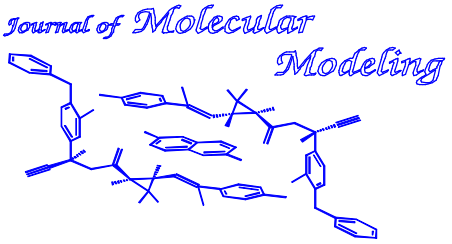
J. Mol. Model. **1996**, *2*, 304 – 306



The Genetic Algorithm Applied as a Modelling Tool to Predict the Fold of Small Proteins with Different Topologies

Thomas Dandekar

Europäisches Laboratorium für Molekularbiologie, Postfach 102209, D-69012 Heidelberg, Germany
(Thomas.Dandekar@Mailserver.EMBL-heidelberg.DE)

Received: 15 May 1996 / Accepted: 6 August 1996 / Published: 27 September 1996

Abstract and Introduction

The genetic algorithm exploits the principles of natural evolution. Solution trials are evolved by mutation, recombination and selection until they achieve near optimal solutions [1].
Our own approach has now been developed [2] after a general overview on the application potential for protein structure analysis [3] to a tool to delineate the three-dimensional topology for the mainchain of small proteins [4], no matter whether they are largely helical, are mixed or β-strand rich [5].
Results on several protein examples for these different modelling tasks are presented and compared with the experimentally observed structures (RMSDs are around 4.5-5.5 Å). To start a modelling trial only the protein sequence and knowledge of its secondary structure is required. The fittest folds obtained after the evolution at the end of the simulations yield the three dimensional models of the fold. Current limitations are protein size (generally less than 100 aminoacids), number of secondary structure elements [7-8] and irregular topologies (e.g. ferridoxins).
Further, preliminary results from current simulations are illustrated. We now want to apply simple experimental or other information, which is available long before the three-dimensional structure of the protein becomes known, to refine the modelling of the protein fold and tackle also more difficult modelling examples by our tool.

Keywords: Genetic algorithm, protein structure analysis, 3D topology

Methods

To achieve protein structures close to observed starting from a population of random structures, selection of structures according to basic protein building principles is applied. They are briefly summarized in the following table (Table 1; for additional details see [2],[4],[5]) and focus around global and hydrophobic packing, avoiding clashes, stabilization of secondary structure (used as input, to avoid bias from bad prediction the DSSP assignment in many trials, but similar trials are also run using standard secondary structure prediction methods) and the build up of strands and sheets.

Results and discussion

With these criteria, we can model in our simulations the main chain topology of a number of different protein structures [5]. Table 2 gives two examples for each of the categories helical, mixed and strand.

Table 1. *Fitness function criteria*

criteria [a]	**des [b]**	**term**	**specific parameters**
constant [c]	C	$weight_C$	adjusted to 10% negative fitness in the first generation
clash	cl	$weight_{cl} \cdot \Sigma$ overlap [d]	$weight_{cl}$ = -500
secondary structure(ss):			
	pf	$weight_{ss} \cdot$ structural preference [e]	
	co	$weight_{ss} \cdot$ cooperativity [f]	$weight_{ss}$ = +12
tertiary structure: global scatter (gs)			
	gs	$weight_{sc} \cdot$ scatter(sc) [g]	$weight_{sc}$ = -24
hydrophobic scatter (hs)			
	hs	$weight_{hd} \cdot$ hydrophobic distribution(hd) [h]	$weight_{hd}$ = -19 hydrophobic residues include: Phe,Tyr, Met,Cys, Ile,Leu,Val,Trp
β-strand criteria [i]:			
hydrogen bond	hyd	$weight_{hyd} \cdot$ hydrogen bond	$weight_{hyd}$ = + 15 bondcount + betapair + bondstrand + revturn + 2 · bondsheet
sheetdir sh		$weight_{sh} \cdot$ sheetdir	$weight_{sh}$ = + 6; within 66^o, reward = +1 within 35^o, additional reward = +6;

[a] The total fitness measures the quality of the structure encoded by an individual bit string. It is the sum of the general fitness terms (C + cl + pf + co + gs + hs) and the §-strand fitness (hyd + sh) plus new fitness terms exploiting experimental information currently investigated.

[b] The term "des" refers to the abbreviated designation for the criteria involved.

[c] The constant keeps the population of prediction trials richer since low fitness individuals may also survive.

[d] Mainchain atom overlaps were counted.

[e] Structural preference rewards all residue conformations encoded in a bit string which agree with the secondary structure (known or predicted) used in the trial.

[f] Cooperativity yields a reward for any two consecutive residues in the same dihedral conformation

[g] scatter of all residues around the center of mass

[h] distribution of hydrophobic residues around the center of mass

The root mean square deviation in Angstrøms of topologically equivalent C_α atoms in the fittest structure relative to those observed is given. Left entry RMSD values include and right value exclude connecting loop residues. For each protein the fittest fold obtained after 10 simulation runs is given, Aa denotes the number of the amino acids in the protein; terminal loop residues were not included in the simulations. The secondary structure is sketched ("a" denotes helices; "T", turns; and "b", beta strands).

To model the topology of a wider range of proteins, in particular more complex topologies or proteins of larger size, we currently investigate additional fitness parameters which can be derived from further, for instance experimental, data available on the protein without knowing its three dimensional structure.

One such example are inclusion of disulfide bonds as a selection criterion. To implement them as a fitness parameter, different potentials and weights have to be tested and simple protein structures such as crambin act as a test fold. Figure 1 shows that a reasonable topology and slight RMSD improvement can be achieved applying this criterion even without having optimized its fitness weight (Figure 1). Another protein investigated is anemona toxin, a very irregular (not much secondary structure) protein fold (Figure2). The topology obtained by including the disulfide bonds as an additional distance constraint here is not too far from observed

Table 2. *Modelling different protein topologies*

helical proteins:				
1HMD (113 Aa)	a_4	4.9	3.7	hemerythrin
1ERP (37 Aa)	a_3	3.5	3.3	mating pheromone
strand rich:				
1BBI (71 Aa)	b_6	5.7	5.1	Bowman-Birk inhibitor
1DEF (30 Aa)	bTbTT	4.5	3.5	defensin
mixed structures:				
1CRN (46 Aa)	baaba	5.4	4.2	crambin
4CRO (66 Aa)	a3TbTb	6.0	5.1	lambda cro-repressor

Figure 1. *Crambin*
(A) Simulation result; the simulation result is given as a brk-file containing the main-chain atoms; RMSD to observed 5.3 Å
(B) crystal structure (1CRN.BRK)

Figure 2. *Anemona toxin*
(A) Simulation result; the simulation result is given as a brk-file containing the main-chain atoms; RMSD to observed 6.2 Å
(B) crystal structure (1ATX.BRK)

in spite of the irregularity of the protein and has an RMSD to the crystal structure of 6.2 Å.

A completely different criterion studied is for instance the formation of a protein core. This criterion can either be derived from studying and comparing the architecture of related folds or by mutagenesis data. Also with these fitness criterion a number of different folds is investigated including barnase and ubiquitin.

These and other new fitness criteria are in the moment examined in detail and in further simulations by us to allow also modelling of more complex and bigger structures.

References

1. Goldberg, D.E. *Genetic algorithms in search, optimization and machine learning*. Addison Wesley Publ., Reading, Massachusetts 1989.
2. Dandekar, T. and Argos, P. *Int.J.Biol. Macro-molecules* **1996**, *18*, 1-4.
3. Dandekar, T. and Argos, P. *Protein Engineering* **1992**, *5*, 637-645.
4. Dandekar, T. and Argos, P. *J.Mol. Biol.* **1994,** *236*, 844-861.
5. Dandekar, T. and Argos, P. *J.Mol. Biol.* **1996**, *256*, 645-660.