Journal of Computational Intelligence in Bioinformatics.
ISSN: 0973-385X Volume 1 Number 1 (2008) pp. 35–43

http://www.ripublication.com/jcib.htm

# Enhancement of Protein Folding Problem using a Genetic Algorithm with Introns

**[1]M.V. Judy, [2]K.S. Ravichnadran and [3]K. Murugesan**

*[1]SOC, SASTRA University, Thanjavur-613402, India*
*E-mail : judynair@mca.sastra.edu*
*[2]SOC,SASTRA University, Thanjavur-613402, India*
*E-mail : raviks@it.sastra.edu*
*[3]Department of mathematics, National Institute of Technology, Trichy*

**Abstract**

Genetic algorithms have proved to be a successful method for predicting the protein structure. In this paper, we report experiments demonstrating that the insertion of introns (non functional sequences of DNA) into bit strings can lead to dramatic increase in the success rate of the genetic algorithm in finding a solution to the protein folding problem. We have demonstrated the superiority of an intron based genetic algorithm on several instances of the protein-folding problem, which not only finds the optimum solution, but also finds them faster than the traditional approach.

## Introduction

Proteins fold rapidly and reliably to their functional state (native state). Whether the native state is kinetically or thermodynamically controlled remains an open question. The native state can therefore be the global energy minimum or a low-lying Meta stable conformer. The energy hyper-surface has high dimensionality and complexity. The primary structure of a protein is the amino acid sequence of its polypeptide chain, while the secondary structure is the local arrangement of a polypeptide's backbone atoms without regard to the conformations of its side chains. Under certain physiological conditions, the primary structure of a protein spontaneously folds into a precise three-dimensional form called its tertiary structure or native state that determines its functional properties. Finding energetically low lying conformations given a sequence of amino acids is termed as "The Protein Folding Problem"[1]. With the completion of the human genome project, new information about the human genome is readily available. But only a small percentage of these sequences have known native states. Efforts aimed at solving the Protein Folding Problem have

involved the optimization of a potential energy function that approximates the thermo dynamic state of a protein macromolecule. Since an algorithm using such a potential function gives insight into how a protein folds, these approaches are also known as Protein Structure Prediction. Human genome is comprised of so-called "junk" DNA, which occurs both in and between functional genes. "Junk" DNA present within genes occurs in short stretches known as introns.

More formally, introns are non-coding section of DNA that occur within functional genes.The model of the second genetic amino acid interaction code has given a possibility to elaborate new methodologies for studies of gene and intron emergence mechanisms, based on comparative amino acid codon root analysis (CAACRA). Codon roots – the second codon letters – are much more conservative and less changed during evolution than amino acids[2]. Twenty natural amino acids determined by the genetic code can be subdivided into four groups of the so called common-root amino acids having identical second codon letters C, G, A and U(T).

Natural selection accepts amino acid exchanges in proteins (as a result of point mutations) mainly between the common-root amino acids, indicating that such amino acids are potentially tantamount[2]. During evolution, as a result of mutations, amino acids in protein structures may change time and time again, maintaining in many cases the same codon Formation of only symmetric (0,0) exons follows that during evolution introns can slide and be gradually eliminated, as introns of natural genes in many cases are outside of the gene knot points and have changed phases (Fig. 1).

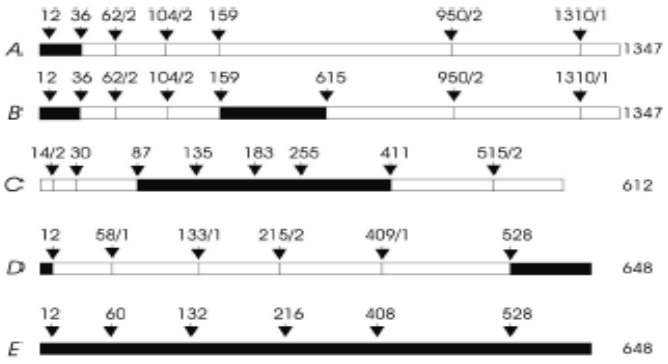



**Figure 1:** Regularity of dimensions of intron and exon maps demonstrate the mechanisms of gene emergence by nucleotide multiplication reactions. A and B, intron maps of ß-tubulin genes of the *Aspergillus parasiticus* and *Aspergillus nidulans*. Intron positions are shown as arrows topped with intron co-ordinates/phases (only in cases when the phase differs from zero). Exons containing the whole number of repeat units (nx12nt) are shown as black parts of ribbons. The length of genes coding parts (exon rows, including a stop codon) is shown beside the maps. C, intron map of the green alga *Volvox carteri* gene yptV1. D and E, intron maps of the *Coprinus cinereus* ras gene before and after restoration of intron phases to phase zero (i.e. by changes of intron co-ordinates by ±1 or ±2nt).

Translating gene exon row nucleotide sequences or protein amino acid sequences to more conservative codon root sequences in separate cases makes it possible to

demonstrate the retained ancestral regularity of genes and proteins. In this study protein conformations and intra molecular interactions are modeled using the simplified HP Bead Model[3] with introns. A genetic algorithm with a modified keep best reproduction selection strategy[6] has been employed to search the hyper-surface. The simulated structure have introns but in the real biological process it is eliminated.

## The Hp Model with Introns

The hydrophobic-hydrophilic model (HP model) by Dill [1] is a simple abstraction that captures the essence of the important concepts of Protein Structure Prediction. In the HP model, amino acids are divided into two categories: hydrophobic (H) and hydrophilic (P). We include the introns (I) into the sequence. The primary sequence of a protein is therefore $S \in (H, P, I)^+$. Using this simplification, optimization models can be developed that seek to maximize interactions between adjacent pairs of hydrophobic amino acids (or hydrophobes). Adjacency is considered only in the cardinal directions of a lattice upon which the sequence is embedded. In an HPI lattice, vertices represent amino acids and edges represent connecting bonds. Black squares at the vertices indicate hydrophobic amino acids, white squares indicate hydrophilic amino acids and white squares with **'i'** marked in the center of the square is an intron .A lattice can be two or three dimensional, and either square, cubic or triangular. The hydrophobic-hydrophobic (HH) contacts are the basis for the evaluation function. Every pair of hydrophobes that are adjacent on the lattice and not consecutive in the primary sequence is awarded a value ε (usually –1). The sum of all such values gives the energy of the conformation. The amino acid sequence is "folded" on a two-Dimensional square lattice on which at each point, the chain can turn 90" left or right, or continue ahead. Figure 1 shows a 40-length sequence embedded on a square lattice, HH contacts indicated by dotted arrows.

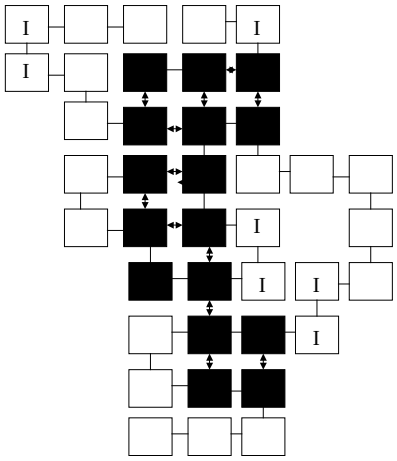



**Figure 2:** HP**i** sequence of length 40 on a square lattice with energy –12

## Implementation Using Genetic Algorithm With Introns

### Genetic Algorithm

Our implementation of the genetic algorithm is same as that described in Unger and Moult [8] but we insert introns to improve the efficiency of the algorithm The solutions are not encoded as binary strings but rather are the conformations themselves, which are treated directly in the spirit of genetic operators. The process starts with V extended structures. In each generation each structure is subject to a number of mutation steps with rate ranging from 0.01 to 0.20. Each mutation is the same as a single Monte Carlo (MC) step [8] and is subject to similar acceptance criteria as in a MC process. At the end of this MC stage [10], the crossover operation is performed. The chance $p\ (S_i)$ of a structure being selected for crossover is proportional to its energy value *Ei, That is*

$$p(S_i) = \frac{E_i}{\sum_{j=1}^{N} E_j},$$

Thus, the lower energy conformations have a higher chance of being selected. For a pair of selected structures *a* random point is chosen along the sequence and the X-terminal portion of the first structure is connected to the C-terminal portion of the second structure (see Fig. 2).As there are three ways to join the parts together (connecting the chains with angles of 0", 90" or 270"), these possibilities are tested in a random order to find one that is valid ( That is, where no residue from one structure occupies a lattice point used by a residue from the other). If none of the three ways led to a self-avoiding structure, then another pair of structures is selected. Once a valid structure *Sk* is created, its energy $E_K$, is evaluated and compared to the averaged energy Eij = *(Ei+ Ej)/2* of its "parents"[7,8]. The structure is accepted if $E_k <= E_{ij}$, , or if the energy will be increased based on the decision:

$$Rnd < \exp\left[\frac{\bar{E}_{ij} - E_k}{c_k}\right].$$

This crossover operation is repeated until *N* - 1 newly accepted hybrid structures have been constructed to constitute the population of the next generation.

The process starts with a population of fully extended structures. Each structure undergoes a MC stage followed by a crossover stage. In the crossover stage. pairs of structures are randomly (based on their energies) cut and pasted. In this example the cut point was randomly chosen to be after residue 14. Joining the first 14 residues of (A) with the last 6 residues of (B) and applying a randomly chosen 270 $^{o}$ rotation at the joint achieves the compact structure in (C). In this case, the energy value of the hybrid (C) is -9, lower than the energies **-5** and **-2** of its "parents". The hybrid is always accepted if its energy is lower than the averaged energies of its parents. or non deterministically accepted according to its energy increase. The selection strategy

involved is the modified keep best reproduction strategy in order to preserve the good genetic material after every generation.

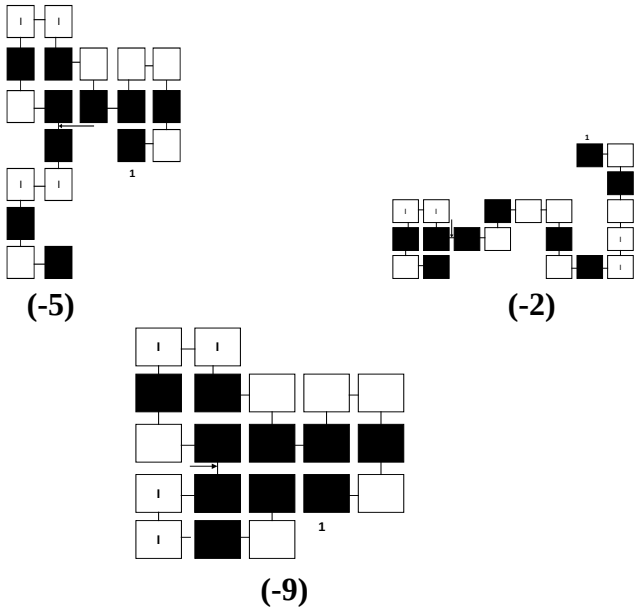



**Figure 3:** The crossover operation in HPI model

**Significance of introns in GA**

Selection should be done as a conservative force as opposed to an agent of change. Among the evidence of this view is that the mutation rate of exons (regions of chromosomes that has functional values)[9] exceeds the mutation rate of introns (regions that are not expressed or intragenic regions). This is because mutations in exons tend to be selected against (since they are maladaptive) whereas mutations in introns are unaffected by selection. Introns protect good building blocks against destructive cross over. Introns affect the survivability of the organism indirectly[10].

Two features which is used to define introns :

i) An intron is a feature of the genotype that emerges from the process if the evolution of variable length structures

An intron does not directly affect the survivability of the GA individual.

The sequence of length 36 without introns and length 40 with introns.

S1=>PPPHHPPHHPPPPPHHHHHHHPPHHPPPPHHPPHPP

S2=>PPPHHPPHHiiPPPPPHHHHiiHHHPPHHPPiiPPHHPPHPP

**The Beneficial Effects of introns**

Introns had differing effects before and after exponential growth of the introns. They promote parsimonious effective solutions and structural protection against destructive crossover .It protects the entire individual from destructive effects of crossover[11,12]. It is implemented when crossover swaps groups of introns that are,

effectively terminals. Structural protection is very beneficial when it allowed building blocks to emerge despite the destructive effects of crossover and mutation.

## Results and Comparison

Genetic Algorithm with introns (GAI) not only found the optimal results but it also converged to optimum conformations in lesser number of energy evaluations.

The experiment was done on the standard amino acid sequence of length 20,24,25,36,48,50,60,64 with a population size of 200. The mutation rate was 2% to 15%. Each application of a genetic operator is counted as a step. Thus, a generation takes 10 X population size times mutation steps plus the number of crossover trials it take. When a valid conformation is encountered, its energy is evaluated. At the end of each generation the chromosome with worst fitness is replaced with the best parent.. It is observed that the GA with introns finds optimum solution with minimum number of energy evaluations. The introns were ruled out as the protein structure evolved as in the biological process. The significance of introns is its presence itself. Eight sequences of amino acids used for testing is given below. Introns was inserted at various locations randomly.

(20) + (6) HPHPPiiHHPHPiiPHPHHiiPPHPH
(24) + (4)HHPPHPPHPPiiHPPHPPHPPHiiiiPPHH
(25) + (6)PPHPPHHPPPiiiPHHPPPPHHPiiiPPPHH
(48)+(10)PPHPPHHPiiPHHPPPPPiiHHHHHHHHHiiHHPPPPPPiiHHPPHHPPiiHPPHHHHH
(50)+(10)HHPHPHPHPHiiHHHPHPPPHPiiPPHPPPPHPPiiPHPPPHPHHHiiHPHPHPHiiPHH
(60)+(15)PPHHHPHHHHiiiiiHHHHPPPHHHiiiiiHHHHHHHHPHPiiiiiPPHHHHHHHHHHHHHPPPPHHHHHHHPHHPHP
(64)+(15)HHHHHHHHHHiiiiiHHPHPHPPHHiiiiiPPHHPPHPPHHiiiiiPPHHPPHPPHHPPHHPPHPHPHHHHHHHHHHHH

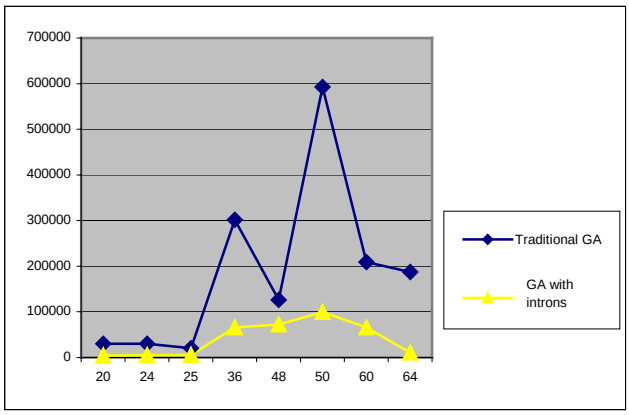



**Figure 4:** The x-axis shows the sequence number while the y-axis shows the number of energy evaluations

The simulation was run for 100 times with different parameter setting. The optimal conformation found for each combination is tabulated along with the mean performance. A comparison of the number of energy evaluations of the other methods is also given

The chart shows the variation in performance of the compared methods (Figure 5). We use higher mutation rates in combination with crossover. The x-axis shows the length of sequence while the y-axis shows the number of energy evaluations The population size was 200. With Introns we were able to speed up the convergence of the GA by using higher mutation rates . MKBR only keeps the best parent and eliminates the worst child. Introns avoids the destructive effects of crossover

## Conclusion

The genetic algorithm with introns outperforms the standard GA with generational replacement technique significantly on protein folding problems, especially as the problem size increases in terms of time and optimality.

It may seem bizarre that extraneous positions in a bit string should increase the efficiency of an algorithm by a considerable factor but the action of crossover operator makes the mechanism apparent. Even though introns are not involved in the manufacture of proteins, they play the role in protecting good building blocks against destructive crossover.

## Further Research and Discussion

This paper describes a research on a GA based system with introns for protein folding problem. . It is very important to note that introns are not functional units of DNA and therefore will not exist in the protein. It is spliced out before translation. We have inserted introns only to accelerate the success rate of genetic algorithm in the energy evaluation.

It helps in the faster convergence of minimum energy. This nomenclature may seem contradictory with the real biological scenario, but it has given remarkable result in the case of simulating the protein folding. Introns may have differing effects before and after exponential growth of introns begins.

Different systems may generate different types of introns with different probabilities.

The extent to which genetic operators are destructive in their effect is likely to be a very important initial condition in intron growth. Mutation and crossover may affect different types of introns differently.

Even though there may be some drawbacks like more memory usage or stagnation, introns provides structural protection against crossover to the building blocks during the earlier stages of evolution.

We need to test larger amino acid sequences. We are confident however, that on these problems we will still be able to get better results than the traditional genetic algorithm.

**Table 1:** Energy evaluations for protein folding problem

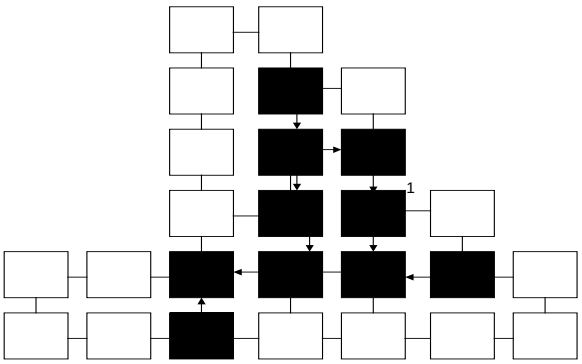
| Length | Optimal Energy | Traditional GA | GA with Introns |
|---|---|---|---|
| 20 | -9 | 30492 | 4230 |
| 24 | -9 | 30491 | 4823 |
| 25 | -8 | 20400 | 5432 |
| 36 | -14 | 301339 | 66222 |
| 48 | -22 | 126547 | 72541 |
| 50 | -21 | 592887 | 99877 |
| 60 | -34 | 208781 | 65888 |
| 64 | -42 | 187393 | 10765 |

## Appendix-A

The sample output of the actual structure of protein got through simulation for the amino acid sequence of length 26 along with introns is given below for the sequence with introns (HPHPPiiiPPPHPiiiPPPHPiiiHHPPPiiiPHPH ).

| | | | | | | |
|---|---|---|---|---|---|---|
| | | 22p | 23p | | | |
| | | 21p | 24h | 25p | | |
| | | 20p | 19h | 26h | | |
| | | 17p | 18h | 1h | 2p | |
| 14p | 15p | 16h | 9h | 8h | 3h | 4p |
| 13p | 12p | 11h | 10p | 7p | 6p | 5p |

The HP square lattice is constructed as the data simulated in the table and the structure is constructed as in fig given below.




The introns are removed when the proteins are evolved. Only the coding region of DNA contribute to the structure of protein. The hydrogen bond corresponding to energy is indicated with pointed arrow.

## References


[1] Dill, K. A. Theory for the folding and stability of globular proteins. Biochemistry, 24, 6 (March 12, 1985), 1501-1509.

[2] Nora Ievina, Gunars Chipens A new approach to study the origin of genes and introns. Acta Universitatis Latviensis, 2003, Vol. 662, pp. 67–79

[3] Unger, R., and Moult, J. A genetic algorithm for three dimensional protein folding simulations. In Proceedings of the fifth international conference on genetic algorithms (ICGA ‘93) (Urbana-Champaign, IL, 17-21 July, 1993).

[4] R. Aroul Selvam and Rajkumar Sasidharan1,Nucleic Acids Research, 2004, Vol. 32, Database issue D193±D195 DOI: 10.1093/nar/gkh047DomIns: a web resource for domain insertions in known protein structures

[5] Inserting Introns Improves Genetic Algorithm Success Rate: Taking a Cue from Biology (1991) James R. Levenick Proceedings of the Fourth International Conference on Genetic Algorithms

[6] Yu, Jun, Yang, Zhiyong, Kibukawa, Miho, Paddock, Marcia, Passey, Douglas A., and Wong, Gane Ka-Shu. (2002) Minimal introns are not "junk." Genome Research, 12, 1185- 189.

[7] Kelsey Byers Biotechnology Academy April 17, 2003An Evaluation of Intron Significance Using Bioinformatics Senior Project Submitted by

[8] K. A. De Jong, “Analysis of the Behavior of a Class of Genetic Adaptive Systems,” Ph.D. Dissertation, The University of Michigan, Ann Arbor, M1, 1975.

[9] Clayton Matthew Johnson, Anitha Katikireddy, “A Genetic Algorithm with Backtracking for Protein structure Prediction. GECCO’06,July 8-12, 2006, Seatle, Washington, USA.

[10] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Reading, MA: Addison-Wesley, 1989.

[11] M.V.Judy and K.S.Ravichandran, ” A soltion to protein folding problem using a modified keep best reproduction strategy”, Proceedings of the CEC2007, IEEE Congress on Evoltionary Computation.

[12] Wiese K. and Goodwin, S.D., “Parallel Genetic Algorithms for Constrained Ordering Problems”, Proceedings of the 1lth International Florida Artificial Intelligence Research Symposium, FLAIRS’98, pp. lOl-105,1998.

[13] Wiese, K. and Goodwin, S.D., “The Effect of Genetic Operator Probabilities and Selection Strategies on the Performance of a Genetic Algorithm”, Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence Vol. 1418, Springer Verlag, Germany, pp. 141-153, 1998.