Genetic Algorithms for Protein Tertiary Structure Prediction

Steffen Schulze-Kremer

Brainware GmbH, Gustav-Meyer Allee 25, D-1000 Berlin 65 and Free University Berlin, Institute of Crystallography, Takustraße 6, D-1000 Berlin 33, Germany
email STEFFEN@KRISTALL.CHEMIE.FU-BERLIN.DE

Abstract

This article describes the application of genetic algorithms to the problem of protein tertiary structure prediction. The genetic algorithm is used to search a set of energetically sub-optimal conformations. A hybrid representation of proteins, three operators MUTATE, SELECT and CROSSOVER and a fitness function, that consists of a simple force field were used. The prototype was applied to the *ab initio* prediction of Crambin. None of the conformations generated by the genetic algorithm are similar to the native conformation, but all show much lower energy than the native structure on the same force field. This means the genetic algorithm's search was successful but the fitness function was not a good indicator for native structure. In another experiment, the backbone was held constant in the native state and only side chains were allowed to move. For Crambin, this produced an alignment of 1.86 Å r.m.s. from the native structure.

Keywords

genetic algorithm / protein tertiary structure / *ab initio* prediction

1. Introduction

The work presented in this article concerns the application of genetic algorithms [Holland, 1975] to the problem of protein structure prediction [Schulz & Schirmer, 1979; Lesk, 1991; Branden & Tooze, 1991]. Genetic algorithms in computer science are heuristic methods that operate on pieces of information like nature does with genes during evolution. Individuals, represented by a linear string of letters of an alphabet (in nature nucleotides, in genetic algorithms bits, strings or numbers) are allowed to mutate, crossover and reproduce. Members of a new generation and their parents are then evaluated by a fitness function. Only the best individuals enter the next reproduction cycle. After a given number of cycles, the population consists of well adapted individuals that each represent a solution for the problem of optimizing the given fitness function. Although it cannot be proven that the final

individuals contain the optimal solution it can be mathematically shown that the overall effort is maximised during a run [Holland, 1975]. In some applications, where the search space was too large for other heuristic or analytic methods, genetic algorithms produced better solutions than those known before [Davis, 1991].

In this work, the individuals are conformations of a protein and the fitness function is a simple force field. What follows is a description of the representation formalism, the fitness function and the operators used. The results of a run on Crambin are then presented and finally, the results of an experiment for side chain placement are shown.

2. Protein Representation

For any application of a genetic algorithm a choice has to be made on the representation of the "genes". In the present work, the so-called hybrid approach was taken [Davis, 1991]. This means, that the objects the genetic algorithm processes are encoded by numbers instead of bit strings, which were used in the original genetic algorithm [Holland, 1975]. A hybrid representation is usually easier to implement and also facilitates the use of operators. However, three potential disadvantages are encountered: strictly speaking, the mathematical foundation of genetic algorithms holds only for binary representations; binary representations run faster in many applications; and an additional encoding / decoding process may be required. For a hybrid representation of proteins there are at least two immediately intuitive choices, one being Cartesian coordinates, the other torsion angles.

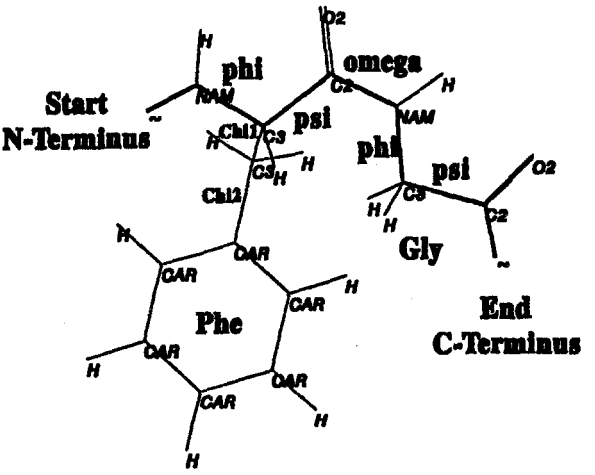
2.1 Cartesian Coordinates

In this representation, the coordinates of all atoms of a protein are stored in a fixed order, i.e. the nth number always refers to the same component of the 3D coordinate of a particular atom. This representation has the advantage of being easily converted to and from the conformation of a protein. However, it faces the disadvantage that the use of a random mutation operator would most of the time create invalid instances, where atoms lie too far apart or collide. To prevent this from happening, a filter which eliminates invalid individuals had to be installed between the operator and the fitness function. This would be a rather time consuming process, especially when a large percentage of the individuals had to be sorted out. Therefore, the representation in Cartesian coordinates would considerably slow down the progress of the genetic algorithm. These considerations have led to another model.

2.2 Torsion Angles

In this representation a protein is described by a set of non-redundant torsion angles. This can be done under the assumption of constant standard binding geometries. Bond lengths and bond angles are taken to be constant and cannot be changed by the genetic algorithm. This assumption is a simplification of the real situation where bond length and bond angle depend on the environment of an atom within a protein. However, a set of non-redundant torsion angles allows enough degrees of freedom and therefore enough variability to represent any native conformation with only small spatial differences when superimposed on the original structure (i.e. small root mean square deviations, abbrev. r.m.s.)

The following diagram illustrates the use of torsion angles. A small fragment from a hypothetical protein is shown. Two basic building blocks, i.e. the amino acids phenyalanine (Phe) and glycine (Gly), are drawn as wire frame models. Atoms are described by their chemical symbols. The thicker bonds indicate the backbone. Main chain torsion angles phi, psi and omega are indicated next to their rotatable bond.



Special to the representation by torsion angles is the fact that small changes in the φ (phi) / ψ (psi) angles can induce large changes in the overall conformation. This is of advantage when creating variability through genetic operators. Also, as a consequence of this representation, one can get relative large r.m.s. differences when comparing a native conformation with a reconstruction of the same molecule under the assumption of constant standard bonding geometries. This difference can be minimized by slight changes in the torsion angles.

In the present work, the representation by torsion angles was used. The torsion angles of 129 proteins from the Brookhaven database [Bernstein *et al*, 1977] were statistically analyzed. The ten most frequently occurring values for each non-redundant torsion angle were collected in 10° intervals. At the beginning of the run, individuals were intitialized with either complete extended conformations, where all torsion angles are 180°, or with a random selection of the ten most frequent intervals for each torsion angle. For the ω (omega) torsion angle the constant value of 180° was used because of the rigidness of the double bond between the atoms C_{2i} and N_{i+1}. An evaluation of ω angles shows that with the exception of prolin average deviations of 5° occur from the mean of 180°, in rare cases 15°.

The genetic operators work on the torsional representation. For the application of the fitness function, it is necessary to translate proteins represented in torsion angles into Cartesian coordinates. Two programs have been written to carry out the conversion of one representation into another. The binding geometries were taken from the molecular modelling package Alchemy [Vinter *et al*, 1987] and the bond lengths from the program Charmm [Brooks *et al*, 1983]. Either a complete form with explicit hydrogen atoms or the so-called extended atom representation can be calculated.

Next to the calculation of the fitness of an individual, the conversion of the Cartesian representation into the torsional representation is the second most time consuming step. The torsion angle of each bond that extends to at least one other atom is calculated. In the reverse process, atoms with standard binding geometries are added to the molecule one by one according to residue type. Then, they are rotated by the amount of their torsion angle. This operation can be replaced by the addition of two vectors because both torsion angle and radius are known. The translation programs were successfully tested by comparison of native and reconstructed conformations. A native structure was translated into its torsion angle representation and back into a Cartesian representation. The native and reconstructed conformations were then superimposed with the FIT routine in the program Alchemy.

2.3 Format of Genes

Genes are stored one conformation in one ASCII file. The number of records per file equals the number of residues of the protein. Each record starts with a three letter identifier of the residue type. Ten floating point numbers in the format -xxx.xx follow, which stand for the torsion angles φ, ψ, ω, χ1, χ2, χ3, χ4, χ5, χ6, χ7, in this order. For residues with less than seven side chain torsion angles the extra fields

carry the value 999.99. The ω angle was kept at the constant value of 180° in most applications.

3. Fitness Function

The fitness function in a genetic algorithm represents the "environment" of all individuals. In order to "survive" an individual has to perform better than its competitors in terms of the fitness function. In the present work, a simple steric potential energy function was chosen as the fitness function.

3.1 Motivation

The main reason for the choice of a simple steric potential energy function as the fitness function of the genetic algorithm was the observation, that up to date it has not been possible to develop a method to efficiently search the conformation space spanned by a force field for its global optimum. The problem with this task lies in the large number of degrees of freedom for a protein of average size. In general, molecules with n atoms have $3n$-6 degrees of freedom. This amounts in the case of proteins with about 100 residues to ((100 residues · approx. 20 atoms per residue) · 3) - 6 = 5994 degrees of freedom. Systems of equations with this number of variables are analytically intractable today [Gunsteren & Berendsen, 1990].

Efforts to empirically find the optimal solution are equally difficult. If there are no constraints for the conformation of a protein and only its primary structure given, then the number of conformations for a protein of medium size (100 residues) can be estimated to be (approx. 5 torsion angles per residue · approx. 5 likely values per torsion angle)^{100} = 25^{100}. Because potential energy function is not monotonous, in the worst case 25 to the power of 100 conformations had to be evaluated to find the global optimum. This is clearly beyond capacity of today's and tomorrow's super computers.

As can be seen from a number of previous applications, genetic algorithms were able to find sub-optimal solutions to problems of equally large search space [Davis, 1991; Lucasius & Kateman, 1989; Tuffery *et al*, 1991]. Sub-optimal in this context means, that it cannot be proven that the solutions generated by the genetic algorithm are in fact the optimum but they were in many cases better than any previously known solution. This can be of much help in n.p.-complete domains, where analytical solutions of the problem are not available. Therefore, it was attempted to apply the genetic algorithm to the *ab initio* protein structure prediction problem.

3.2 Conformational Energy

The steric potential energy function was taken from the program Charmm [Brooks *et al*, 1983]. It is the sum of the expressions for E_{bond} (bond length potential), E_{phi} (bond angle potential), E_{tor} (torsion angle potential), E_{impr} (improper torsion angle potential), E_{vdW} (van der Waals pair interactions), E_{el} (electrostatic potential), E_H (hydrogen bonds), and of two expressions for solvent interaction, E_{cr} and E_{cphi}

$$E = E(bond) + E(phi) + E(tor) + E(impr) + E(vdW) + E(el) + E(H) + E(cr) + E(cphi).$$

As constant bond lengths and bond angles were assumed, the expressions for E_{bond}, E_{phi} and E_{impr} are constant for each protein. The expression E_H was omitted because it would have required to exclude the effect of hydrogen bonds from the expressions for E_{vdW} and E_{el}. This, however, was not done by the authors of Charmm. In all runs, folding was simulated in vacuum with no ligands or solvent, i.e. E_{cr} and E_{cphi} are constant. This is certainly a crude simplification which will have to be extended in future. Thus, the potential energy function simplifies to:

$$E = E(tor) + E(vdW) + E(el).$$

If only the three expressions E_{tor}, E_{vdW} and E_{el} were calculated, there would be no force to drive the protein to a compact folded state. An exact solution to this problem is the inclusion of entropy. Unfortunately, measuring the difference in entropy between folded and unfolded state requires taking into account interactions of the protein with solvent. This cannot be done in a simple way. To have a running prototype, it was therefore decided to introduce *ad hoc* a pseudo entropic force, that drives the protein to a globular state. Analysis of a number of globular proteins reveals the following empirical relation between the number of residues and the diameter:

$$\text{expected diameter} = 8 \cdot \text{length}^{1/3}\ \text{Å}$$

The pseudo entropic potential for a conformation is a function of its actual diameter. The diameter is defined to be the largest distance between any C_α atoms in a given conformation. A positive exponential of the difference between expected and the actual diameter E_{pe} is added to the potential energy, if that difference is less than 15 Å. If the difference is greater than 15 Å a fixed amount of energy is added (10^{10} kcal/mol). If the actual diameter is smaller than the expected diameter, E_{pe} is zero. This has the effect, that more extended conformations have more positive potential energy values and are therefore less fit for reproduction.

$$E_{pe} = 4^{(\text{actual diameter - expected diameter})} \text{ kcal/mol}$$

Occasionally, if two atoms are very close, the E_{vdW} can become very large. The maximum value for E_{vdW} in this case is 10^{10} kcal/mol and the expressions for E_{el} and E_{tor} are not calculated.

Runs have been performed with the potential energy function E as described above, where lower values mean fitter individuals and with a variant, where the four expressions E_{tor}, E_{vdW}, E_{el} and E_{pe} were given individual weights. The results were similar in all cases. Especially, scaling down the dominant effect of electrostatic interactions did not improve the results.

4. Operators

In order to combine individuals of one generation to produce new offspring, nature as well as genetic algorithms apply several operators. In the present work, individuals are protein conformations in their torsion angle representation under the assumption of constant standard binding geometries (see above). Three operators were invented to process these individuals: SELECT, MUTATE and CROSSOVER. The decision about the application of one operator is made at run time and can be controlled by various parameters.

4.1 SELECT

The first operator is the SELECT operator. If SELECT gets activated for a particular torsion angle, this angle will be replaced by a random choice of one of its ten most frequently occurring values. The decision, whether a torsion angle will be modified by SELECT is made independently for each torsion angle in a protein. A random number between 0 and 1 is generated and if this number is greater than the SELECT parameter at that time, SELECT is triggered. The SELECT parameter can change dynamically during the run. The values for SELECT to choose from are from a statistical analysis of 129 proteins from the PDB database. The number of values occurring in each of the 36 10°-intervals for that torsion angle was counted. The ten most frequent intervals, each represented by its left boundary, are available for substitution.

4.2 MUTATE

The MUTATE operator consists of three components: the 1°, 5° and 10° operator. After application of the SELECT operator and independently from it, for each torsion angle in a protein two decisions are made; first, whether a MUTATE operator will be applied and second, if the first decision was in favor of the

MUTATE operator, which of the three components will be carried out. Mutation is done by incrementing or decrementing (always an independent random chance of 1:1) the torsion angle by 1°, 5° or 10°. Care is taken that the range of torsion angle values is always in the [-180°, 180°] interval. The probability of applying the MUTATE operator is controlled by the MUTATE parameter, which can change dynamically during the run. Similarly, three additional parameters control the probability for choosing among the three components.

4.3 CROSSOVER

The CROSSOVER operator has two components: the two point crossover and the uniform crossover. It is applied to two genes (individuals) independently of the SELECT and MUTATE operators. First, the parent generation of individuals, possibly modified by SELECT and MUTATE, are randomly grouped pairwise. For each pair, an independent decision is made whether or not to apply the CROSSOVER operator. The probability for this is controlled by the CROSSOVER parameter, which can change dynamically during the run. If the decision is "no", the two individuals are not further modified and added to the list of offspring. If the decision is "yes", a choice between the two point crossover and the uniform crossover has to be made. This decision is controlled by two other parameters, that also can change dynamically during the run.

The two point crossover selects randomly two sites (residues) on one of the individuals. Then, the fragment between the two residues is exchanged with the corresponding fragment of the second individual. The uniform crossover decides independently for each residue, whether or not to exchange the torsion angles of that residue. The chance for exchange is always 1:1.

4.4 Parameterization

As indicated in the previous sections, there are a number of parameters to control the run of a genetic algorithm on protein conformations. The parameter values which were used for the runs presented in the results section are summarized in the following table. The ω torsion angle was kept constant at 180°. The intitial generation was created by a random selection of the torsion angles from the list of the ten most frequently occurring values for each angle. There were ten individuals in one generation. The genetic algorithm terminated after 1000 generations. At the start of the run, the chance for a torsion angle to be modified by the SELECT operator is 80%, at the end of the run 20%. The probability decreases linearly with the number of generations. In contrast, the chance of applying the MUTATE operator increases from 20% at the start to 70% at the end of the run. The 10°

component of the MUTATE operator is dominant at the start of the run (60%), whereas it is the 1° component at the end (80%). Likewise, the chance of performing the CROSSOVER operator rises from 10% to 70%. At the beginning of the run mainly uniform CROSSOVER is done (90%), at the end mainly two point CROSSOVER (90%).

This parameter setting has a small number of individuals but a large number of generations. This was chosen to keep cpu time low while allowing a maximum number of crossover events. This run took about 12 hours on a SUN SPARC station. At the beginning of the run, SELECT and uniform CROSSOVER are applied most of the time. This is to create some variety in the population. At the end of the run, the 1° component of the MUTATE operator dominates the scene. This is intended for fine tuning the conformations that have survived the fitness pressure so far.

Parameter	Value
ω angle constant 180°:	on
initialize start generation:	random torsion angels
number of individuals:	10
number of generations:	1000
SELECT (start)	80%
SELECT (end)	20%
MUTATE (start)	20%
MUTATE (end)	70%
MUTATE (start 10°)	60%
MUTATE (end 10°)	0%
MUTATE (start 5°)	30%
MUTATE (end 5°)	20%
MUTATE (start 1°)	10%
MUTATE (end 1°)	80%
CROSSOVER (start)	70%
CROSSOVER (end)	10%
CROSSOVER (start uniform)	90%
CROSSOVER (end uniform)	10%
CROSSOVER (start two point)	10%
CROSSOVER (end two point)	90%

4.5 Generational Replacement

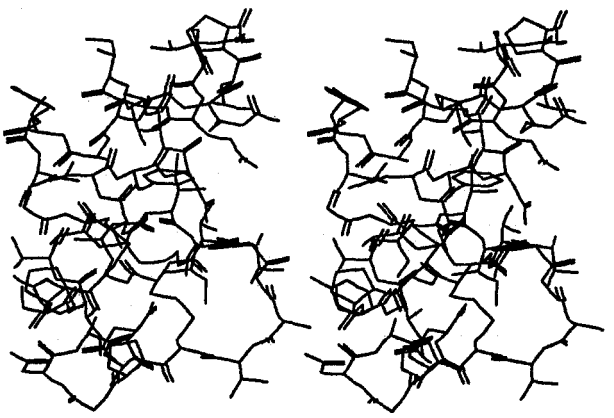
There are a number of ways of how to select from the individuals of one generation and its offspring the parents for the next generation. Given the constraint, that the number of individuals should remain constant, inevitably, some individuals have to be discarded. Two ways of controlling the transition are complete replacement and elitistic replacement. In the first case, all offspring become parents in the next generation. The parents of the old generation are completely discarded. This has the disadvantage, that a fit parent can be lost, if it only once produces bad offspring. With elitistic replacement all parents and offspring of one generation are sorted according their fitness. If the size of the population is n, then the n fittest

individuals are selected as parents for the following generation. This mode has been used in the present work.

5. *Ab initio* Prediction Results

A prototype of a genetic algorithm with the representation, fitness function and operators as described above has been implemented. For *ab initio* prediction the sequence of Crambin was given to the program. Crambin is a plant seed protein from the cabbage *Crambe Abyssinica*. The structure was determined by W.A. Hendrickson and M.M. Teeter up to a resolution of 1.5 Å [Hendrickson & Teeter, 1981]. Crambin has a strong amphiphilic character, which makes it especially difficult to predict its tertiary structure with a simple force field. Because of its good resolution and its small size of 46 residues it was decided to use Crambin as a first candidate to start with. Independently from this work Scott Le Grand in the laboratory of Prof. Karplus at MIT did similar experiments, using a GA and a different force field. The results were basically the same as those presented here [Le Grand & Merz 1991]. The following structures are displayed in stereo projection. If the observer manages to look cross eyed at the picture in a way that superimposes both halves, a three dimensional impression can be perceived.

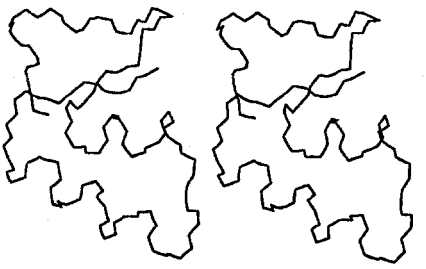
Crambin with side chains



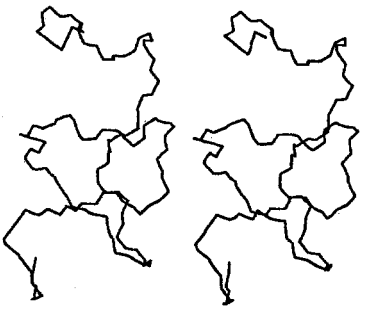
5.1 Conformations

In the following, the backbone structure of the ten best individuals generated by the genetic algorithm are shown in stereo projection.

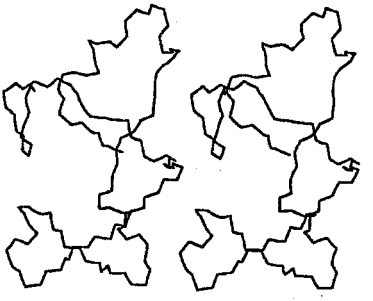
Crambin, native



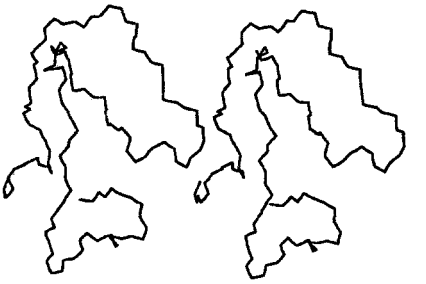
Individual P1



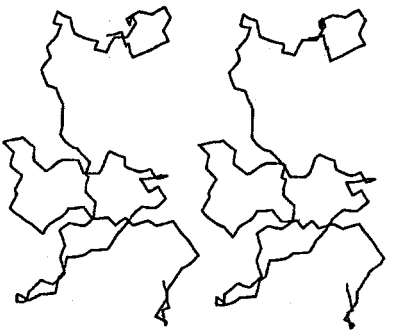
Individual P2



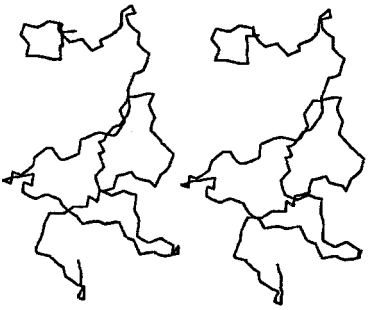
Individual P3



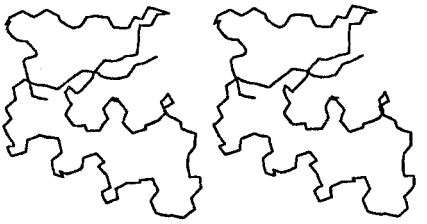
Individual P4



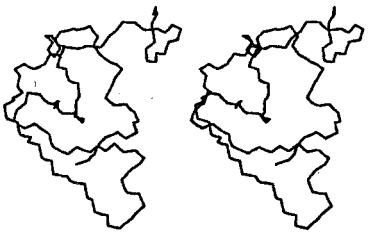
Individual P5



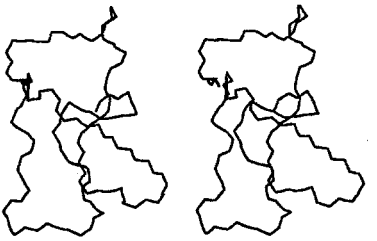
Crambin, native



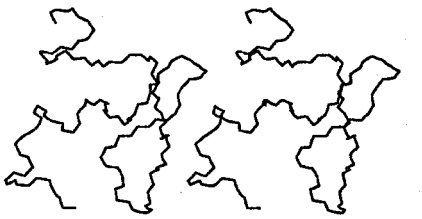
Individual P6



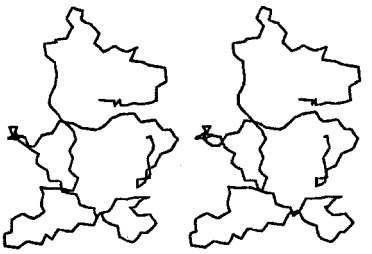
Individual P7



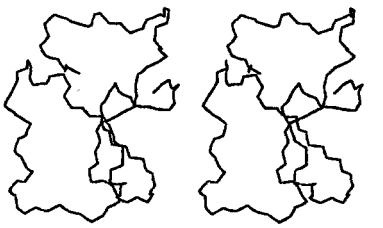
Individual P8



Individual P9



Individual P10



It can be seen from the graphs that none of the individuals generated show significant structural similarity to the native Crambin conformation. This can be confirmed by superpositioning the generated structures with the native conformation. The following table shows the r.m.s. differences between individuals

P1 to P10 and the native conformation. All values are in the range of 9 Ångström, which rejects any structural homology.

r.m.s. Deviation to Native Crambin

P1	10.07 Ångström	P6	10.31 Ångström
P2	9.74 Ångström	P7	9.45 Ångström
P3	9.15 Ångström	P8	10.18 Ångström
P4	10.14 Ångström	P9	9.37 Ångström
P5	9.95 Ångström	P10	8.84 Ångström

The following table shows the r.m.s. differences between the generated individuals. They can be grouped into two classes. The members within each class are similar, whereas structures from both classes have no similarity. One class holds the individuals P1, P2, P4, P5, P6, P8 and P9. The other class has P3, P7 and P10. The fact, that two unrelated classes of conformations were generated, means that the genetic algorithm did simultaneously search in different regions of the conformation space and thus was *not trapped in one a local optimum.*

r.m.s. Deviation within Generated Individuals

P1	P10	8.40	P10	P8	8.57	P4	P5	1.10
P1	P2	1.73	P10	P9	7.30	P4	P6	0.71
P1	P3	9.52	P2	P3	8.96	P4	P7	8.71
P1	P4	0.86	P2	P4	1.44	P4	P8	0.93
P1	P5	1.43	P2	P5	1.13	P4	P9	2.15
P1	P6	1.20	P2	P6	1.63	P5	P6	1.28
P1	P7	9.03	P2	P7	8.46	P5	P7	8.98
P1	P8	0.58	P2	P8	2.12	P5	P8	1.75
P1	P9	2.30	P2	P9	1.15	P5	P9	2.04
P10	P2	7.75	P3	P4	9.24	P6	P7	8.71
P10	P3	1.57	P3	P5	9.52	P6	P8	1.28
P10	P4	8.18	P3	P6	9.23	P6	P9	2.44
P10	P5	8.39	P3	P7	1.13	P7	P8	9.11
P10	P6	8.15	P3	P8	9.61	P7	P9	8.21
P10	P7	1.74	P3	P9	8.57	P8	P9	2.68

5.2 Energies

The following table lists the values of the four contributions to the potential energy function for the ten individuals generated. The total energy of all individuals is much lower than the energy for the native conformation of Crambin: E(vdW) -12.8 kcal/mol, E(el) 11.4 kcal/mol, E(tor) 60.9 kcal/mol, E(pe) 1.7 kcal/mol and E(total)

61.2 kcal/mol. It is obvious, that the largest contribution comes from electrostatic interactions. This is due to the six partial charges in Crambin. For a more elaborate force field these charges have to be neutralized.

Energy Contributions of the Generated Individuals

Individual	E_{vdw}	E_{el}	E_{tor}	E_{pe}	E_{total}
P1	-14.9	-2434.5	74.1	75.2	-2336.5
P2	-2.9	-2431.6	76.3	77.4	-2320.8
P3	78.5	-2447.4	79.6	80.7	-2316.1
P4	-11.1	-2409.7	81.8	82.9	-2313.7
P5	83.0	-2440.6	84.1	85.2	-2308.5
P6	-12.3	-2403.8	86.1	87.2	-2303.7
P7	88.3	-2470.8	89.4	90.5	-2297.6
P8	-12.2	-2401.0	91.6	92.7	-2293.7
P9	93.7	-2404.5	94.8	95.9	-2289.1
P10	96.0	-2462.8	97.1	98.2	-2287.5

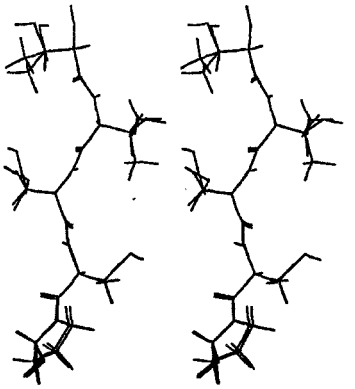
6. Side Chain Placement Results

Crystallographers often face the problem of positioning the side chains when the primary structure and the conformation of the backbone is known. At present, there is no method that automatically does side chain placement with sufficient fidelity for routine, practical use. The side chain placement problem is much easier than *ab initio* prediction but still too complex for analytical treatment.

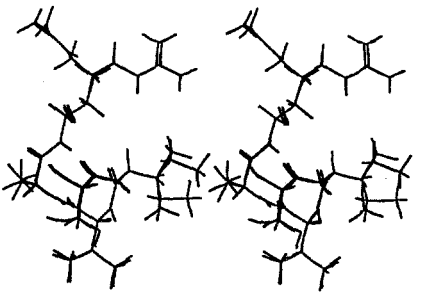
The genetic algorithm approach, as described above, can also be used for side chain placement. The torsion angles φ, ψ, and ω are kept constant at the values for the given backbone. Side chain placement by the genetic algorithm was done for Crambin. For each five residues, a superposition of the native and the predicted conformation was done. This is shown in stereo projection graphs on the following pages.

As can be seen, the predictions are quite well in agreement with the native conformation in most of the cases. The overall r.m.s. difference in this example is 1.86 Å. This is comparable to the results from a simulated annealing approach (1.65 Å) [Lee & Subbiah, 1991] and a heuristic approach (1.48 Å) [Tuffery *et al*, 1991].

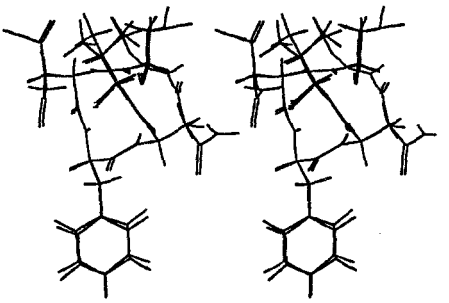
Superposition of Residues 1-5



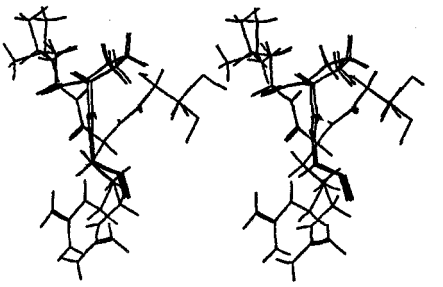
Superposition of Residues 6-10



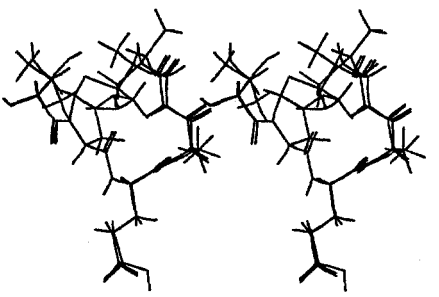
Superposition of Residues 11-15



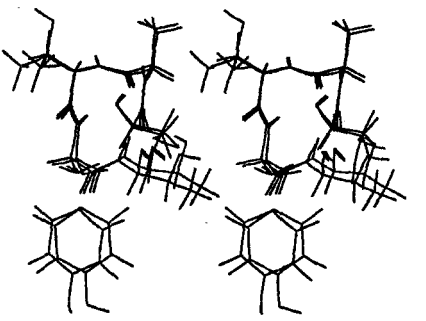
Superposition of Residues 16-20



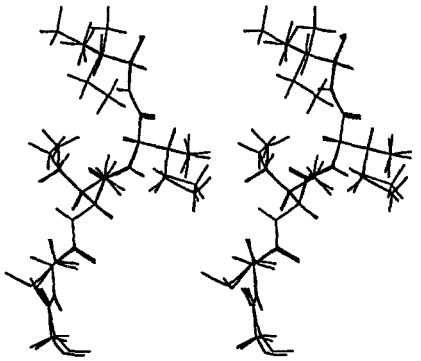
Superposition of Residues 21-25



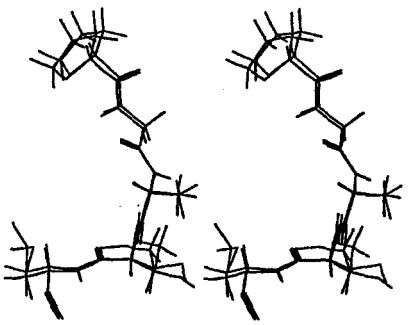
Superposition of Residues 26-30



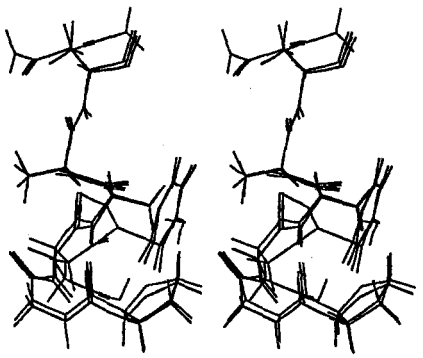
Superposition of Residues 31-35



Superposition of Residues 36-40



Superposition of Residues 41-46



7. Discussion and Conclusion

A prototype for the application of a genetic algorithm to the problem of protein tertiary structure prediction is presented. The genetic algorithm searches for energetically favorable conformations. A hybrid representation of proteins and three operators MUTATE, SELECT and CROSSOVER to manipulate the "genes" of a genetic algorithm were developed together with a fitness function, that consists of a simple force field. The work was motivated by the fact that present attempts to find *ab initio* an energetically optimal conformation of a protein face the problem of a very large search space. If no constraints are given, it is virtually impossible to

systematically evaluate all valid conformations in order to find the one with the lowest energy. Genetic algorithms have been shown to work efficiently on certain function optimization problems, where the search space was too large for other methods.

The prototype was applied on the *ab initio* prediction of Crambin. The genetic algorithm produced ten conformations, which could be grouped into two classes. Structures within one class are similar in structure but differ substantially from members of the other class. Electrostatic interactions were much higher than in the native conformation. This is likely to result from the six partial charges in Crambin, which were not neutralized. None of the conformations generated are similar to the native conformation. However, all conformations generated by the genetic algorithm show much lower energy than the native structure on the same force field. This means, that the genetic algorithm's search was successful as it produced "good" structures in terms of the fitness function and was *not trapped in one local minimum* but also that the *fitness function was not a good indicator for native structure.* Crambin has a strong amphiphilic character whereas the simple force field used is more suitable for globular, cytosolic proteins. Work has started to improve the model at this point.

In a side chain placement experiment, the backbone of Crambin was held constant in the native state and only side chains were allowed to move. The genetic algorithm produced an alignment of 1.86 Å r.m.s. from the native structure, which is reasonable when compared with other methods.

The results indicate, that in the domain of *ab initio* prediction of protein conformation the genetic algorithm could be an efficient instrument to produce likely candidates for sub-optimal solutions. Certainly, the algorithm cannot do better than the fitness function given to it. It seems therefore possible that with a fitness function that is a good indicator of native structure *ab initio* prediction might become feasible on present day computers.

8. Acknowledgments

This work was supported by the Bundesminister für Forschung und Technologie, grant number BEO 21 / 17405 A.

9. References

[Bernstein *et al*, 1977] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, *The Protein Data*

Bank: A Computer-based Archival File for Macromolecular Structures, Journal of Molecular Biology, **112**, pp. 535-542, 1977

[Branden & Tooze, 1991] C. Branden, J. Tooze, *Introduction to Protein Structure*, Garland Publishing New York, 1991

[Brooks *et al*, 1983] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, *Charmm: A program for Macromolecular Energy, Minimization and Dynamics Calculations*, J. Comp. Chem., vol 4, no 2, pp. 187-217, 1983

[Davis, 1991] L. Davis, (ed.) *Handbook of Genetic Algorithms*, New York, 1991

[Gunsteren & Berendsen, 1990] W. F. Gunsteren, H. J. C. Berendsen, *Computer Simulation of Molecular Dynamics: Methodology, Applications and Perspectives in Chemistry*, Angew. Chem. Int. Ed. Engl., vol 29, pp. 992-1023, 1990

[Hendrickson & Teeter, 1981] W. A. Hendrickson, M. M. Teeter, *Structure of the Hydrophobic Protein Crambin Determined directly from the Anomalous Scattering of Sulphur*, Nature, vol **290**, pp. 107, 1981

[Holland, 1975] J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, 1975

[Le Grand & Merz 1991] S. M. Le Grand, K. M. Merz, *The Application of the Genetic Algorithm to the Minimization of Potential Energy Functions*, submitted to The Journal of Global Optimization, 1991

[Lee & Subbiah, 1991] C. Lee, S. Subbiah, *Prediction of protein side chain conformation by packing optimization*, J. Mol. Biol., no **217**, pp. 373-388, 1991

[Lesk, 1991] A. M. Lesk, *Protein Architecture - A Practical Approach*, IRL Press, 1991

[Lucasius & Kateman, 1989] C. B. Lucasius, G. Kateman, *Application of Genetic Algorithms to Chemometrics*, Proceedings 3rd International Conference on Genetic Algorithms, George Mason University, 1989

[Schulz & Schirmer, 1979] G. E. Schulz, R. H. Schirmer, Principles of Protein Structure, Springer Verlag, 1979

[Tuffery *et al*, 1991] P. Tuffery, C. Etchebest, S. Hazout, R. Lavery, *A new approach to the rapid determination of protein side chain conformations*, J. Biomol. Struct. Dyn., vol 8, no 6, pp. 1267-1289, 1991

[Vinter *et al*, 1987] J. G. Vinter, A. Davis, M. R. Saunders, *Strategic approaches to drug design. An integrated software framework for molecular modelling*, J. Comput.-Aided Mol. Des., no 1, pp. 31-51, 1987