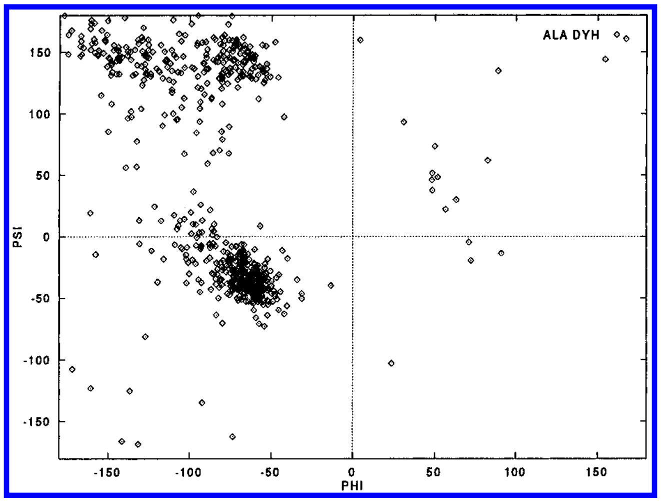

Figure 4. Steric overlap contour map for alanine.

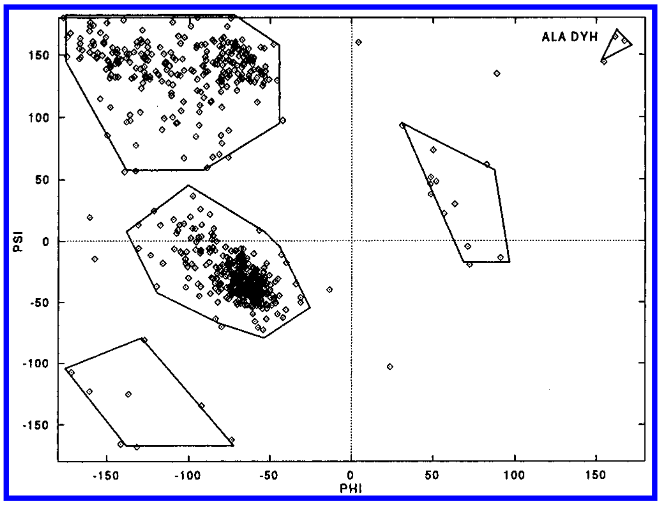

Figure 5. Clusters of sterically allowed *f* and *Y* angles for alanine.

of the hull and the line perpendicular to the axis corresponding to any of the two dihedral angles. The dihedral angles used to generate the populations of conformers are selected randomly from the interval of angles represented by the line segment. When more than one class of sterically allowed angles is intersected, the algorithm decides randomly on the class used in the computations. The selection of these torsional angles is performed (i) at the initial step of the algorithm, to generate the initial population, and (ii) when a random mutation is performed.

3.2. Reproduction. After the evaluation of the fitness of each chromosome from the initial population, the GA selects those "individuals" or chromosomes (conformers) that are best fit, i.e., have the best score among the chromosomes in the entire population. This operation mimics natural selection, i.e., the "individuals" which best fit to the environment survive. In our GA, the selected individuals are reproduced into the next generation or recombine among them to produce better fit offspring. The implementation of this operation in our GA is similar to the one found in the original elementary GA algorithm designed by Goldberg.[8] This is a linear search through a roulette wheel with slots weighted in proportion to the individual fitness values (Figure 6).

3.3. Crossover. This operation is implemented here by mating two members of the pool of newly reproduced individuals selected at random. After the selection of the mating partners the crossover point is determined randomly. This crossover point is between two torsional dihedral angles. The mating operation produces two "child" conformers which inherit the coordinates of the two parent conformers from the first atom to the atom at the crossover point. The coordinates of the atom from the crossover point to the last one in each offspring are those belonging to the mating partner, translated to the position of the atom before crossover, and rotated by an angle equal to that of the mating parent, respectively. This operation is illustrated in Figure 7.

3.4. Mutation. This operation, abound in the conventional GA is also implemented in our algorithm. Mutation here is performed by rotating one of the angles of the main chain of the polypeptide. Whenever the probability of mutation of one "allele" is higher than a present threshold, then the operation is performed only in one of the ϕ or ψ angles of the backbone, i.e., the angle between the N$-$C$^{\alpha}$ or C$^{\alpha}-$C$'$ atoms of the polypeptide backbone. An angle is selected within the region of the steric energy contour map which is in correspondence with that of the nonmutated angle, that is, an angle that is not sterically and energetically forbidden and that corresponds to the value of the nonmutated angle for that particular amino acid. In other words a value

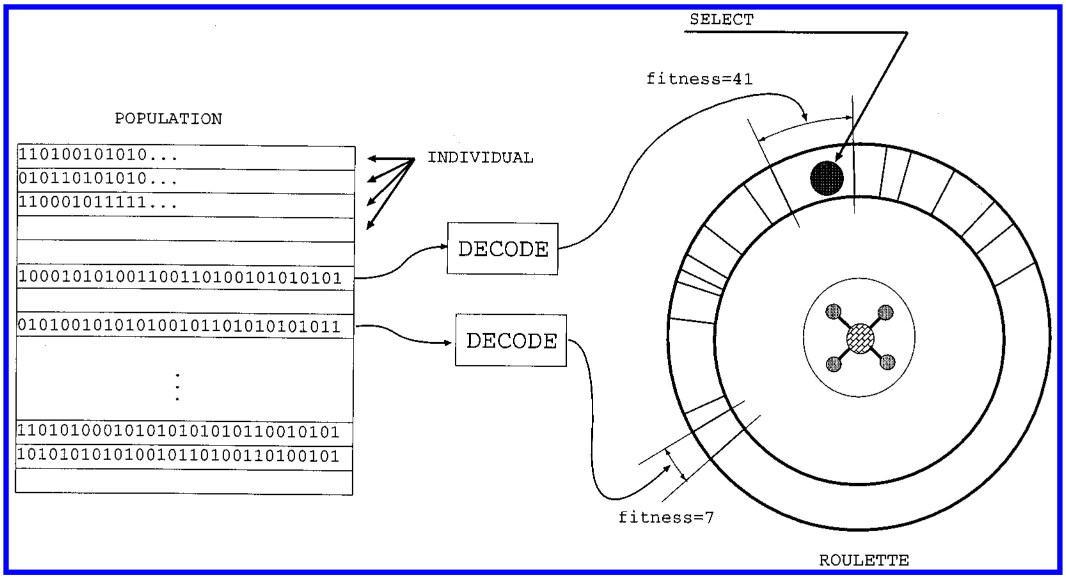

Figure 6. Selection and reproduction in the genetic algorithm.

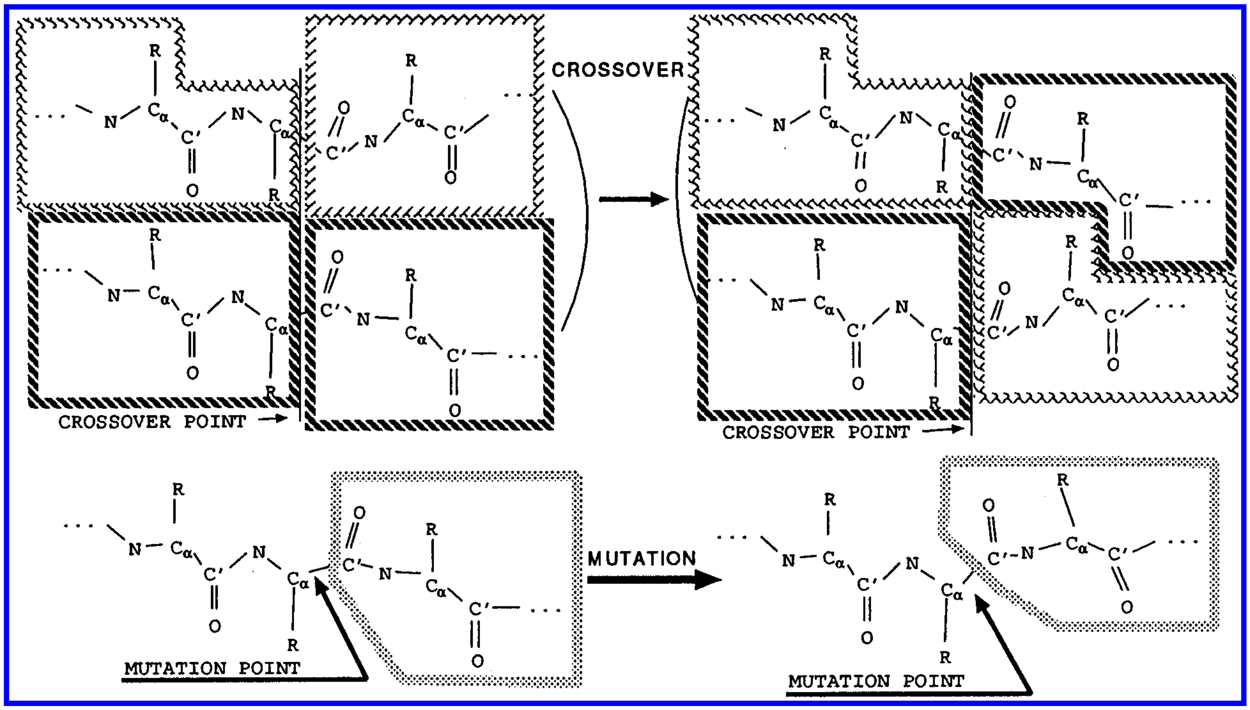

Figure 7. Crossover in the genetic algorithm.

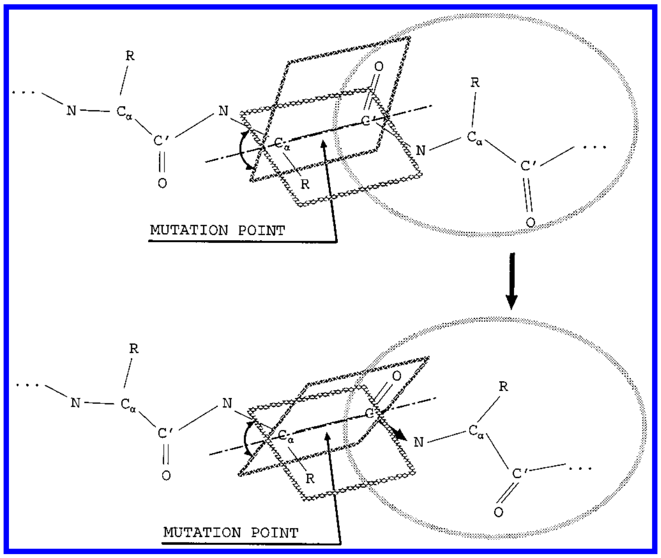

Figure 8. Mutation in the genetic algorithm.

for ψ is selected if the mutated angle is ϕ, or vice versa if the mutated angle is ψ. Figure 8 illustrates the mutation operation implemented in our GA. Here, all cartesian coordinates of atoms lying at the right of the mutation site are rotated by an angle equal to the selected value minus that of the original individual.

This three operations are repeated for a number of generations until the halting condition is achieved.

4. THE FITNESS FUNCTION AND THE HYBRID GENETIC ALGORITHM

Studies on the tertiary structure of proteins suggest that various and complex factors determine protein conformation. Consequently, in addition to the dependency of the three dimensional structure on the internal potential energy of the molecule, there is the necessity to consider also other constraints peculiar to the folding process of the biomolecule. Here we take into account a series of factors that influence the formation of the native protein 3D structure and try to quantify them based on a detailed analysis of crystallographic data of proteins as well as theoretical considerations from the mechanics of biomolecules. The details of the construction, calibration, and evaluation of the fitness function are described elsewhere.[7] Here, we describe briefly some of these factors which are found to be the most relevant for the evolutionary process that we propose in the present work.

4.1. The Potential Energy Term. A. J. Hopfinger presented and described in detail the terms of the potential energy function utilized in conformational studies of polypeptides.[10] Molecular force fields provide a suitable framework to evaluate the steric energy and determine the molecular three dimensional structure. Several programs for calculating the steric energy of a molecule, based on these approaches, have been already developed. Among them, the MM2 and MM3 force fields developed by N. L. Allinger[11] are the most used in conformation analysis.

We enumerate some of the terms of the steric energy function that are used in our GA. The parameters needed for the calculation of these terms can be read into the program from a file, but the default values are, however, those belonging to the MM2 force field.[12]

4.1.1. Van der Waals Potential. Interatomic attractions and repulsions are described by a Lennard-Jones type function

$$E_{\mathrm{vdw}} = \sum_{i=1}^{n}\sum_{j>i}^{n}\left[\left(\frac{A_{ij}}{r_{ij}}\right)^{12} - \left(\frac{B_{ij}}{r_{ij}}\right)^{6}\right] \tag{1}$$

Here, A_{ij} and B_{ij} are empirical constants proper to each pair of atoms, while r_{ij} is the interatomic distance, and n is the number of atoms in the molecule.

4.1.2. Electrostatic Potential. Electrostatic repulsions and attractions are expressed by the Coulomb law

$$E_{\mathrm{electrostatic}} = \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{q_i q_j}{\epsilon r_{ij}} \tag{2}$$

Here, q_i and q_j are the atomic partial electric charges, r_{ij} is the atomic distance, and ϵ the dielectric constant of the medium.

4.1.3. Torsional Potential. This potential energy is computed by a Fourier series of the form

$$E_{\text{torsional}} = \sum_{i=1}^{n_\tau}\sum_{j=1}^{3} V_j^{(i)}(1 + (-1)^j \cos(j\omega)) \qquad (3)$$

Here, V_j is an empirical constant and ω is the angle of rotation. As stated above, in the present work only the angles ϕ and ψ of the main chain of the polypeptide are rotated by the GA.

The fitness function presented in this work operates in the torsional space, i.e., Fitness $= f(\psi, \varphi)$, and, as stated above, it involves terms, such as the steric energy of the molecule, and terms that represent diverse factors influencing the protein folding process. The steric energy term is a measure of the internal force field of the molecule, which is composed mainly of the stretching, bending, and torsional energies. A change in a torsional angle will also induce small changes in the bond length and bond angles. Removing stretching and bending terms from the fitness function would lead to an incomplete force field which would not yield the correct optimized geometry. Furthermore, as discussed later, a process of local minimization is performed to calculate the optimal packing of the side chains after the main chain conformation has been settled by the operations of GA. Hence, the torsional energies are an indispensable factor for the determination of the conformation that side chains may adopt for a determined conformer in the population of main chain conformers represented as chromosomes in the present algorithm.

4.1.4. Bond Stretching Potential. This potential is computed by the following expression

$$E_r = \sum_{i=1}^{n_b}\frac{1}{2}k_r(r_i - r_{0_i})^2 \qquad (4)$$

Here, k_r is a constant that depends on the bond type and r_0 is the equilibrium bond length.

4.1.5. Angle Bending Potential. This potential energy is calculated by means of the following expression

$$E_\theta = \sum_{i=1}^{n_\theta}\frac{1}{2}k_\theta(\theta_i - \theta_{0_i})^2 \qquad (5)$$

where, k_θ is a constant particular to the three atoms forming the angle θ, and θ_0 is the equilibrium angle.

4.1.6. Total Potential Energy. The total potential energy is computed as the sum of the terms described above

$$E_{\text{total}} = E_{\text{vdw}} + E_{\text{electrostatic}} + E_{\text{torsional}} + E_r + E_\theta \qquad (6)$$

4.2. The Hydrogen Bond Energy Term. It is well-known that among the major factors in deciding the tertiary structures of proteins the hydrogen bond plays a critical role. Hence, we have taken into account this fact and have considered the hydrogen bond potential as an independent term from the rest of the potential function. The hydrogen bond energy comes mostly from the electrostatic interactions; however, the directionality of the bond is also considered. The hydrogen bond potential function is computed as

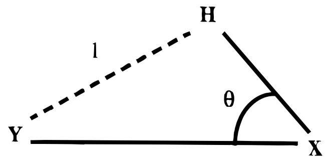

Figure 9. The dependence on directionality of the hydrogen bond.

$$E_{\text{HB}} = \sum_{i=1}^{n_{\text{H}}}\frac{C_{\text{HB}}}{\epsilon}\left[Ae^{-12R_i/r_{\text{HB}}} - B\left(\frac{r_{\text{HB}}}{R_i}\right)^6\left(\frac{l_i}{l_0}\right)\cos\theta_i\right] \qquad (7)$$

where C_{HB} is the hydrogen bond energy parameter, and A and B are empirical constants, r_{HB} is the equilibrium hydrogen bond distance, R is the distance between the donor and acceptor atoms, l is the length of the X−H bond, and l_0 is the natural bond length of bound X−H. θ is the angle between donor and acceptor atoms as shown in Figure 9. Equation 7 is similar to the one used in MM3 force field, and the parameters used in the calculations were taken from the MM3 Reference Manual.[13] The hydrogen bond energy term is taken as an independent term from the rest of the steric energy in the present work due to its special importance in the process of protein folding which is conveyed later in a weighting scheme proposed here for separating factors influencing the folding process.

4.3. The Compactness. Although the initial population of individuals are generated randomly the individuals tend to be less dense than the native structure. The tendency is also observed along the evolutionary process, as there are individuals in later generations where the atoms are sometimes scattered, i.e., they do not show the bulkiness characteristic to globular proteins. We propose thus a factor in the potential energy function to account for the bulkiness or compactness of the polypeptide conformer. This factor is computed as the sum over all pair of atoms within the molecule of the inverse of the interatomic distances between two atoms. This factor is defined as

$$C = \sum_{i=1}^{n-1}\sum_{j>i}^{n}\frac{1}{r_{ij}} \qquad (8)$$

The minimal interatomic distance allowed is that of a bond length. Smaller distances mean a reduction in the compactness score of the conformer.

4.4. The Disulfide Bond. Disulfide bonds also play an important role in the process of protein folding when cystine amino acids are present in the primary sequence. A native protein structure forms only one particular arrangement of disulfide linkages. As proteins fold into their most stable conformation, the disulfide linkages snap in to reinforce this particular configuration. Thus, a particular term to score conformers that enhance disulfide bond formation is introduced here.

This terms scores higher conformers where pairs of half cysteines are at shorter distances

$$S = \sum_{i=1}^{N_s-1}\sum_{j>i}^{N_s}\frac{1}{r_{ij}} \qquad (9)$$

Here the minimal distance considered is that of a S−S bond. Distances smaller to this value are scored negatively. r_{ij} is the distance between two sulfur atoms in the protein, and N_s is the number of half cysteine amino acids.

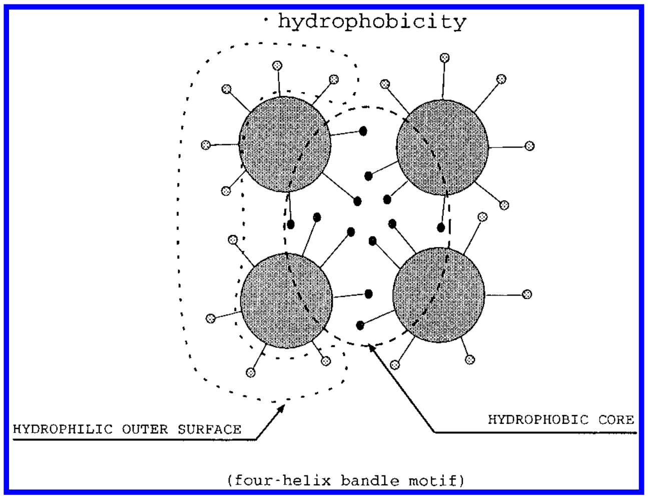

Figure 10. Hydrophobic cores and solvent accessible surfaces in proteins.

4.4. The Solvent Effect. The solvent effect in the formation of the tertiary structure of polypeptides is also critical as is indicated by the high sensitivity of the tertiary structure of the protein to interaction with solvents in denaturation studies. The changes in the shape of the protein, resulting from an unfolding or refolding process, are caused by the transfer of free energy between protein and solvent.

Several studies to understand the interaction of solvent with polypeptide have been carried out in recent years; however, as it happens with the other factors, it is extremely difficult to determine the contribution from any single force to the overall shape of the macromolecule.

Here, we propose a rather empirical function to take into account the solvent effect from a macroscopic point of view.

The different degrees of hydrophobicity and hydrophilicity of the 20 amino acids constituting proteins have them concentrate in regions of relative high hydrophilicity and relatively high hydrophobicity according to their hydrophobic tendencies. Hydrophobic amino acids will then form a hydrophobic core within the protein, while hydrophilic amino acids will escape to zones of easy accessibility for the solvent Figure 10.

We propose here a function that takes into account these factors as well as the surface tension of each atom constituting the amino acids. The function is based on the calculation of the solvent accessible surface area, and the tension parameters are those proposed by Cramer et al.[14]

Analytical calculation of the solvent accessible surface area (*ASA*) for each atom composing the protein is performing using Richmond's algorithm.[15] The surface tension is calculated by the expression

$$\mathrm{St}_i = \sum_{i=1}^{n} (\sigma_k)_i ASA_i \qquad (10)$$

where St_i is the tension surface of the *i*th atom, $\mathrm{ASA_i}$ is its solvent accessible surface area, and σ_k an empirical parameter for atom *i* when its type is *k*.

The energy contribution due to the hydrophobic effect is computed by weighting the total St for an amino acid, with the hydrophobic coefficient. Thus, conformers having a large surface area for hydrophilic amino acids will score higher than hydrophobic amino acids having large solvent accessible surface areas. Similarly, conformers which have hydrophobic amino acids with small solvent accessible surface areas will score higher than their hydrophilic counterpart having approximately equal solvent accessible surface areas. This is computed as

$$\mathrm{Hp} = \sum_{i=1}^{n_{\mathrm{ac}}} \mathrm{Hh}_i (\sum_{j=1}^{n_i} \mathrm{St}_j) \qquad (11)$$

Here, Hp is the total hydrophobic score for the conformer, Hh_i is the hydrophobic coefficient for the *i*th amino acid in accord to hydrophobicity tables. $\mathrm{St_j}$ is the surface tension for the *j*th atom of the amino acid, while n_{ac} is the number of amino acids in the molecule and n_i is the number of atoms for a particular amino acid. Further details of the function for hydrophobic expression are described elsewhere.[7]

4.5. The Overall Fitness Function. The terms described above are calculated for every conformer from the population of "individuals" at every generation obtained with the evolutionary algorithm.

The effect of each factor on the formation of the tertiary structure of the protein is pondered by a normalization scheme resulting in the following expression

$$F_{\mathrm{tt}} = f_e E'_{\mathrm{total}} + f_{\mathrm{HB}} E'_{\mathrm{HB}} + f_c C' + f_s S' + f_{hp} \mathrm{Hp}' \qquad (12)$$

where F_{tt} is the fitness value for an individual from the population of conformers. The *f* factors are the experimental weights standing for the relevance of each factor in the formation of the tertiary structure of the molecule. The primes stand for normalized factor. The normalization of each factor is performed computing the minimal and average values of that factor for the population in a certain generation according to the following equation

$$X' = \frac{X - X_{\mathrm{min}}}{X_{\mathrm{av}}} \qquad (13)$$

where *X* is any of the five factors of the tertiary structure of the protein. X_{min} and X_{av} are the minimum and average values for the generation in course, respectively.

4.6. A Hybrid Genetic Algorithm. Many methods have been proposed to predict the way of packing of the side chains in proteins. For example C. Lee et al.[16] applied the simulated annealing method to optimize the side packing interaction and predict the side chain optimal conformation. Dickett et al.[17] performed an analysis of the configurational entropy of side chains, and, based on this, they suggested sites and configurations for side chains in proteins. Other studies include packing the side chains into bulky ball shaped structures.

In our work, after a GA operation is performed, the perturbation caused by rotation or translation of a part of the molecule is removed by minimizing the steric energy of the structure. This minimization is performed by the Newton−Raphson second derivative method.[18] All the atoms are allowed to move except those of the main chain (Figure 11). This minimization leads to a local energy minima. The protein conformers constituting the population at a certain generation are refined both energetically and structurally by this local minimization (Figure 12). This procedure makes the search for the global minima a much easier task, because conformers at a local energy minimum have parts of the structure already minimized, and by recombination with other conformers of the same type they transmit to the "children" the already optimized sequences.

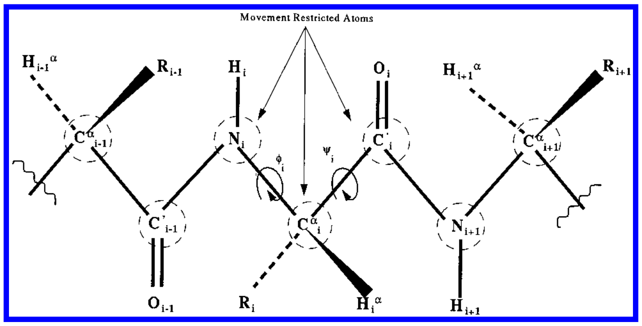

Figure 11. Movement restricted atoms during the local minimization process.

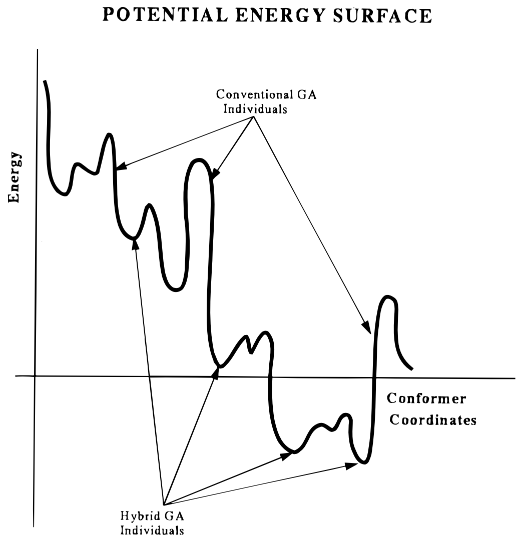

Figure 12. The hybrid genetic algorithm and the individuals' steric energies.

Thus, the hybrid GA is able to locate low energy protein conformers in shorter computational times, bringing down the number of population generations required for the full exploration of the conformational space. Although the hybrid GA proposed here does not operate directly on the representation of the chromosomes, as they are the same before and after the local minimizations, it does on the fitness value which is used to decide on their survival onto the next generation, modifying thus the entire artificial evolutionary process.

5. PARALLEL IMPLEMENTATION OF THE HYBRID GENETIC ALGORITHM

The prediction of the tertiary structure of proteins by a heuristic method such as the genetic algorithm requires the consideration of a large number of conformers. Furthermore, the evaluation of an elaborated potential function such as the one proposed here requires intensive computations. The terms of the potential function which are especially computer power consuming are the local minimization of the potential function and the calculation of the solvent accessible surface area.

Developments in parallel and distributed computing offer a mean to overcome some of the limitations of single processor machines. The development of both, software and hardware in parallel processing has been remarkable in recent years. Mainframe computers composed of hundreds of processors, able to run programs in parallel, are now used routinely to solve problems whose solutions were thought to be impossible a few years ago. Parallel processing may be regarded as the commencement of a new era in computer technology.

The requirements of computer power of the algorithm described here as well as the intrinsic parallelism of the GA enhance the development of a parallel algorithm.

We have developed a parallel hybrid GA for polypeptide 3D structure prediction, and we have implemented it in a network of transputers.

The algorithm is illustrated in Figure 13. Here a population of chromosomes is distributed among the processors (transputers) constituting the network.

The architecture of a transputer, a processor with on-chip RAM and private storage memory, consists in a processor endowed with four links (hard channels) that allows it to communicate (transfer and receive data) with four other processors of the same type. A network of such processors makes a MIMD (multiple instruction multiple data) type of data processing. Figure 14 illustrates an array of four TRAMS or QUADPUTER.[19]

The algorithm developed here was tested on a network of five transputers.

The population of the GA is divided among the processors. The root processor where the master process runs, reads data

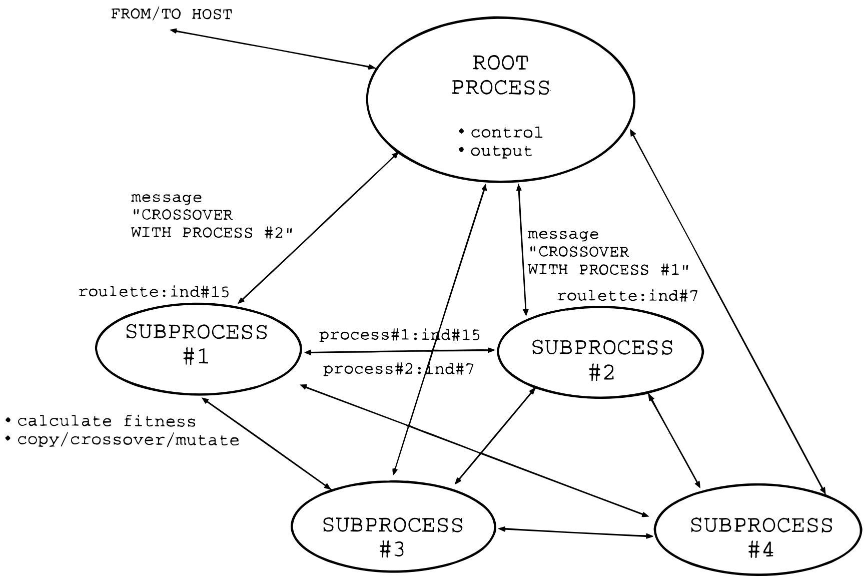

Figure 13. Flow diagram for the parallel hybrid GA.

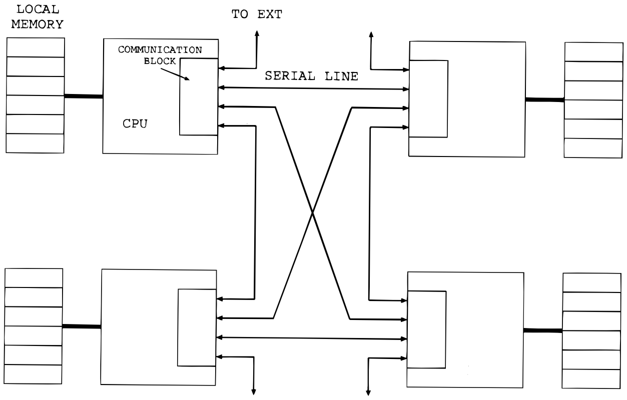

Figure 14. Architecture of a network of four transputers.

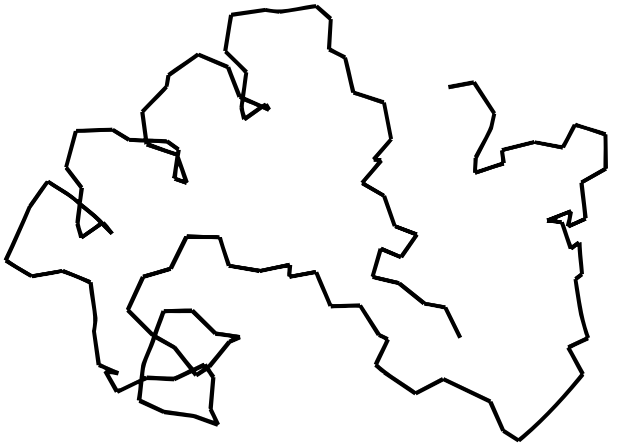

Figure 15. Native structure of CRAMBIN.

Table 1. Parameters for the Genetic Algorithm

parameter	value
no. of generations	500
no. of individuals in the population	200
mutation rate	0.03
crossover rate	0.60

Table 2. Parameters for the Fitness Function

coefficient	value
f_e	2.0
f_{HB}	2.0
f_c	1.5
f_S	1.5
f_{hp}	1.0

from the host server, performs assignment of parameters to the molecule, and transmits data to the slave processes to perform the evaluation of the potential function.

The slave processor performs the task of sending data to any of its four neighbor processes or receive data from any of them. A generation is built, in the network by collecting the results of the three GA operations (selection, mutation, crossover) performed by the network of transputers. A slave process running on one transputer selects the best individuals from its share of individuals (its population) and sends them to a neighbor processor waiting for data. Similarly it gets the signal to receive individuals from any of its neighbors. Each slave process possesses routines to receive, send, and copy an individual, to mate the incoming individual from other processor with one of its own individuals, and to perform random mutations on the offsprings.

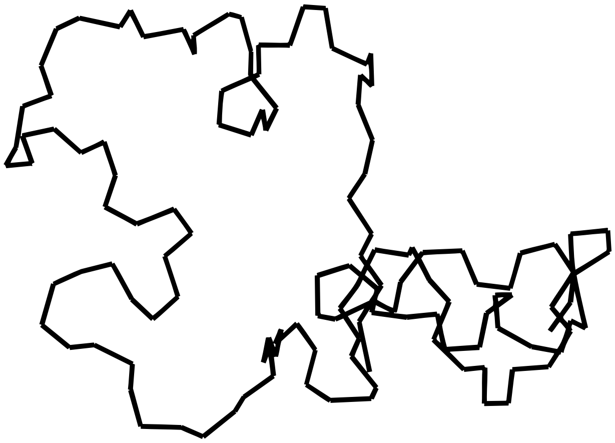

Figure 16. Predicted structure of CRAMBIN.

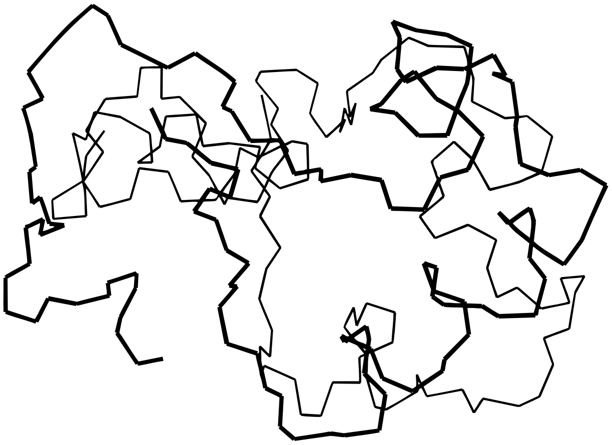

Figure 17. Superposition of native (boldline) and predicted structures for CRAMBIN (RMS = 6.69 Å).

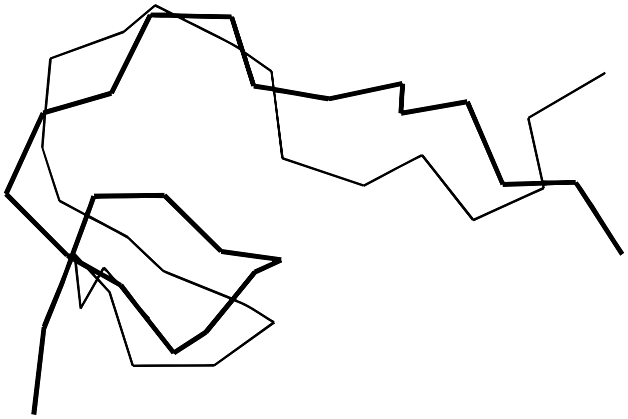

Figure 18. Superposition of fragments of predicted and of native conformers for CRAMBIN (frament between atoms 74 and 94. RMS = 1.73 Å).

6. RESULTS AND DISCUSSION

The parallel hybrid GA described here was tested elucidating the 3D structure of diverse polypeptides and comparing predicted conformers with the structure of the native conformers recorded in the crystallographic data base. Parameters for the genetic algorithm as well as coefficients for the potential energy function developed here are summarized in Tables 1 and 2, respectively. Although it is necessary to do a thorough study to determine the effects of these parameters on the evolutionary process, which is actually being carried out in our laboratories, the parameters shown in Tables 1 and 2 have been set on a qualitative basis.

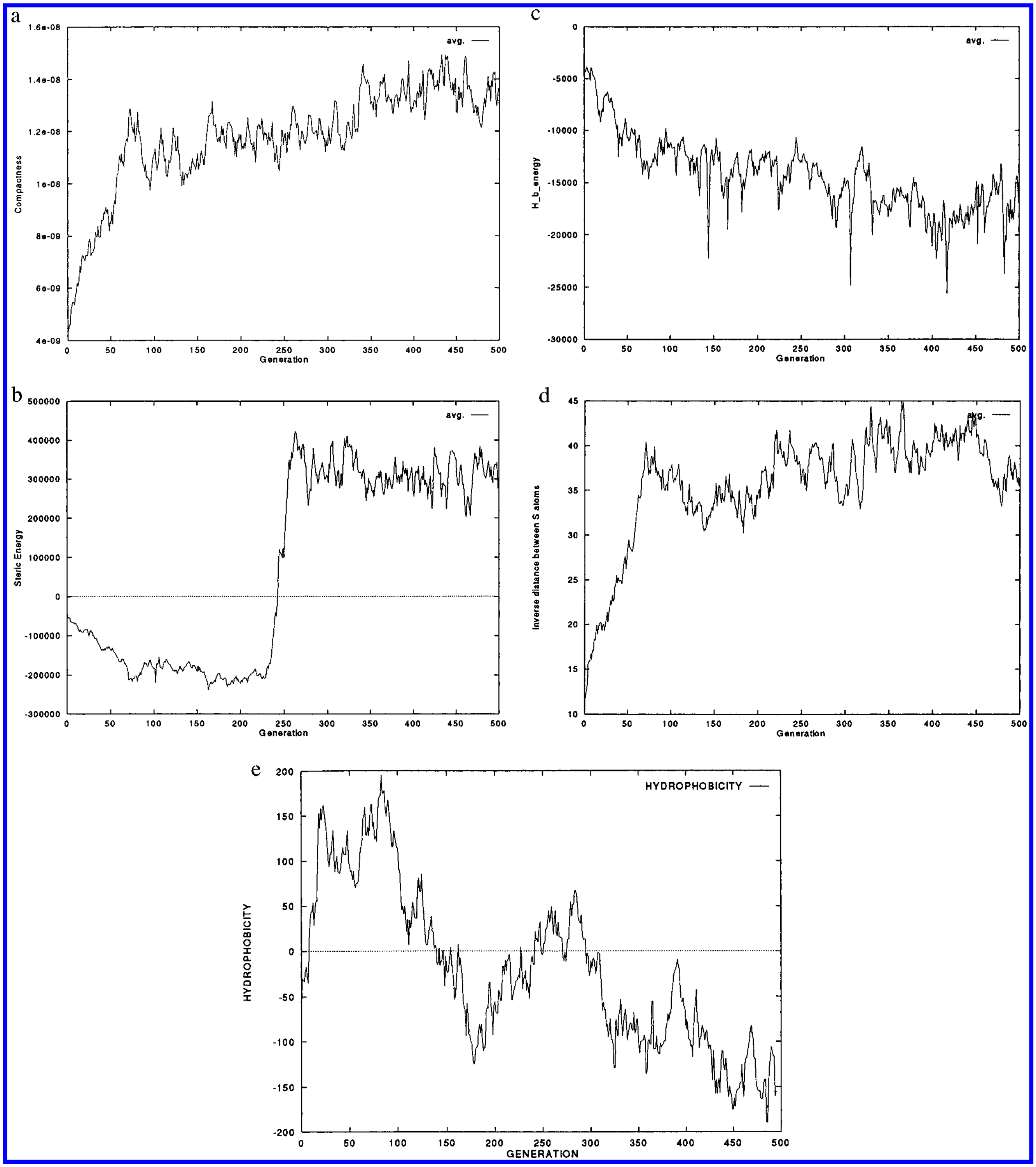

Figure 19. Evolution of the terms of the fitness function for BPTI: (a) evolution of the compactness term, (b) evolution of the steric energy term, (c) evolution of the hydrogen bond energy term, (d) evolution of the disulfide bridge term, and (e) evolution of the solvent effect term.

The first structure used to test the algorithm was the protein CRAMBIN constituted by 36 amino acid residues and which crystallographic data were extracted from the PDB.

Figure 15 displays the native structure, and Figure 16 shows the best conformer of the 250th generation of the evolutionary process. In Figure 17 both structures are superimposed (the native structure in bold line). The overall RMS[20] is 6.69 Å. Similarly to this individual of the last generation many others are within 9.90 Å of RMS.

Although the generated structure of CRAMBIN does not fit completely with the crystal structure, segments of well defined secondary structure similar to those of the native structure have been found. Figure 18 shows this type of segments. The smallest is 25 atoms of length of the main chain, and the RMS is as low as 1.7 Å.

Finally, we have used our implementation of the hybrid GA to predict the structure of bovine pancreatic trypsin inhibitor (BPTI) from its primary sequence. This protein is composed of 46 amino acids, and the most relevant factors in determining the tertiary structure of the protein are the disulfide bonds.

Figure 19 illustrates the variation of the steric energy, hydrogen bond energy, the compactness, the hydrophobic potential, and the disulfide bond terms of the evaluation

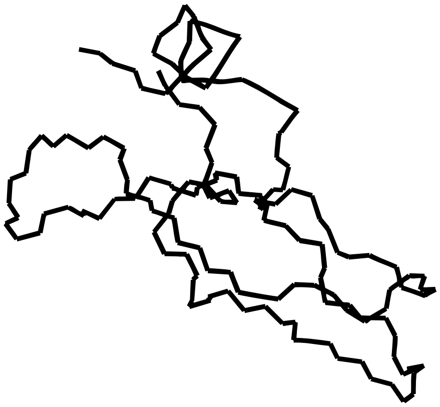

Figure 20. Native form of BPTI from the PDB.

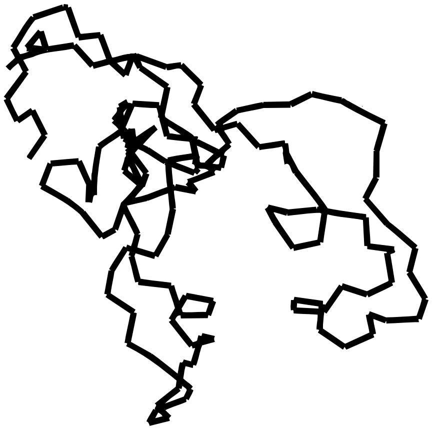

Figure 21. Best conformer of BPTI at the 500th generation of the GA.

function for the genetic process. While the compactness term (Figure 19a) increases, and those counting for the distance between sulfur atoms (Figure 19d) and the steric energy (Figure 19b) decrease, the hydrogen bonding energy (Figure 19c) and the hydrophobic factor (Figure 19e) show certain abnormalities in their evolution. It is especially remarkable the variation of the steric energy during the artificial evolution, which at approximately generation 250 passes from a positive to a negative value, that is, energetically more stable conformers begin to appear as the operations of the evolution process proceed to generate better generations of conformers.

The crystal structure of BPTI, extracted from the PDB is shown in Figure 20. Values for the energy (with hydrogen coordinates supplied by the computer) show a value well over the values of the conformers of evolved populations. That is, conformers at lower steric energies are found by the algorithm. A comparison of the best fit individual of the 500th generation with the crystal structure gives a RMS of 7.81 Å. Here again, the structures possess dissimilarities; however, patterns observed in the crystal structure are also in formation in the population of conformers that evolves. Figure 21 shows the best fit conformer of the last generation, and Figure 22 is a superposition of both structures. Fragments as large as 30 atoms of the main chain are also superimposed. The RMS for the fragment between atoms 84 and 103 is 1.43 Å. The superposition of this fragment in the predicted structure with the respective fragment in the native structure is shown in (Figure 23a). Similarly, fragments between atoms 48 and 67 for six conformers of the

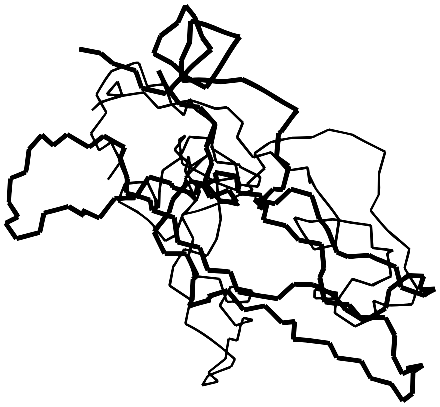

Figure 22. Superposition of predicted and crystal conformers of BPTI (RMS = 7.81 Å).

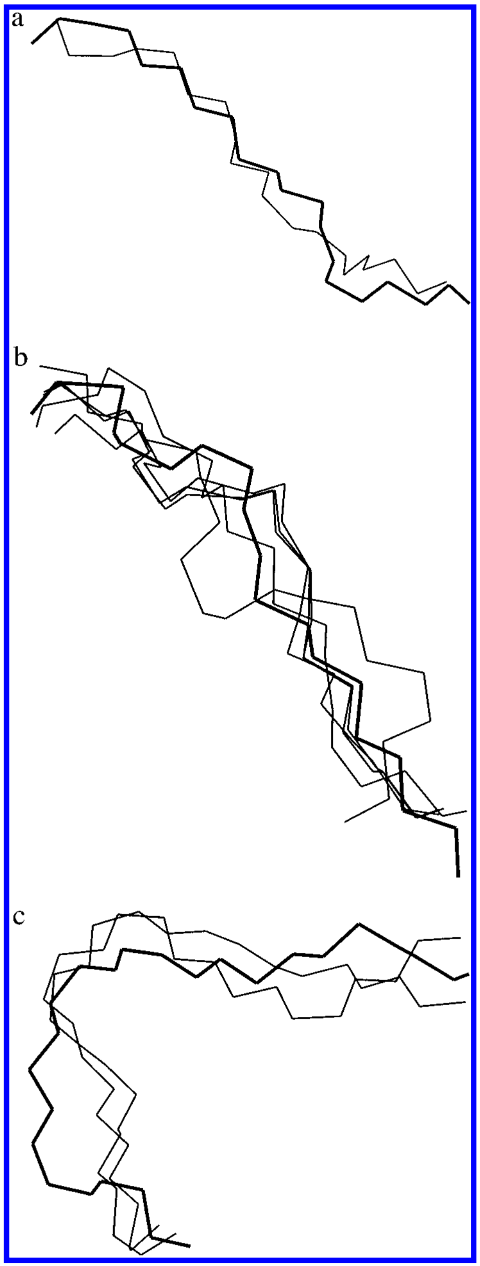

Figure 23. Superposition of segments of predicted conformers of BPTI and native BPTI (boldline): (a) fragment between atoms 84 and 103 (RMS = 1.43 Å), (b) superposition of diverse fragments between atoms 48 and 67 (1.20 < RMS < 1.89), and (c) fragment between atoms 97 and 121 (RMS = 1.8456 Å).

last generation are shown superposed to the crystal fragment (boldface line) (Figure 23b) where the RMS ranges from

1.20 to 1.89 Å. Finally Figure 23c show fragments between atoms 97 and 121 for two conformers with a RMS of 1.8456 Å.

The distance between pairs of sulfur atoms of cysteines is also a remarkable factor in the folding of BPTI, and this is reproduced by the hybrid GA we introduce here. As the evolutionary process proceeds, better conformers of BPTI are found, when distances among pairs of sulfur atoms are smaller.

Although the values of the factors corresponding to the hydrogen bond energy as well as the solvent were not as we expected during the evolution of the conformers populations, the fitness function successfully reproduces many of the characteristics of the tertiary structure of the polypeptide. Its success can be attributed to the fact that it takes into account many of the factors influencing the folding process and the use of parameters which include more accurate experimental information on the folding process of proteins. The new energy terms which were added to the fitness function (i.e., compactness factor, solvent effect, disulfide bond term) seem to be critical for predicting folded conformations of proteins.

Furthermore, we propose a hybrid genetic algorithm to have populations whose individuals are more than mere unrealistic three dimensional strands of amino acids. Local minima conformations, as already shown by many of the examples mentioned above, retain already some structural elements of the global solution. Thus, local minimization, although computer power consuming, assists the global conformational mapping operation, with clues on structural schemata of the amino sequence. This facilitates the evolutionary process, and at the end results in less generations created and less processing time.

Furthermore, we introduce a parallel algorithm that not only can solve the problem of huge processing times but also assists the genetic algorithm maintaining a diversity of population, allowing a more extended search of the entire conformational space for a molecule.

A rather complete match of the crystal and predicted structures, the goal of the present study, can be attained, we believe, with a further calibration of a potential energy function similar to the one presented in this work. In fact, here we have tried to incorporate only the most relevant factors from the macroscopic point of view. It is still necessary to examine the behavior of the hydrophobic effect and hydrogen bond energies and express in a more effective way the mode in which they affect the folding process. It is true that the determination of the effects of only one factor in the whole process of folding is difficult; however, the program developed here can also play a role of a learning tool, with which many experiments can be carried out to simulate more exactly the phenomena of protein folding.[7]

The factors obtained in this way will certainly improve the potential function and give a deeper insight into the process of protein folding and thus a better tool to predict 3-D structures of polypeptides and proteins.

ACKNOWLEDGMENT

The present work was carried out under the financial support of the "Grant for Research of the University President" of Toyohashi University of Technology. Part of this work was presented at the XIth International Conference of Computers in Chemical Research and Education thanks to the grant from the Foundation for C&C Promotion, funded by the NEC corporation of Japan.

REFERENCES AND NOTES

(1) Anfinsen, C. *The Molecular Basis of Evolution*; John Wiley & Sons: 1959.

(2) Holley, L. H.; Karplus, M. [10] *Neural Networks for Protein Structure Prediction. Methods Enzymol.* **1991**, *202*, 204−224.

(3) Davis, L. *Handbook of Genetic Algorithm*; Van Unestranged Reynold: 1991.

(4) Le Grand, S. M.; Merz, K. M., Jr. *The application of the Genetic Algorithm to the minimization of potential energy functions. J. Global Optimization* **1993**, *3*, 49−66.

(5) Dandekar, T.; Argos, P. *Potential of Genetic Algorithms in Protein Folding and Protein engineering simulations. Protein Engineering* **1992**, *5*, 637−645.

(6) Weissman, J. S.; Kim, P. S. *Reexamination of the Folding of BPTI: Predominance of Native Intermediates. Science* **1991**, *253*, 1386−1393.

(7) Del Carpio, C. A., manuscript in preparation.

(8) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley Publishing Company, Inc.: 1989.

(9) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekhara, V. *Stereochemistry of Polypeptide Chain Configurations. J. Mol. Biol.* **1963**, *7*, 95−99.

(10) Hopfinger, A. J. *Conformational Properties of Macromolecules*; Academic Press: 1973.

(11) Burkert, U.; Allinger, N. L. *Molecular Mechanics*; American Chemical Society: 1982.

(12) Allinger, N. L. *Calculation of Molecular Structure and Energy by Force-Field Methods. Adv. Phys. Org. Chem.* **1976**, *13*, 1−82.

(13) Allinger, N. L. *Operating Instructions for the MM3 Program*; **1992**, *105*.

(14) Cramer, C. J.; Truhlar, D. G. *General Parameterized SCF Model for Free Energies of Solvation in Aqueous Solution. J. Am. Chem. Soc.* **1991**, *113*, 8305.

(15) Richmond, T. J. *Solvent Accessible Surface Area and Excluded Volume in Proteins. J. Mol. Biol.* **1984**, *178*, 63−89.

(16) Lee, C.; Subbiah, S. *Prediction of Protein Side-chain Conformation by Packing Optimization. J. Mol. Biol.* **1991**, *217*, 373.

(17) Pickett, S. D.; Sternberg, M. J. E. *Empirical Scale of Side-chain Conformational Entropy in Protein Folding. J. Mol. Biol.* **1993**, *231*, 825−839.

(18) Allinger, N. L.; Tribble, M. T.; Miller, M. A.; Wertz, D. H. *Conformational Analysis. LXIX. An Improved Force Field for the Calculation of the Structures and Energies of Hydrocarbons. J. Am. Chem. Soc.* **1971**, *93*, 1637−1648.

(19) *Quadputer Owner's Manual*; Microway, Inc.; 1993.

(20) Here the RMS (root mean square deviation) is computed by the following expression

$$\mathrm{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(x_i - x_{ic})^2 + (y_i - y_{ic})^2 + (z_i - z_{ic})^2]}$$

where x_i, y_i, and z_i are the Cartesian coordinates for the atoms of the predicted conformer and x_{ic}, y_{ic}, and z_{ic} are the Cartesian coordinates for corresponding atoms of the crystal structure.

CI950106R