An artificial life model for predicting the tertiary structure of unknown proteins that emulates the folding process

Raffaele Calabretta [§] **Stefano Nolfi*** **Domenico Parisi***

[§] Centro di Studio per la Chimica del Farmaco
Dipartimento di Studi Farmaceutici, University of Rome "La Sapienza"
National Research Council, Piazzale A. Moro 5, 00185 Rome, Italy
tel.: (+39) 6 86 09 02 33

*Institute of Psychology
National Research Council, Viale Marx 15, 00137 Rome, Italy
tel.: (+39) 6 86 09 02 31
fax : (+39) 6 82 47 37

e-mail: raffaele@gracco.irmkant.rm.cnr.it
stefano@kant.irmkant.rm.cnr.it
domenico@gracco.irmkant.rm.cnr.it

Abstract

We present an "ab initio" method that tries to determine the tertiary structure of unknown proteins by modelling the folding process without using potentials extracted from known protein structures. We have been able to obtain appropriate matrices of folding potentials, i.e. 'forces' able to drive the folding process to produce correct tertiary structures, using a genetic algorithm. Some initial simulations that try to simulate the folding process of a fragment of the crambin that results in an alpha-helix, have yielded good results. We discuss some general implications of an Artificial Life approach to protein folding which makes an attempt at simulating the actual folding process rather than just trying to predict its final result.

Keywords: Protein Folding, Genetic Algorithms, Artificial Life.

An artificial life model for predicting the tertiary structure of unknown proteins that emulates the folding process

Raffaele Calabretta [§] Stefano Nolfi* Domenico Parisi*

[§] Centro di Studio per la Chimica del Farmaco
Dipartimento di Studi Farmaceutici, University of Rome "La Sapienza"
National Research Council, Piazzale A. Moro 5, 00185 Rome, Italy
tel.: (+39) 6 86 09 02 33

*Institute of Psychology
National Research Council, Viale Marx 15, 00137 Rome, Italy
tel.: (+39) 6 86 09 02 31
fax : (+39) 6 82 47 37

e-mail: raffaele@gracco.irmkant.rm.cnr.it
stefano@kant.irmkant.rm.cnr.it
domenico@gracco.irmkant.rm.cnr.it

Abstract

We present an "ab initio" method that tries to determine the tertiary structure of unknown proteins by modelling the folding process without using potentials extracted from known protein structures. We have been able to obtain appropriate matrices of folding potentials, i.e. 'forces' able to drive the folding process to produce correct tertiary structures, using a genetic algorithm. Some initial simulations that try to simulate the folding process of a fragment of the crambin that results in an alpha-helix, have yielded good results. We discuss some general implications of an Artificial Life approach to protein folding which makes an attempt at simulating the actual folding process rather than just trying to predict its final result.

Introduction

The prediction of the three-dimensional structure of proteins is a great challenge both for the difficulty of the task and for the importance of the problem. While computational approaches appear to be natural candidates to solve it, optimization techniques that try to predict the result of the folding process by ignoring the specificity of the process itself (Qian & Sejnowski, 1988; Fariselli *et al.*, 1993) have produced limited results. We claim that approaches in the spirit of Artificial Life (Alife) that try to reproduce, even if in extremely simplified ways, the natural processes as they actually occur could be more fruitful.

The protein folding problem presents many similarities with the kind of problems that have been investigated in the Alife literature in the last few years. Proteins, like the simple artificial creatures studied by several researchers in this field (Parisi *et al.*, 1990; Wilson, 1991; Taylor & Jefferson, 1994), are physical entities that have their own structure, which interact with an external environment (the solution), and which are made of sub-components which interact among themselves (the amino acids). In addition, proteins "behave" by folding into a stable structure and such "behaviour" depends on the interaction among the sub-components of the protein itself and between these sub-components and the external environment. Finally, as in most Alife models, to each individual protein corresponds a given fragment of DNA and the mapping between the genetic information and the final stable three-dimensional structure of the protein is very complex and non-linear (Langton, 1992).

The protein folding problem

Many researchers have tried to predict the three-dimensional structure of proteins on the only basis of the amino acid sequence. The attempt has been defined as trying to decipher the second half of the genetic code (Gierasch & King, 1990). Success in this area would be the starting point for new research directions with promising results and possible applications in many fields (biology, genetics, drug-design, etc.).

Proteins chemically consist of the sequencing of structural units which are amino acids: each protein is constructed with the same twenty amino acids which are arranged according to a unique and well defined order. Each protein differs from any other in the number of amino acids linked together (generally between 50 and 3000) and the sequence in which the various amino acids occur. The amino acids are linked to each other by the peptide bond to form a typical linear polypeptide chain. The polypeptide backbone is a repetition of the basic unit common to all amino acids. What changes is the side-chain which is characteristic for each one of the twenty amino acids and is different in shape, bulk and chemical reactivity.

The protein structure can be discussed in terms of three levels of complexity. The primary structure refers simply to the linear amino acid sequence. The secondary structure describes the presence in the protein of regular local structure (alpha-helix and beta-sheets), built with segments of the protein chain. Finally, the tertiary structure represents the real three-dimensional structure of the entire protein. Thanks to the possibility of alternating the twenty amino acids, proteins differ in amino acid sequence (primary structure) and therefore in three-dimensional structure (tertiary structure). In other words, the primary structure of a protein, as it is codified exactly in DNA, contains all the information to determine the three-dimensional structure, on which the function of that protein finally depends. The proteins are necessary macromolecules for the normal deployment of almost all biological processes, but for this to happen it is necessary that the proteins, at the end of a folding process, assume their characteristic spatial structure, which varies from protein to protein. In fact, after the ribosomal biosynthesis of a protein as a linear chain of amino acids, the chain folds up rapidly until it assumes a stable and functional three-dimensional structure. A linear or randomly folded chain would not be biologically active.

On one hand, molecular biology methods have allowed us to identify the amino acid sequence of over 30,000 proteins (Swiss-Prot Data Bank; Bairoch & Boeckmann, 1992). On the other hand, by means of X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR), we have been able to identify the high-resolution structure of only over 1,300 of them (Brookhaven Data Bank; Bernstein *et al.*, 1977). In the next few years the gap is expected to increase due to the great mass of data originated from the Human Genome Project.

Computational approaches to protein folding

Currently there is an increasing interest in the field of computational approaches to protein folding. As Wodak and Rooman (1993) claim, this appears to be due to several factors:

(a) experimental mutagenesis studies have demonstrated that the overall fold of a protein is much more tolerant to sequence modification (Sondek & Shortle, 1990);

(b) analyses of known three-dimensional structures have revealed structural similarity for proteins with different functions (Farber & Petsko, 1990; Kabsch *et al.*, 1990);

(c) the number of known high-resolution protein structures has significantly increased allowing computational models to lie on more solid grounds;

(d) there is an widening gap between the increase in known protein sequences and the lack of information about the structure and function of most of them;

(e) finally, new computational approaches have been developed (Rumelhart & McClelland, 1986; Holland, 1975) that appear to be promising for the protein folding problem and computational power has increased significantly as well.

We will review some of the most significant attempts in this direction and then we will describe our own model.

Extracting knowledge-based potentials

Several researchers have used computational models to design pseudo-energy functions that represent a reduced description of detailed atomic force fields. These pseudo-energy functions or potentials are usually expressed as a sum of several terms and mostly ignore side-chain atomic details.

Examples of such potentials are:

(a) Residue-specific secondary structure propensities (i.e. the tendency of a given residue to fold in a helix, beta sheet or random coil structure; Rost *et al.* ,1994);

(b) Residue-residue potentials (i.e. the tendency of a given residue to end-up close to another one; Maiorov & Crippen, 1992);

(c) Hydrophobicity (tendency of a given residue to interact with water; Casari & Sippl, 1992);

(d) Phi-psi backbone angle probabilities (the probability that two subsequent amino acids can assume a certain relative position; Rooman *et al.*, 1991).

These pseudo-energy function potentials can be derived from known protein tertiary structures by using different computational methods (Statistics, Monte Carlo, Neural Networks, Genetic Algorithm).

The way in which statistics is used to extract potentials is straightforward: the probabilities of observing the parameter of interest are computed and then normalised to correct for sample bias and finally translated into scores (e.g. Bryant & Lawrence, 1993).

Neural networks, given their ability to classify noisy stimuli and generalize to new ones, have also been used to predict the secondary structure of proteins (e.g. Rost & Sander, 1994).

Powerful optimization methods can also be used. Maiorov and Crippen (1992), for example, used an optimization procedure to extract the residue-residue potentials. They derived the strengths of individual contacts starting with non-correct values and then changing such values so that the potential energy of any native structure in the training set would be lower than the potential of any alternative conformation generated from segments of known protein structures.

The extracted function potentials can in turn be used in order to build models which are able to predict the second or the tertiary structure of other sequences (see next paragraph). In other cases, potential extraction and prediction of tertiary structure of unknown sequences can be realized at the same time using a single model.

Application of knowledge-based potentials to prediction of folded structures

The availability of knowledge-based potentials allows us to go beyond the classical approaches based on sequence alignment for predicting secondary and tertiary structure. The main idea is that the extracted potentials can be used to choose between alternative predicted structures by measuring which of them results in lower energy value (e.g. which of them best conforms to the known residue-structure propensity, residue-propensity, hydrophobicity, and virtually to any known potential). In other words, the knowledge based potentials that are extracted from known protein structures can be used to evaluate predicted protein structures.

There are two ways of using knowledge-based potential to predict the tertiary structure of sequences, a hybrid method that combines the classical alignment procedure with the use of potentials and a pure method that use the potential in order to derive the tertiary structure directly from the sequence.

The first approach involves scanning a library of sequences and corresponding known structure motifs in search of compatible sequence-structure combinations, i.e. those which correspond to structures which best conform to the known potentials (see for example Sippl & Weitckus, 1992). In this case potentials are used to choose the best combination of chain folds present in the databases. The combinations of folds that best align with the given sequence and best conform to potentials are selected. This method produces good performance for proteins closely related to those present in the used database but, as the distance increases, performance progressively deteriorates and it becomes unreliable when the sequence identity is lower than 30% (Wodak & Rooman, 1993).

The second approach based only on potentials, by not restricting the space of possible tertiary structure to a known limited set, is much more demanding because it is necessary to assess the value of the potentials of a huge number of possible alternative configurations from which the correct fold needs to be singled out. In this approach an initial wrong structure configuration is chosen and then the structure is progressively modified for a given number of trials until the final configuration, which represents the predicted structure, is obtained. In each trial the actual structure is evaluated by using potentials in order to preserve good modifications (i.e. changes that result in a better configuration from the extracted potential point of view) and to reject bad modifications. The search in the conformation space of a given protein can be implemented by using different algorithms. In particular Monte Carlo (e.g. Godzik *et al.*, 1992) and genetic algorithms (e.g. Dandekar & Argos, 1994) have been used.

In the model of Dandekar and Argos (1994), an initial population of different hypothetical three-dimensional structures for a given sequence are generated. Each individual of the population consists of a vector of dihedral and rotation angles which in turn determines the folding of the main chain of the corresponding protein. Individuals are evaluated according to a set of extracted potentials (secondary structure propensities, presence of hydrogen bonds, hydrophobicity), and ad-hoc criteria (undesired overlapping of C atoms) by determining if and how much a given structure conforms to each potential or criterion. The sum of all these positive and negative contributions constitute the individual's fitness that determines which individuals are allowed to reproduce by generating copies of their vectors with the addition of mutations and combinations between two 'parent' vectors. By repeating this process for a certain number of generations, three-dimensional structures which have better and better fitness and closely resemble the actual folded structure may be obtained.

Emulating the folding process by evolving abstract folding potentials

We think that while using potentials extracted by folded sequences may be adequate in choosing between alternative final structures as is necessary in hybrid approaches that combine folded sub-parts of known protein structures, it may be less useful in "pure" or "ab initio" approaches in which the final folded state is progressively determined through successive modifications starting from the initial amino acid sequence. In fact, the type of conformations that a protein assume during the folding process may differ from the final folded conformation. In other words, it may happen that a structure, in order to reach its final stable state, is forced to pass through a state which even if it does not resemble the final folded state is crucial in order to reach that state. Dandekar and Argos (1994), for example, in order to limit search space, restricted the dihedral and rotation angles to a set of 7 standard conformations extracted from the topology of known folded proteins. However, it is not known whether during the folding process significant different conformations of angles occur.

In our own work, we used an "ab initio" method that does not use pre-extracted potentials and that tries to determine the tertiary structure of unknown proteins by modelling the folding process itself. In other words, we did not want only to predict the final tertiary structure of proteins but we also wished to model the temporal process of folding that results in such a structure. We are aware of the difficulty of the task and of the fact that our results are very preliminary. But we believe that the method can have some validity because a better understanding of the folding process itself, even in the limited case of very short sequences, can have useful results.

For the present time we, as many others (e.g. Lau & Dill, 1990; Unger & Moult, 1993; Šali *et al.*, 1994), have modelled the primary structure of proteins in an extremely simplified way. Amino acid side-chains are represented as spheres connected to the corresponding C_α of backbone with a link of fixed length (see Figure 1); the backbone is represented as a chain of C_α atoms linked by pseudobonds between the C_α atoms of successive amino acid residues (for a similar approach, cf. Oldfield & Hubbard, 1994) .

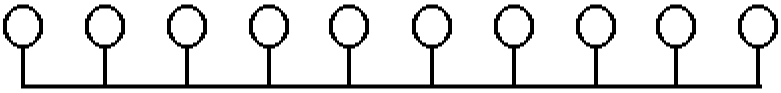


Figure 1. Simulated protein at the beginning of the folding process.

The length of the link between the amino acid side-chain and the backbone is 25A and the pseudobond between two succeeding C_α is 15A (which approximates the average length in real proteins), but can slightly vary during the folding process. Different amino acid side chains all have the same dimensions but can differ in the way they interact with other amino acid side chains and possiby with other substances (e.g. water, but we have not explored this possibility yet). Side chains (spheres) by being attracted or repulsed by other side chains can move in the three-dimensional space. However, in doing so, because of the physical links, amino acid side chains can either (a) rotate around the backbone in the three dimensions modifying the angles between their link and the backbone and/or (b) bend the local portion of the backbone (see Figure 2).

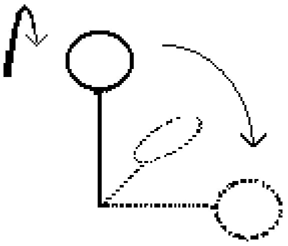


Figure 2a. A side-chain (sphere) rotating around the corresponding C_α.

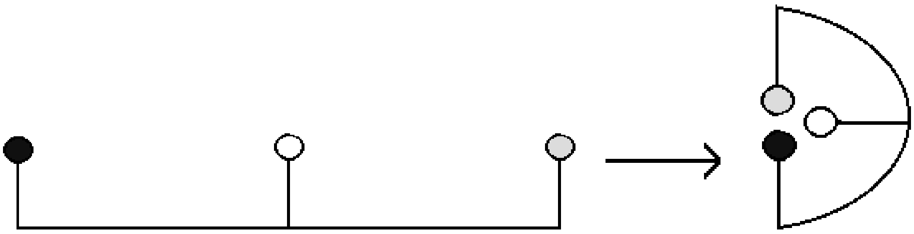


Figure 2b. Side-chains (spheres) that bend the backbone by reciprocal attraction.

A matrix of 20 x 20 values, which were initially randomly specified, determine for each amino acid how much it attracts or repulses other amino acids within a given distance (100A). The attraction or repulsion force is a function of both the value specified in the matrix and of the distance. The process starts with the backbone and the amino acids aligned (see Figure 1) then, depending on the types of amino acids and of the matrix of interaction forces, amino acids start to interact and as a consequence move and fold the backbone. Amino acids are let free to interact for 100 steps. During each cycle, all the interaction forces between neighbouring amino acids (spheres) are computed and then used to move and fold the structure. It is important to notice that while at the beginning of the folding process only amino acids close in the sequence are also close in the three-dimensional space and therefore interact, during the folding process also amino acids distant in the sequence can end up close and start to interact. As a consequence, the final folded structure is the result of the potential interaction of all the amino acids that constitute the sequence.

The problem now is how to determine the matrix of interaction forces in order to emulate the folding process. Once we have obtained a matrix able to fold primary structures into the right tertiary structures we can use such a matrix to predict the tertiary structure of unknown proteins by artificially folding them. For these reasons we can call this matrix of 'forces' *folding potentials*, i.e. potentials that do not extract regularities of known tertiary structures but instead represent 'forces' able to drive the folding process in order to produce correct tertiary structures.

In order to determine such folding potentials we used a genetic algorithm (Holland, 1975; Goldberg, 1989). We started with a population of 100 different matrices of folding potentials randomly generated that represent Generation 0. We then used these potentials to artificially fold proteins with known tertiary structures. In this way we obtained 100 different tertiary structures. The similarity of such tertiary structures with the known right tertiary structure was measured (see below) and used to determine which are the best individuals, i.e. the folding potentials that result in the best tertiary structures. The best 20 individuals were allowed to reproduce by generating 5 offspring each that are copies of the parent matrix of folding rules with the addition of mutations (i.e. random modifications of 10% of the folding potential values). These 20x5 individuals will constitute Generation 1. The process is then repeated for a certain number of generations. The folding potentials of each generation will tend to differ from the previous generation for 2 reasons: because they are the copies of the best individuals of the previous generation and because they receive mutations. Mutations may produce better or less good offspring with respect to the corresponding parents. However, selective reproduction will ensure that only individuals that received good mutations will be able to reproduce.

The evaluation of tertiary structures can be realized in different ways. For each residue segment one can measure the discrepancy of the alpha-carbon bend and torsion angles (see Oldfield & Hubbard, 1994) between the real known tertiary structure and the artificially folded structure produced by each individual matrix of folding potentials. The sum of these discrepancies represent a measure of the error produced by the corresponding folding potentials. Therefore the lower the error is, the higher the probability will be that the corresponding matrix of folding potentials will produce offspring. Alternatively, one can use the sum of discrepancies in distance measured between all combinations of C_α atoms in the three-dimensional space. In our first empirical attempts, the first method appeared to produce better results. In addition, one should decide whether to consider only the discrepancies at the end of the folding process (i.e. after 100 steps) or also as it is taking place. We decided to pay attention to the discrepancies as the folding process is taking place but we weighted the discrepancies at the end of the folding process more in order to force the evolutionary process to select potentials that result in stable folded structures. Hence, the final evaluation of an individual is a weighted sum of the discrepancies throughout the folding process.

In a first attempt to test this model we have tried to simulate the folding process of a fragment of the crambin made of a sequence of 13 amino acids that result in an alpha-helix. We ran 10 simulations starting with different randomly generated folding potentials. As Figure 3 shows, the error, i.e. the discrepancy between artificially folded structures and the real tertiary structure, progressively decreases across generations.

Figure 4 shows six (not immediately) successive stages of the folding process generated by the evolved potentials. A tertiary structure close to the expected one is obtained. In addition, it is interesting to note that in most of the simulations the tertiary structure stabilizes after a certain number of folding steps. How early the folding process reaches a stable stable state could be another component of the 'fitness formula' used to select folding potential matrices for reproduction.

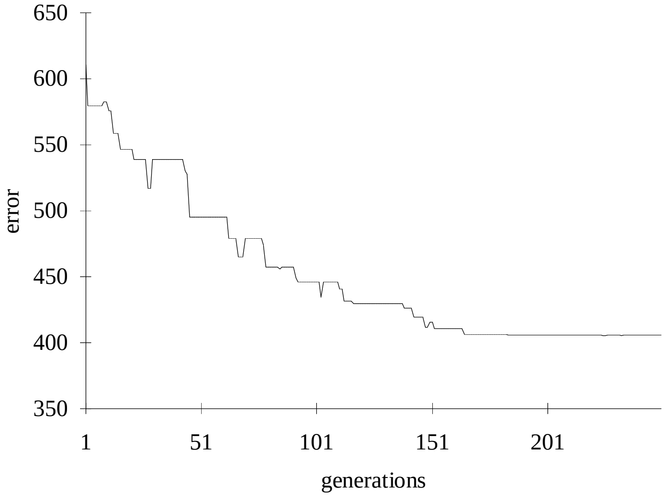


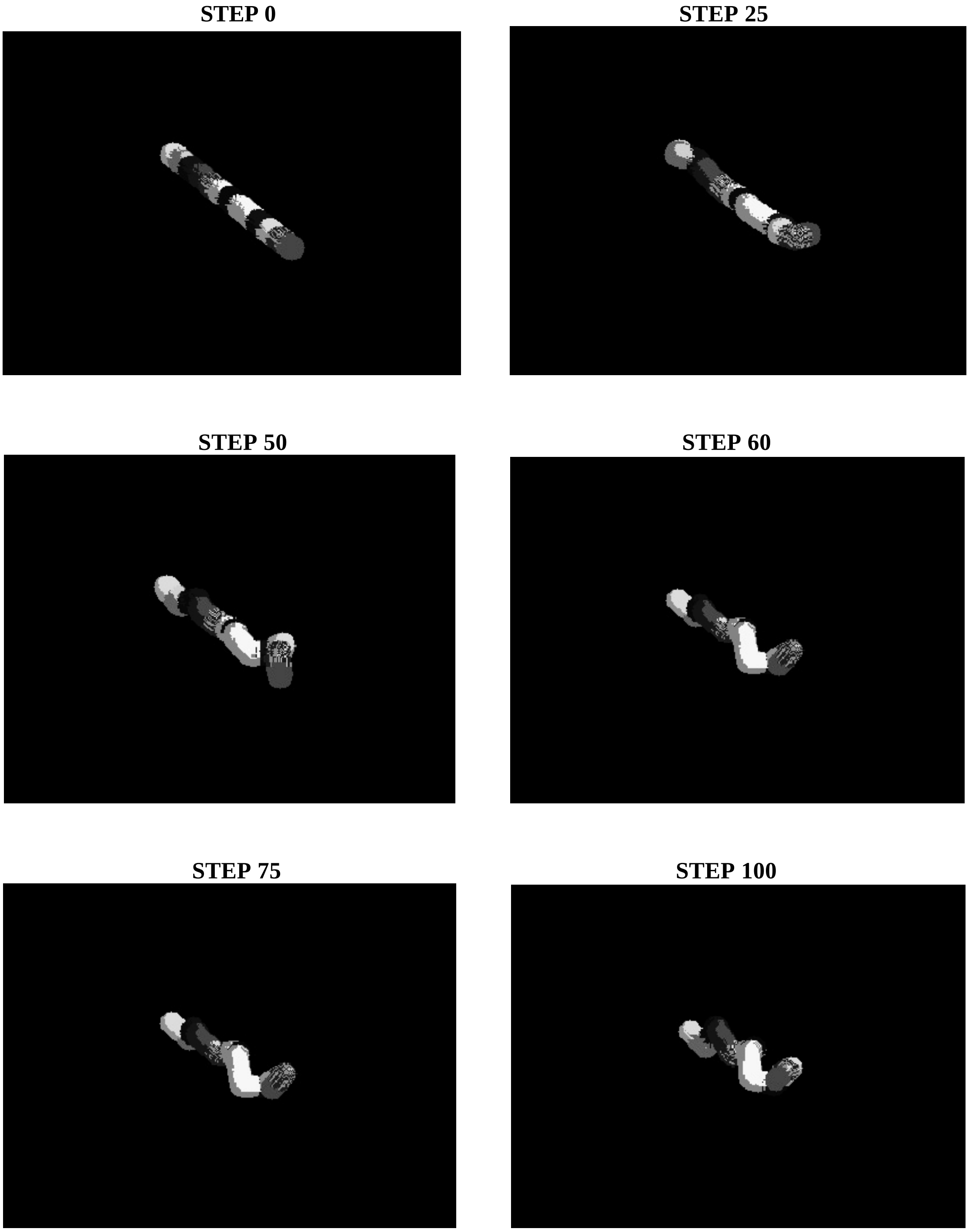
Figure 3. Discrepancies between simulated folded proteins across generations in one of the most successful simulations. For each generation the error of the best individual of the population is shown.

Discussion

Artificial Life is an attempt at understanding all biological phenomena through their reproduction in artificial systems, e.g. computer simulations. More specifically, Artificial Life simulates life phenomena at various levels of biological entities (molecules, cells, organs, organisms, etc.) and tries to understand how phenomena at one level are related to phenomena at other levels. At the same time, Artificial Life is interested in determining similarities and differences in what happens at the various biological levels.

Computational approaches to the protein folding problem are often interpreted as alternative techniques for predicting the tertiary structure of proteins given their amino acid sequence. There is no implication that one is modeling or simulating the actual physico-chemical process that results in a given three-dimensional configuration starting from a linear sequence of amino acids. An Alife approach to protein folding suggests that one should try to model this process. The ability to predict the tertiary structure of unknown proteins should come as a by-product of these modeling efforts.

Assuming that one is modeling the aminoacid sequence-to-tertiary structure mapping process, one can ask potentially useful questions about how this process is related to other biological processes and to their causes. For example, we have used a genetic algorithm to search for appropriate matrices of folding potentials that would give us the correct tertiary structure given an aminoacid sequence. The genetic algorithm can be interpreted either as a search or optimization technique which is only 'inspired' by biological evolution or it can be taken to be a model of biological evolution. For example, when the genetic algorithm is applied to populations of neural networks, by using the genetic algorithm one may want to model the process of evolutionary change in a population of nervous systems, or organisms, or behaviors, etc., and to study such phenomena as the shape of evolutionary change (e.g. gradualilty or punctuated equilibria), evolutionary divergence, speciation, etc. Now, we can ask: When the genetic algorithm is applied to the protein folding problem, are we modeling some actual process of evolution which has taken place (and is taking place) at the molecular level and has shaped the mechanism that maps a linear aminoacid sequence into a three-dimensional structure? Can the population of folding potential matrices be assimilated to a population of genotypes for neural networks?



STEP 0

STEP 25

STEP 50

STEP 60

STEP 75

STEP 100

Figure 4. Folding process of the best individual of the last generation of one of the most successful simulations. For space reasons only 6 of the 100 time steps are shown.

The protein folding process can be viewed as a part of the larger process of mapping from the genotype to the phenotype of an organism which is called development. Can we find similarities between the process of protein folding which results in the 'adult' three-dimensional shape of the protein and the process of development which takes place during the developmental age of a multicellular organism and which result in the adult, mature form of the individual? For example, at the level of the organism all the successive phenotypical forms that are realized during development appear to be subject to an evaluation in terms of fitness. Is this the case also for the successive spatial conformations that a sequence of aminoacid assumes before the final stable conformation? This problem is technically related to the choice of the 'fitness formula' when one is applying the genetic algorithm to the protein folding problem. We have adopted a fitness formula which takes into consideration all intermediate conformations but evaluates them only in function of their degree of approximation to the final conformation. Is this solution correct? We have advanced the hypothesis that during the folding process 'odd' conformations may appear that deviate from the final shape but are useful as stepping stones to arrive to the final shape. In this case a more sophisticated fitness formula would be more appropriate. (Notice that fitness formulae should not be necessarily decided by the researcher but can be viewed as evolvable - or co-evolvable - traits as any other trait; cf. Lund and Parisi, 1994).

In actual proteins the main force that drives the folding process appears to be hydrophobicity (i.e. the aversion for water of nonpolar residues) while Van der Waals interactions (i.e. interactions between dipoles), hydrogen bond interactions (i.e. sharing of an hydrogen atom between to two electronegative atoms), and electrostatics interactions in general appear to play a secondary role (Dill, 1990). However, as Dill states very clearly, driving forces are only half of the story. Another fundamental component that determines the folding process appears to be a opposing forces, e.g. the impossibility that two chain segments simultaneously occupy the same volume of space. Because the folding process involves the collapse of the chain from a large volume to a small one the role of this opposing force appear to be essential.

However, how the driving and the opposing forces produce the known three-dimensional structures is a controversial matter. Dill (1990) claims that any driving force, given the volume constraint (i.e. the fact that two elements cannot occupy the same space) would produce a structure with helices and sheets with hydrophobic interaction as the likely driving force. In fact, he claims that "there are very few possible ways to configure a compact chain, and most of them involve helices and sheet". In other words, he hypothesizes that the tertiary structure drives the secondary structures and not vice versa. Computer simulations appear to support this hypothesis because they show that the amount of secondary structure (helices and sheets) increases as a chain becomes increasingly compact (Dill, 1990). However this hypothesis is in contrast with NMR analysis which appear to suggest that "stable secondary structure first forms the framework necessary for the subsequent formation of the complete tertiary structure" (Udgaonkar & Baldwin, 1988). In addition, it remains to be determined what is the role of the other forces and why is the native structure unique given the fact that hydrophobicity cannot alone determine a unique native structure.

We think that models that reproduce the folding process like the one we have presented in this paper could shed some light on these issues. To pursue our objectives we certainly need to complicate our model by simulating the solution and by allowing the emergence of hydrophobic interactions between amino acids and the solution itself. This could be done by using an additional matrix that specifies for each amino acid the type and the strength of the hydrophobic interaction and by letting the genetic algorithm select the values contained in the matrices. It would then be interesting to observe which type of force will result the dominant one in the simulation, in particular if the hydrophobic forces will outnumber in strength the forces between amino acids.

We also claim that it might be misleading to try to predict the tertiary structures of unknown proteins by using minimization energy techniques based on potentials extracted by folded sequences. In fact, as we have observed, the type of conformations that proteins assume during the folding process may differ from final folded conformations. In addition, this approach requires that native conformations of proteins are at global energy minima (Anfinsen, 1973). But, as Baker & Agard (1994) have observed, "there are good reasons to think that the native states of proteins may not be at global energy minima.....there may be large regions of conformational space that are kinetically inaccessible in which a more stable state might exist". If this hypothesis is true, all computational efforts which try to find the global minimum of a specified potential function would be unable to predict the native state of proteins.

We think that our approach which is not based on minimization of energy but tries to select a set of abstract forces which are able to induce the correct folding may avoid this problem.

References

Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, **181**, 223-230.

Bairoch, A. & Boeckmann, B. (1992). The SWISS-PROT Protein Sequence Data Bank. *Nucleic Acids Res.*, **20**, 2019-2022.

Baker, D. & Agard, D. A. (1994). Kinetics versus Thermodynamics in Protein Folding. *Biochemistry*, **33**, 7505-7509.

Bartel, D. P. & Szostak, J. W. (1993). Isolation of new ribozymes from a large pool of random sequences. *Science*, **261**, 1411-1418.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535-542.

Bryant, S. H. & Lawrence, C. E. (1993). An Empirical Energy Function for Threading Protein Sequence through Folding Motif. *Proteins*, **16**, 92-112.

Casari, G. & Sippl, M. J. (1992). Structure-Derived Hydrophobic Potential. Hydrophobic Potential Derived from X-Ray Structures of Globular Proteins Is Able to Identify Native Folds. *J. Mol. Biol.*, **224**, 725-732.

Dandekar, T. & Argos, P. (1994). Folding the Main Chain of Small Proteins with the Genetic Algorithm. *J. Mol. Biol.*, **236**, 844-861.

Farber, G. K. & Petsko, G. A. (1990). The Evolution of α/β Barrel Enzymes. *Trends Biochem. Sci.*, **15**, 228-234.

Fariselli, P., Compiani, M. & Casadio, R. (1993). Predicting Secondary Structures of Membrane Proteins with Neural Networks. *Eur. Biophys. J.*, **22**, 41-51.

Gierasch, L. M. & King, J. (1990). *Protein Folding: Deciphering the Second Half of the Genetic Code*. American Association for the Advancement of Science, Washington, DC.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Reading, MA.

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

Kabsch, W., Mannherz, H. G., Suck, D., Pai, E. F. & Holmes, K. C. (1990). Atomic Structure of the Actin: DNase I Complex. *Nature*, **347**, 37-44.

Langton, C. G. (1992). Artificial Life. In *1991 Lectures in Complex Systems, SFI Studies in the Sciences of Complexity*, Lect. Vol. IV (Nadel, L. & Stein, D. eds.), Addison-Wesley, Reading, MA.

Lau, K. F. & Dill, A. (1990). Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. Usa*, **87**, 638-642.

Lehman, N. & Joyce, G. F. (1993). Evolution in vitro of an RNA enzyme with altered metal dependence. *Nature*, **361**, 182-185.

Lund, H. H. & Parisi, D. (1994). Simulations with an Evolvable Fitness Formula. *Technical Report* PCIA-1-94, C.N.R., Rome.

Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of Protein Models with Three-Dimensional Profiles. *Nature,* **356**, 83-85.

Maiorov, V. N. & Crippen, G. M. (1992). A Contact Potential that Recognizes the Correct Folding of Globular Proteins. *J. Mol. Biol.,* **227**, 876-888.

Oldfield, T. J. & Hubbard, R. E. (1994). Analysis of C_α Geometry in Protein Structures. *Proteins,* **18**, 324-337.

Parisi, D., Cecconi, F. & Nolfi, S. (1990). Econets: Neural Networks that Learn in a Environment. *Network,* **1**, 149-168.

Qian, N. & Sejnowski, T. J. (1988). Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J. Mol. Biol.,* **202**, 865-884.

Rooman, M. J., Kocher, J-P. A. & Wodak, S. J. (1991). Prediction of Protein Backbone Conformation Based on Seven Structure Assignments: Influence of Local Interactions. *J. Mol. Biol.,* **221**, 961-979.

Rost, B. & Sander, C. (1994). Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure. *Proteins,* **19**, 55-72.

Rost, B., Sander, C. & Schneider, R. (1994). Redefining the Goals of Protein Secondary Structure Prediction. *J. Mol. Biol.,* **235**, 13-26.

Rumelhart, D. E. & McClelland, J. L. (1986). *Parallel Distributed Processing. Explorations In the Microstructure of Cognition*. MIT Press, Cambridge, MA.

Šali, A., Shakhnovich, E. & Karplus, M. (1994). Kinetics of Protein Folding. A Lattice Model Study of the Requirements for Folding to the Native State. *J. Mol. Biol.,* **235**, 1614-1636.

Sippl, M. J. & Weitckus, S. (1992). Detection of Native-Like Models for Amino Acid Sequences of Unknown Three-Dimensional Structure in a Data Base of Known Protein Conformations. *Proteins,* **13**, 258-271.

Sondek, J. & Shortle, D. (1990). Accommodation of Single Amino Acid Insertions by the Native State of Staphylococcal Nuclease. *Proteins,* **7**, 299-305.

Taylor, C. & Jefferson, D. (1994). Artificial Life as a Tool for Biological Inquiry. *Artificial Life,* **1**, 1-13.

Udgaonkar, J. B. & Baldwin, R. L. (1988). NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A. *Nature*, **335**, 694-699.

Unger, R. & Moult, J. (1993). Genetic Algorithms for Protein Folding Simulations. *J. Mol. Biol.,* **231**, 75-81.

Wilson, S. W. (1991). The Animat Path to AI. In *From animals to animats:Proceedings of the First International Conference on Simulation of Adaptive Behavior* (Meyer, J.-A. and Wilson, S. W., eds), pp. 15-21, MIT Press, Cambridge, MA.

Wodak, S. J. & Rooman, M. J. (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol.,* **3**, 247-259.