Improving Protein Structure Prediction by New Strategies: Experimental Insights and the Genetic Algorithm$

Thomas Dandekar

EMBL, Postfach 102209, D-69012 Heidelberg; Phone +49-6221-387-372 (dandekar@embl-heidelberg.de)

Abstract

Three different approaches to improve tertiary fold prediction using the genetic algorithm are discussed: (i) Refinement of the search strategy, (ii) combination of prediction and experiment and (iii) inclusion of experimental data as selection criteria into the genetic algorithm. Examples from our current work are presented for refined strategies against crowding in solution space, definition of domain boundaries and secondary structure in combination with experiment, and direct incorporation of experimentally known distance constraints into the fitness function.

Keywords: Genetic algorithm, Fold prediction, Disulfide bonds, Crowding

Introduction

We study the genetic algorithm (GA) [1] for protein structure prediction as it has a large application potential [2]. The field, including the advantages of GA in the large combinatorial space of protein folding is reviewed in [3,4]. With our approach we could achieve tertiary fold prediction from sequence and predicted secondary structure for four helix bundles (RMSD around 6 Å,[5]). Further, exploiting available secondary structure information (DSSP) the fold for 19 different protein topologies was succesfully delineated (proteins less than 100 amino acids in size, not more than eight secondary structure elements, RMSD around 4.5-5.5 Å on average; [6]).

Though promising, these results as well as efforts from other groups including blind tests [7], predictions of small helical and strand containing proteins [8] and peptide libraby

assemblies for fold prediction [9] illustrate at the same time that despite the advantages of the GA, protein structure prediction is still a challenge.

Thus we currently explore prediction improvement by refining the search strategy,combining the GA prediction with experimental data for a more complete picture of the protein and feeding experimental information as additional fitness criteria directly into the model prediction made by the GA.

Methods

Protein folding simulations are achieved modeling the mainchain using internal coordinates and standard conformations [5]. Starting from random structures and known or predicted secondary structure, the fitness function selected for growth and cooperativity in the secondary structure, global and hydrophobic packing, clash free structures and promoted formation of beta-strand regions. Details of the genetic algorithm and the fitness function used are described in

[6]. The simplified hp standard protein model is used in the first part of our results.It considers only two types of amino acids,hydrophobic and hydrophilic, details are described in [10].

Results and Discussion

Development of improved search strategies

Starting from the simple genetic algorithm [1] different parameters such as mutation rate, number of cross-over sites and population size are examined. However, focus is on modified search strategies which should improve searching for the correct protein structure by the genetic algorithm. Different crowding and elitist searching strategies are examined to keep the population rich and diverse while not compromising effective searching and convergence. To test these efficiently, simulations are run in the context of square lattice hp protein models in two dimensions. Similar versions in three dimensions are also investigated.

Table 1 shows a search strategy against crowding found to be promising, "pioneer search", which was advantageous in comparison to our standard simulations under most of the parameter conditions tested and will be examined in detail further including its performance in more complex models.

The strategy "pioneers" new search space every ten generations. At this check point, all individuals of the new generation have to be different from any individual of the previous generation to search for new protein structures during the run. Individuals which are identical to previous ones are discarded and two new parents are randomly selected according to fitness (using normal roulette-wheel parent selection) from the old population to generate an alternative individual which again is only accepted if it is different from the previous population.

Several variations of this scheme were also tested, for instance "incest" (replace the individual which is found to be identical to previous generation by one from remating with the fitter parent) or "promiscuity" which accepts the result of any random mating within the previous population as a replacement without another check but were less successful in this comparison.

Combining experimental approaches and GA model prediction

Independent experimental approaches can advantageously complement the GA prediction. They can be exploited to rule out wrong structures and to point to domains where the GA prediction should concentrate on. Such combined predictions depend heavily on a close interaction between experiment and molecular model. Rejecting predictions by conflicting experimental data is in many parts still non automated and thus includes some element of subjectivity. However, violation of experimentally measured and known sec-

Table 1. *Results for a 20 residue chain hp model simulation in 2D comparing the genetic algorithm in standard form (Goldberg, 1989) versus its improvement by an anti-crowding strategy (see text). Given is the amount of times the global energy minimum conformation was found in ten simulations under various conditions of population size (popsize),mutation rate (pmutation) and cross-over (pcross).*

popsize	pmut	pcross	GA unmodified	GA PS [a]
200	0.05	0.2	6	10
200	0.1	0.2	9	9
200	0.15	0.2	3	2
200	0.05	0.5	5	8
200	0.1	0.5	9	9
200	0.15	0.5	2	2
200	0.05	0.8	5	9
200	0.1	0.8	6	8
200	0.15	0.8	3	2
400	0.05	0.2	8	10
400	0.1	0.2	8	9
400	0.15	0.2	4	2
400	0.05	0.5	7	10
400	0.1	0.5	10	7
400	0.15	0.5	5	2
400	0.05	0.8	6	10
400	0.1	0.8	8	7
400	0.15	0.8	2	2

[a] pioneer search

ondary structure content or violation of known distance constraints by model predictions can be objectively measured and quantified and were useful to reject wrong predictions by these criteria.

A case in point is a strongly hydrophobic viral protein (EP5) which hampers accumulation of sufficient quantities for crystallization. Spectroscopic data indicate a low amount of helix content and help to define domainal boundaries. Figure 1 shows our current model which best fits the experimental data accumulated so far. Further studies will aim to verify the different topological vicinities predicted by the genetic algorithm model.

Direct incorporation of experimental data as additional fitness criteria

Experimental data can be even more tightly connected to the GA model if used as additional fitness criteria. A systematic
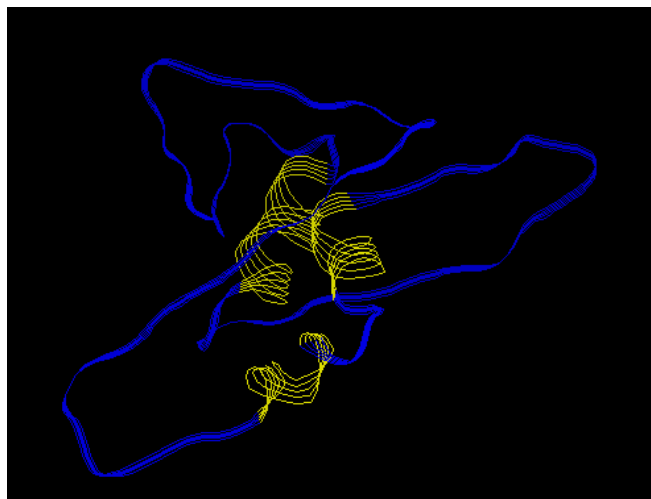
Figure 1. *Viral hydrophobic protein EP5. Result of the genetic algorithm simulation. The low helical content and overall shape agree well with the available experimental data collected so far, further tests will examine the details of the topological vicinities predicted. The main chain trace is shown in blue, helical regions are shown in yellow.*

Figure 2. *Predicted structure for a non pathogenic amoebapore peptide. Using standard secondary structure prediction, RMSD error on four helical bundles with known crystal structure was found to be around 6 Å [5].*

Supplementary material: 3D coordinates of the structures shown in Figure 1 and 2 are available as PDB-file.

exploration of the effect of experimental information on prediction accuracy on smaller protein structures and domains, utilizing readily available experimental data such as cross-linking data, iron-sulfur clusters, chelating residues, core residues and important catalytic residues is currently conducted.

In this context we also examine disulfide bonds and their effect on improving protein structure resolution. Different potentials were tested. The square root of the sum of the distance square deviation from the optimal distance yielded best results in test runs on proteins with known three dimensional structure and exploiting a known connectivity of disulfide bonds.

Further, including this fitness parameter, predictions for pathogenic amoebapore proteins and mammalian NK-lysin could be obtained [11]. Figure 2 shows the structure of a related non pathogenic amoebapore protein as currently predicted based on our genetic algorithm approach. The somewhat loosened up bundle structure could be implicated in its non pathogenicity. This result will be followed up by further simulations and comparisons. Different non pathogenic amoebapore sequences are studied together with experimental tests to reveal and understand differences to pathogenic amoebapores.

Outlook

Future work will extend our efforts in the three areas of research presented. By combining the advantages of each we also want to tackle larger protein structures.
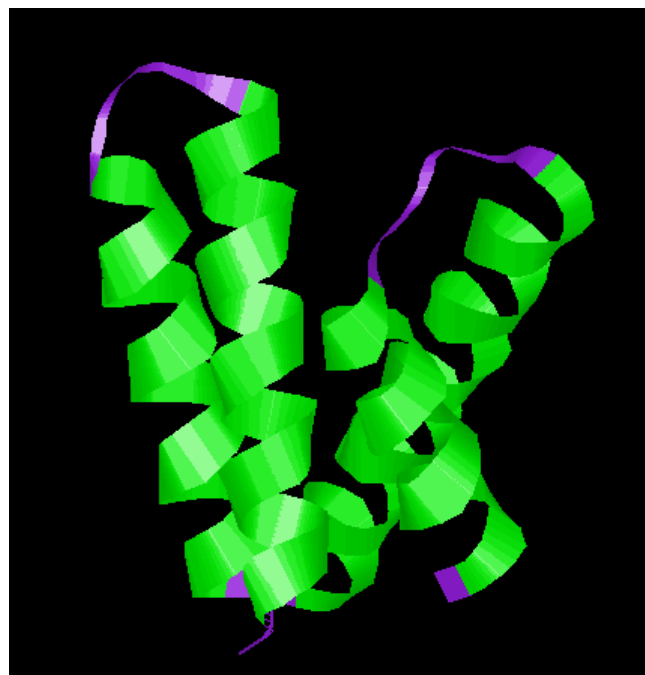
References

1. Goldberg, D. *Genetic algorithms in search, optimization and machine learning*; Addison Wesley: Reading, Massachusetts, 1989.
2. Dandekar, T.; Argos, P. *Protein Eng.* **1992**, *5*, 637.
3. Clark, D.E.; Westhead, D.R. *J.Comp.-Aided Mol. Design* **1996**, *10*, 337.
4. Pedersen, J.T.; Moult, J. *Curr.Op.Struc.Biol.* **1996**, *6*, 227.
5. Dandekar, T.; Argos, P. *J.Mol.Biol.* **1994**, *236*, 844.
6. Dandekar, T.; Argos, P. *J.Mol.Biol.* **1996**, *256*, 645.
7. Pedersen, J.T.; Moult, J. *Proteins* **1995**, *23*, 454.
8. Sun, S.; Thomas, P.D.; Dill, K.A. *Protein Eng.* **1995**, *8*, 769.
9. Bowie, J.U.; Eisenberg, D. *Proc.Natl.Acad.Sci. USA* **1994**, *91*, 4436.
10. Unger, R.; Moult, J. *J.Mol.Biol.* **1993**, *231*, 75.
11. Dandekar, T.; Leippe, M. *Folding and Design* **1997**, *2*, 47.