Using Crowding-Distance in a Multiobjective Genetic Algorithm for Protein Structure Prediction

Gregório Kappaun Rocha
gregorio@lncc.br

Fábio Lima Custódio
flc@lncc.br

Helio J.C. Barbosa
hcbm@lncc.br

Laurent Emmanuel Dardenne
Laboratório Nacional de Computação Científica - LNCC/MCTI
Rua Getúlio Vargas, 333
25651-070 Petrópolis RJ, Brazil
dardenne@lncc.br

ABSTRACT

In this paper the insertion of the crowding-distance technique in a multiobjective genetic algorithm with phenotypic crowding is carried out for the protein structure prediction (PSP) problem. The main goal is obtain a more diversified and well distributed Pareto frontiers at the end of the optimization process. Three classical force field potentials, three hydrogen bond potentials and a hydrophobic compactation term were combined in two configurations with different objectives for the fitness function. A set of 45 proteins was used to evaluate the performance of the predictions. The results were compared against the previous mono- and multiobjective approaches, and with QUARK and MEAMT, two consolidated free-modeling PSP methodologies. The strategy proposed here was able to obtain improvements in the predicted models relative to the previous mono- and multiobjective approaches, proving to be quite promising in dealing with the PSP problem.

CCS Concepts

•**Applied computing** → **Molecular structural biology;**
•**Mathematics of computing** → *Evolutionary algorithms;*

Keywords

Protein Structure Prediction, Multiobjective Optimization, Crowding-distance, Genetic Algorithm

1. INTRODUCTION

Proteins are macromolecules which exhibit a wide variety of functions, operating in practically all biological processes. They are formed by 20 amino acids, combined into sequences of several lengths. The protein structure prediction (PSP) problem consists in predicting the native three-dimensional conformation of the molecule (tertiary structure) from the information stored in its primary structure (the amino acids sequence) and it is often described as the most challenging problem in computational biology [21]. The 3D structure of a protein highlights crucial information about its function, and thus, a better understanding of its role in complex biological systems can be inferred [2].

Computational methods for PSP are justified because the experimental determination is costly and can take several months [25]. Furthermore, large scale sequence determination, mainly from genome projects, results in a growing number of proteins with unknown structure [35].

The efficiency of a PSP method depends (i) on a "free energy function" (the evaluation function), which models the forces participating in the folding process and, ideally, ranks the native structure as the global minimum [1, 25], and (ii) on a efficient search strategy, which should be able to deal with thousands of degrees of freedom in a multimodal energy landscape with large regions of unfeasible conformations [23, 10, 29, 25]. As a result, the PSP problem has a high computational cost and cannot be approached with exhaustive searches, the most common approaches to the problem employing metaheuristics [34, 4].

Genetic algorithms (GAs), a particular kind of evolutive algorithm [19], are a broadly applied metaheuristic in optmization and search problems, such as the PSP. The GA choice is justified because it (i) is a population based method capable of simultaneously exploring multiple regions of the search space, (ii) is capable of working efficiently with multimodal problems and (iii) does not require a differentiable, or even continuous, evaluation function [19, 21].

Those features make the GA well suited for application in multiobjective (MO) optimization problems [5]. A monoobjective GA can be easily modified to find an approximation of the Pareto set. A MO problem has a number of evaluation functions where each one can represent a distinct objective to be optimized. Some objectives can be conflicting, and in that case there is no optimal solution, but rather a set of "efficient" solutions, the Pareto-optimal set. This

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

GECCO'16 Companion, July 20 - 24, 2016, Denver, CO, USA

© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ISBN 978-1-4503-4323-7/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2908961.2931717

set is comprised of solutions where there is no possibility of improving one objective without diminishing another. That way, multiobjective genetic algorithms guide a population to the Pareto-optimal set, ideally avoiding premature convergence, that is, maintaining a diverse population of solutions [13, 5].

Recently, three prominent methods to achieve greater diversity in the Pareto-set were proposed: the archive truncation procedure of the Strength Pareto Evolutionary Algorithm (SPEA2) [36], the crowding-distance technique of the Non-dominated Sorting Genetic Algorithm (NSGA-II) [14], and the grid-based density of the Adaptive Grid Algorithm (AGA) [24]. These methods supplant the previous strategies [15, 16, 26] which generally use fitness sharing, and had low efficiency and high computational cost.

The PSP can be appropriately represented as a multiobjective optimization problem (MO) because frequently there is a conflict scenario in the energy landscape between different objectives in a single conformation [12, 18, 8, 22]. Furthermore, by combining (in objectives) the different potentials in varied ways the MO approach allows an interesting (and novel) exploration of the distinct terms constituting the energy function.

Another point is that the MO approach allows the use of non force field additional terms (e.g. hydrogen bond potentials, hydrophobic compactation [22]) in a more natural and straightforward way by avoiding the use of weighting coefficients to couple them with classical force fields (FF) [33, 9, 7], as is necessary for the mono-objective approach.

In the last few years, some strategies have been presented for the study of PSP problems using MO approaches [12, 21, 6, 22].

In our recent work [22], a multiobjective steady-state genetic algorithm with phenotypic crowding for free-modeling PSP (i.e., prediction without the use of template structures) was developed. This MO approach showed better prediction results when compared to the mono-objective approach (SG) [11]. The configuration with three objectives ($MO3_{HB}$) showed better exploration of the energy landscape and also obtained significant improvements in the quality of the predicted models when compared to other well established free-modelling methods, such as QUARK [34]. However, that algorithm [22] had no explicit strategy for maintaining a diverse and well distributed Pareto frontier. Furthermore, it utilized on one of its steps a single objective to choose between two solutions, which makes that approach dependent on the knowledge of the researcher in choosing one objective over the others and hinders further tests involving modifications on the energy potentials modelling the problem.

In this work a modification of our previous algorithm is proposed aiming at increasing the diversity of the solutions in the Pareto front and making the method independent from the selection of an objective. For these purposes, the Crowding-Distance technique (originally proposed in the NSGA-II) [14] is employed.

The paper is organized as follows. Section 2 describes the multiobjective genetic algorithms and the implementation of crowding-distance during parental replacement. Section 3 shows the experiments carried out, the results obtained, and pertinent discussions. Finally, conclusions are given in section 4.

2. A MULTIOBJECTIVE STEADY-STATE GA WITH PHENOTYPIC CROWDING ENHANCED BY CROWDING-DISTANCE

Rocha et al. [22] developed a multiobjective steady-state genetic algorithm with phenotypic crowding for PSP, which was implemented in the GAPF (Genetic Algorithm for Protein Folding) program [11]. In the next sections the basic operation of the MO algorithm is introduced and the modifications proposed here with the insertion of the crowding-distance technique on the parental replacement procedure are highlighted.

2.1 The Algorithm

The algorithm is briefly described in Algorithm 1, where the maximum number of function evaluations is the stop criterion ($N_{EvalsMax}$) and the current iteration (current number of function evaluations) is denoted by t.

Algorithm 1 Basic operation of the MO genetic algorithm.

```
Begin
Generate initial population (sec. 2.4)
Evaluate the population (sec. 2.5)
for t = 0 until N_EvalsMax do
  Parental selection tournament (sec. 2.2)
  Apply genetic operator on parents (sec. 2.6)
  Evaluate offspring (sec. 2.5)
  Parental replacement (Fig. 1) (sec. 2.3)
end for
End
```

The dominance and Pareto-front criteria were applied to compare two candidate solutions involving multiple objectives. Given two solutions x and y, it can be said that x dominates y if the following two conditions are satisfied: (i) x is at least equal to y in all objectives; (ii) x is better than y at least in one objective.

$D(x, y)$ is the measure of dominance between two solutions x and y applied here. It returns 1 if x is not dominated by y, and 0 if x is dominated by y [22].

The frontier value of a given solution x, $FR(x)$, denotes the number of individuals who do not dominate it, and is calculated as follows [22]:

$$FR(x) = \sum_{y=1, y \neq x}^{n} D(x, y) \quad (1)$$

where n is the population size. Thus, the higher the $FR(x)$, the better the frontier.

Two steps were modified in Rocha et al. [22] to make the original mono-objective GA applicable in a multiobjective optimization: the Parental Selection Tournament and the Parental Replacement.

2.2 Parental Selection Tournament

In this tournament, n parents are randomly selected from the current population, P_t. Let y_1 and y_2 be two possible progenitors. y_1 wins y_2 if: (i) $D(y_2, y_1) = 0$, or (ii) $FR(y_1) > FR(y_2)$. If conditions (i) and (ii) are not fulfilled (no dominance between them) then one of the parents is selected by a random draw.

2.3 Parental Replacement with Phenotypic Crowding enhanced with Crowding-Distance Technique

In PSP methodologies a structural proximity criterion in the crowding procedure is an interesting strategy [11]. The GAPF program uses the phenotypic crowding in the parental replacement and it has been maintained in the MO approach. To verify the similarity between individuals the distance matrix error (DME) of the position of the α-carbons of hydrophobic residues is used [22].

The phenotypic crowding strategy provides a good exploration of the energy landscape by allowing the concurrent identification of multiple minima and promoting the maintenance of useful population diversity. These significantly improve the chances of global minimum identification. When a MO approach is employed, care must be taken about the distribution and diversity of solutions in the Pareto-front itself.

The previous approach [22] did not employ any strategy to maintain diversity within the Pareto-front during parental replacement, furthermore, in one of its steps, the criterion used to select candidate solutions is based in a chosen objective, in that case the hydrogen bonding potential HB (Algorithm 3). Here a modification to the phenotypic crowding is introduced with the application of the crowding-distance strategy [14] that has been successfully applied in MO optimization problems resulting in diverse and well distributed solution sets. Furthermore, it is no longer required to select an objective as a criterion during the parental replacement step.

The crowding-distance is a technique proposed by Deb [14] and employed in NSGA-II (Non-dominated Sorting Genetic Algorithm II). This strategy guarantees the diversity of solutions with an estimated density of solutions in the neighborhood of each population individual. That is achieved by calculating the crowding-distance CD, that is, the sum of the normalized distances between two adjacent solutions for each individual, relative to all objectives. The least crowded solution is the one with the highest crowding-distance value (Algorithm 2).

Algorithm 2 Crowding-distance technique.

```
Begin
set T = population size
set M = number of objectives
for i = 0 until T do
  set CD[i] = 0
end for
for m = 0 until M do
  sort the population according to the m-th objective;
  F_m[1] = F_m[T] = ∞
  for i = 2 until T − 1 do
    CD[i] = CD[i] + (F_m(i+1) − F_m(i−1)) / (F_max(m) − F_min(m))
  end for
end for
End
```

In algorithm 2, $CD[i]$ is the crowding-distance value of the i-th solution; $F_m(i)$ is the value of the m-th objective for the i-th solution; $F_{max}(m)$ and $F_{min}(m)$ are the maximum and minimum values of the m-th objective, respectively. The boundary points $F_m(1)$ and $F_m(T)$, for each objective function, are initialized equal to infinity. Thus, they have the maximum crowding-distance value and are always selected.

2.3.1 The Comparison Between The Algorithms

Let Y be the new offspring, and W the most similar (lowest DME with respect to Y) solution in the parental population P_t. The algorithms 3 and 4 depict previous and current versions of the parental replacement procedure.

In the previous algorithm, from Rocha et al. [22], when there is no dominance between Y and W, and both have the same $FR(x)$ (Eq. 1) value, the HB potential is used to choose the best solution, i.e., the one that will remain in the population. Here, under such conditions the CD is the metric used to choose between the two solutions. Figure 1 shows the complete parental replacement procedure using both phenotypic crowding and crowding-distance.

Algorithm 3 Phenotypic Crowding with No Standard Decision (Based on the expert) [22]. Y replaces W if:

(i) $D(W, Y) = 0$;
(ii) $FR(Y) > FR(W)$, if there is no dominance between them;
(iii) if $FR(Y) = FR(W) =$ best front, W is kept and Y replaces a random element in the $P_{(t+1)}$, if the previous conditions are not fulfilled;
(iv) $HB(Y) < HB(W)$, if the previous conditions are not fulfilled.

Algorithm 4 Phenotypic Crowding with Crowding-Distance (with Standard Decision). Y replaces W if:

(i) $D(W, Y) = 0$;
(ii) $FR(Y) > FR(W)$, if there is no dominance between them;
(iii) $CD(Y) >= CD(W)$, if the previous conditions are not fulfilled.

2.4 Representation of Solutions and Initial Population

The backbone dihedral angles ϕ, ψ and ω are able to describe the protein 3D organization. In the program, a chromosome containing a triplet $\{\phi, \psi, \omega\}$ for each residue in the sequence represents a possible solution (3D conformation).

A coarse-grained representation [27, 17] is used to reduce the computational cost required to evaluate the energy of a particular conformation. In this representation, the side chain atoms are replaced by a super-atom located at its geometric center, while all polar backbone atoms are explicitly included.

New candidate solutions are achieved by genetic operators which change the dihedral angles ϕ and ψ. The peptide bond angle ω does not change during the search, being kept in its trans configuration (180 °).

The initial population is created from a fragment library (set of angles ϕ, ψ and ω) using the information of the secondary structure prediction provided by PSIPRED program [20, 30].

2.5 The Fitness Function

To evaluate each candidate solution, the following components of the energy function, available in the GAPF pro-

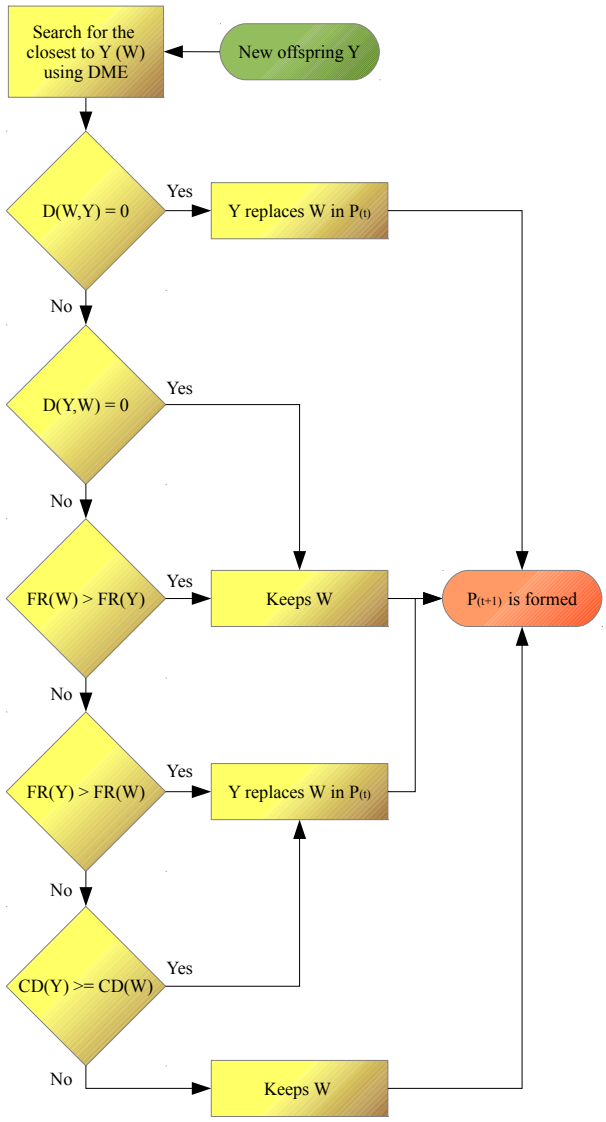


Figure 1: Parental Replacement with Phenotypic Crowding enhanced with Crowding-Distance Technique. P_t is the population in the t^{th} evaluation function.

gram, were used:

1) Classical Force Field Terms (FF): composed of three potentials of GROMOS96 force field [33, 31] as follows:

$$FF = LJ + Coul + Dihed \quad (2)$$

where, LJ, $Coul$ and $Dihed$ represent Lennard-Jones, Coulomb and Dihedral potentials, respectively.

2) Hydrogen Bonds Terms (HB): three potentials were used to model the hydrogen bonds contribution [22, 28]:

$$HB = HB_{hx} + HB_{st} + HB_{att\_st} \quad (3)$$

where, HB_{hx} is the Hydrogen Bond Potential for α-helix, HB_{st} is the Hydrogen Bond Potential for β-sheet and HB_{att_st} is the Attractive Hydrogen Bond Potential for β-sheet. These potentials require previous information about the secondary structure arrangements of the target protein and PSIPRED was used as a predictor.

3) Hydrophobic Compaction Term (Cpk): models the burial of hydrophobic residues minimizing the distances between the side-chain super-atoms of these residues.

More information about such potentials can be found in Rocha et al. [22].

2.6 Genetic Operators

An adaptive scheme based on the quality of the generated structure determines the probability of using each of the six following operators [11]: (i) two crossovers (Two-point and Multiple-point); (ii) three mutations (Incremental, Compensatory, Segments); (iii) fragments insertion [32].

3. RESULTS

3.1 Test Set and Parameters

To allow for fair comparisons, the same settings used in the previous mono-objective (referred to here as SG) and multiobjective (referred to here as MO_{HB}) GAPF prediction protocol were applied in this work.

The parameters for the genetic algorithm were: 30 independent runs per target sequence, maximum of 300,000 function evaluations, population size of 200 and parental selection tournament with four candidates ($n = 4$).

A test set of forty five proteins with known and diverse 3D structures, obtained from the Protein Data Bank [3] (PDB), were used to evaluate the method's performance (Table 1).

Table 1: The protein test set.

PDB	Length	Class	PDB	Length	Class	PDB	Length	Class
2rlg	18	α	2jzq	57	α	1f7m	46	β
1l2y	20	α	1bdd	60	α	1k36	46	β
1sol	20	α	1i2t	61	α	1msi	70	β
2xl1	24	α	1uzc	71	α	1qjo	80	β
1wqc	26	α	1xzy	90	α	1fna	91	β
1amb	28	α	1sxd	91	α	2kp0	33	$\alpha\beta^*$
1fsd	28	α	1k40	126	α	1crn	46	$\alpha\beta$
1psv	28	α	2evq	12	β	1e0g	48	$\alpha\beta$
1vii	36	α	1niz	16	β	2hbb	51	$\alpha\beta$
1erc	40	α	1e0n	27	β	2gb1	56	$\alpha\beta$
2p81	44	α	1g26	31	β	3fil	56	$\alpha\beta$
1f4i	45	α	1e0l	37	β	1dtk	57	$\alpha\beta$
1bbl	51	α	1i6c	39	β	1orc	71	$\alpha\beta$
1enh	54	α	2dmv	43	β	2rpv	75	$\alpha\beta$
1fyj	57	α	1ed7	45	β	1h5p	95	$\alpha\beta$

PDB: protein code in Protein Data Bank. Length: number of amino acid residues. Classes of proteins: α: mainly-α; β: mainly-β; $^*\alpha\beta$: includes proteins of classes α/β and $\alpha+\beta$.

The fitness function (described in section 2.5) was broken down into different objectives, and two configurations were evaluated using (Table 2): (i) two objectives ($MO2_{CD}$); (ii) three objectives ($MO3_{CD}$).

The division was carried out in order to keep the potentials of the classical force field together on the same objective, and separated from the other additional terms.

Table 2: Distribution of the fitness function terms in objectives.

Algorithm	Objective 1	Objective 2	Objective 3
$MO2$	FF	$HB + Cpk$	—
$MO3$	FF	HB	Cpk
SG	$FF + HB + Cpk$	—	—

SG (single-objective): mono-objective approach, is the GAPF original version used as reference. $MO2$ and $MO3$: multiobjective approaches developed with, respectively, two and three objectives.

3.2 Predicted 3D-Conformations: Structural Quality

To assess the quality of the predicted models a metric of structural similarity called Root Mean Square Deviation (RMSD) was applied. The lower the value of RMSD, the closer the model is to the native structure. Models with RMSD values lower than 4.0 Å are considered good, and values larger than 6,0 Å show no structural similarity. Only backbone atoms were used in the RMSD calculations computed for all structures of the final population of each run (Figure 2).

Table 3: Number of proteins with good predicted structures.

Class	SG	$MO3_{HB}$	$MO2_{CD}$	$MO3_{CD}$
Mainly-α	11	12	15	15
Mainly-β	4	5	5	5
$\alpha\beta$	0	1	0	2
Total	15	18	20	22

SG (single-objective): mono-objective approach, is the GAPF original version used as reference. $MO3_{HB}$: multiobjective approach of Rocha el al. [22] with three objectives. $MO2_{CD}$ and $MO3_{CD}$: multiobjective approaches developed in this work with, respectively, two and three objectives. A good predicted structure shows < 4.0 Å relative to the native reference structure.

Table 4: Number of targets per range of RMSD.

	SG	$MO3_{HB}$	$MO2_{CD}$	$MO3_{CD}$
< 4.0 Å	15	18	20	22
< 5.0 Å	24	28	28	31
< 6.0 Å	32	34	33	36

SG (single-objective): mono-objective approach, is the GAPF original version used as reference. $MO3_{HB}$: multiobjective approach of Rocha el al. [22] with three objectives. $MO2_{CD}$ and $MO3_{CD}$: multiobjective approaches developed in this work with, respectively, two and three objectives. A good predicted structure shows < 4.0 Å relative to the native reference structure, and a informative prediction shows RMSD value < 6.0 Å.

The proposed algorithm (MO_{CD}), using two or three objectives, was able to improve the quality of the predicted models compared to the mono-objective SG and the previous best MO version ($MO3_{HB}$) (Figure 3).

Using the MO_{CD} during parental replacement, GAPF generated good models ($RMSD < 4.0$ Å) and informative models ($RMSD < 6.0$ Å) for a larger number of targets than the SG and $MO3_{HB}$ algorithms (Tables 3 and 4).

Table 4 shows that the crowding-distance version with three objectives, $MO3_{CD}$, produced the best results increasing the number of targets predicted with good models, compared to $MO3_{HB}$, from 40.0% of the targets (18 targets) to 48.9% (22 targets).

This result agrees with the proposed in Rocha et al. [22], that the division of the fitness function in three objectives is the best configuration for this approach to the PSP problem.

The results were compared against two other free-modeling PSP methodologies: QUARK [34], which uses a replica-exchange Monte Carlo simulation; and the MEAMT[6], which is also a multiobjective genetic algorithm approach. The best model sent by QUARK server was used in the comparisons. For MEAMT, the values refer to the best model (selected with the lowest RMSD) of the round with the best average RMSD, among the ten rounds performed. The algorithms $MO3_{CD}$ produced better results than: QUARK in 57.50% (23 of 40) of the targets and than MEAMT in 80% (28 of 35) of the test proteins in common (Table 5).

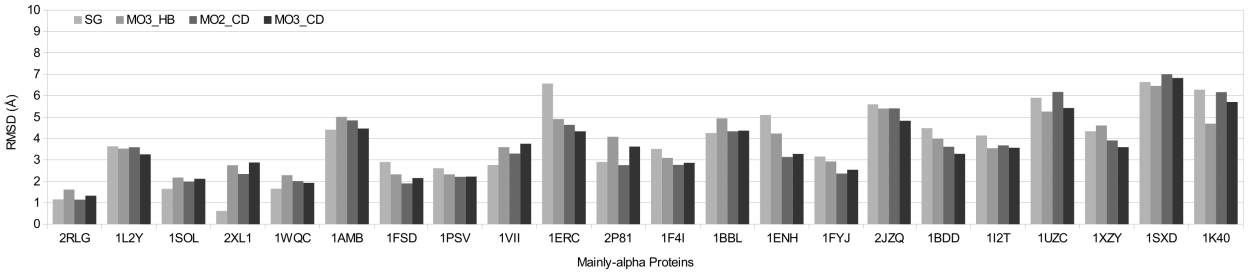
Table 5: Comparative analysis with other methods.

PDB	$MO3_{CD}$	QUARK	MEAMT	PDB	$MO3_{CD}$	QUARK	MEAMT
2RLG	**1.33**	–	1.55	1NIZ	2.54	–	**1.64**
1L2Y	3.26	4.10	**2.31**	1E0N	**3.51**	6.34	–
1SOL	2.12	1.34	**0.98**	1G26	**3.73**	4.16	5.69
2XL1	2.88	**0.34**	1.12	1E0L	**3.90**	7.20	5.84
1WQC	1.94	**1.74**	4.08	1I6C	4.82	7.37	**4.02**
1AMB	**4.47**	7.77	4.82	2DMV	**4.21**	5.77	5.45
1FSD	**2.16**	2.33	–	1ED7	**5.29**	5.60	5.52
1PSV	**2.22**	4.21	–	1F7M	6.89	**5.43**	6.13
1VII	3.76	**3.40**	5.13	1K36	**6.15**	10.18	7.09
1ERC	**4.34**	8.53	5.75	1MSI	**7.22**	7.38	8.67
2P81	**3.62**	7.79	–	1QJO	**8.13**	13.43	11.84
1F4I	2.87	**2.66**	–	1FNA	7.76	**2.97**	10.93
1BBL	4.36	**2.35**	4.69	2KP0	**3.85**	6.50	5.26
1ENH	3.29	**2.43**	6.32	1CRN	**4.97**	5.28	6.00
1FYJ	2.54	**2.00**	6.30	1E0G	3.94	**2.51**	5.44
2JZQ	4.83	**4.67**	6.44	2HBB	4.91	**1.42**	6.66
1BDD	**3.29**	5.37	6.40	2GB1	**5.72**	14.17	8.19
1I2T	3.57	**0.92**	–	3FIL	**4.79**	–	–
1UZC	**5.43**	5.62	7.91	1DTK	**5.10**	8.58	7.41
1XZY	3.60	**2.94**	–	1ORC	**6.60**	–	7.76
1SXD	6.82	**3.80**	–	2RPV	**6.99**	12.54	8.84
1K40	5.70	**2.07**	–	1H5P	**6.88**	7.59	9.67
2EVQ	2.10	–	**0.76**				

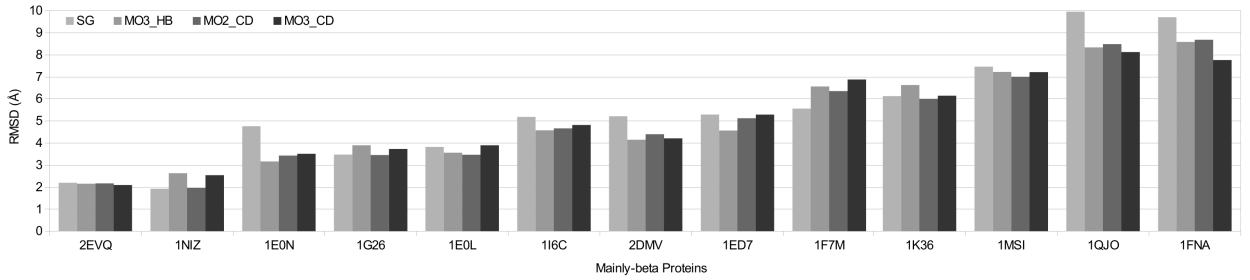
PDB: Protein Data Bank code. The best prediction is highlighted in bold. The data for MEAMT were obtained from [6] and for QUARK were obtained from the web server (http://zhanglab.ccmb.med.umich.edu/QUARK/). Values are RMSD (Å) of backbone atoms.

4. CONCLUSIONS

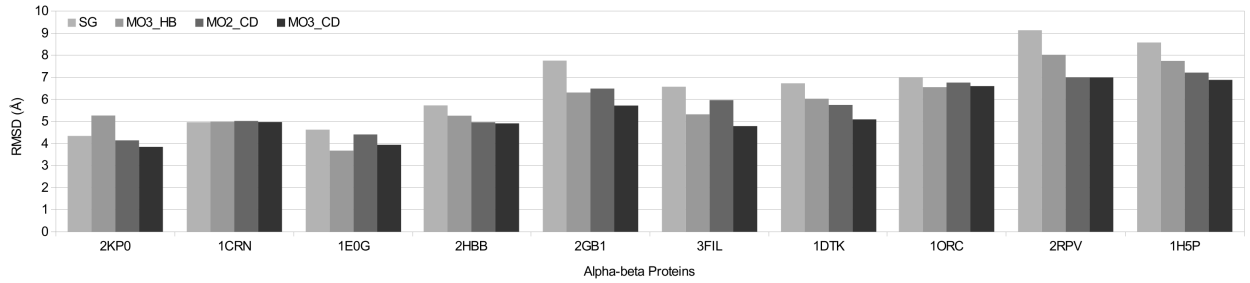
The insertion of the crowding-distance technique in the GAPF multiobjective steady-state genetic algorithm proposed in this work was able to produce improvements in the predicted models relative to the previous mono-objective and multiobjective approaches, and proved to be quite promising in dealing with the PSP problem. Furthermore, results comparable to other well established free-modeling methods, such as QUARK, were achieved. The insertion of the crowding-distance technique was also important to make the strategy used in MO algorithm of GAPF independent of the



(a)



(b)



(c)

Figure 2: Structural variations (measured with the RMSD) of the predicted models relative to the native reference structure for following protein classes: (a) mainly-α, (b) mainly-β and (c) $\alpha\beta$.

expert choice's, which is important when possible changes and introductions of new objectives are considered.

5. ACKNOWLEDGMENTS

This work was supported by: CNPq (grants 307062/2010-4 and 310778/2013-1) and FAPERJ (grants E-26/100.280/2013 and E26/102.443/2009).

6. REFERENCES

[1] C. B. Anfinsen. Principles that govern the folding of proteins. Science, 181:187, 1973.

[2] D. Baker and A. Sali. Protein structure prediction and structural genomics. Science, 294(5540):93–96, 2001.

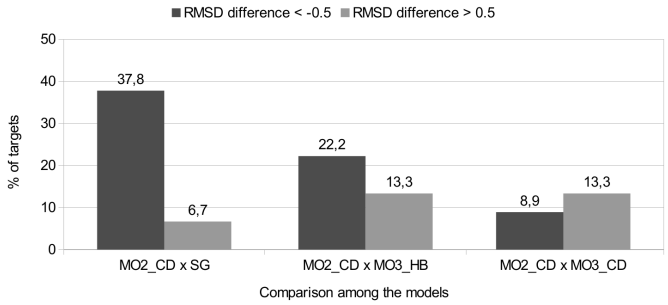
[3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. Nucleic Acids Res, 28(1):235–242, Jan 2000.

[4] R. Bonneau and D. Baker. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct, 30:173–189, 2001.

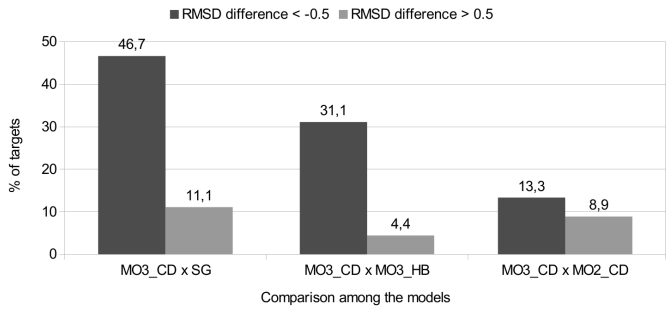
[5] C. C. Borges and H. J. C. Barbosa. A non-generational genetic algorithm for multiobjective optimization. In Evolutionary Computation, 2000. Proceedings of the 2000 Congress on, volume 1, pages 172–179. IEEE, 2000.

[6] C. R. S. Brasil, A. C. B. Delbem, and F. L. B. da Silva. Multiobjective evolutionary algorithm with many tables for purely ab initio protein structure prediction. Journal of Computational Chemistry, 34(20):1719–1734, 2013.

[7] B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The Biomolecular Simulation Program. Journal of Computational Chemistry, 30(10, Sp. Iss. SI):1545–1614, JUL 30 2009.

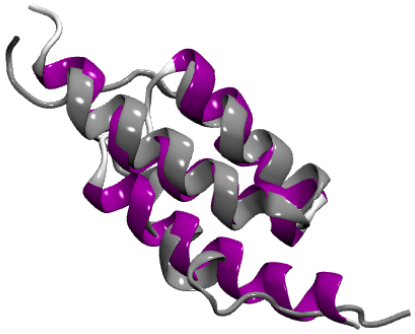


(a)

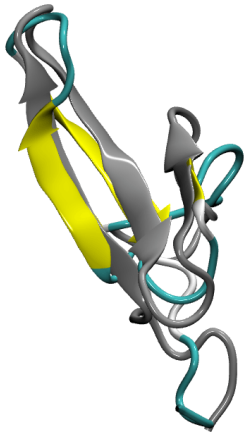


(b)

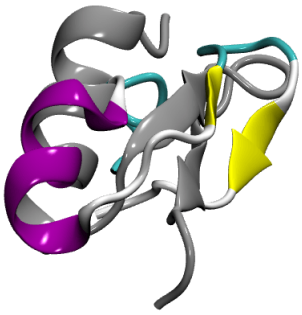
Figure 3: Comparison between $MO2_{CD}$ (a) and $MO3_{CD}$ (b) with the other algorithms considering only RMSD differences greater than 0.5 Å.



(a)



(b)



(c)

Figure 4: Predicted models using $MO3_{CD}$ approach for proteins of three distinct classes, respectively: (a) 1BDD (3.29Å), mainly-α, (b) 1E0L (3.90Å), mainly-β and (c) 2KP0 (3.85Å), $\alpha\beta$. The $MO3_{CD}$ predicted model is represented in colors and the native reference structure in silver.

[8] J. C. Calvo, J. Ortega, and M. Anguita. Comparison of parallel multi-objective approaches to protein structure prediction. The Journal of Supercomputing, 58(2):253–260, 2011.

[9] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society, 117(19):5179–5197, 1995.

[10] F. L. Custódio, H. J. C. Barbosa, and L. E. Dardenne. Investigation of the three-dimensional lattice HP protein folding model using a genetic algorithm. Genetics and Molecular Biology, 27(4):611–615, 2004.

[11] F. L. Custódio, H. J. C. Barbosa, and L. E. Dardenne. A multiple minima genetic algorithm for protein structure prediction. Applied Soft Computing, 15:88–99, 2014.

[12] V. Cutello, G. Narzisi, and G. Nicosia. A multi-objective evolutionary approach to the protein structure prediction problem. Journal of The Royal Society Interface, 3(6):139–151, 2006.

[13] K. Deb. Multi-objective genetic algorithms: Problem difficulties and construction of test problems. Evolutionary Computation, 7(3):205–230, 1999.

[14] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. Evolutionary Computation, IEEE Transactions on, 6(2):182–197, 2002.

[15] D. Goldberg. Genetic algorithms in search, optimization, and machine learning. AddisonWesley Publishing Company, Inc, Reading, MA, 1989.

[16] D. E. Goldberg and J. Richardson. Genetic algorithms with sharing for multimodal function optimization. In Genetic algorithms and their applications: Proceedings of the Second International Conference on Genetic Algorithms, pages 41–49. Hillsdale, NJ: Lawrence Erlbaum, 1987.

[17] P. V. Z. C. Goliatt. Desenvolvimento e Implementação

de um Modelo Coarse-Grained para predição de estruturas de proteínas. PhD thesis, LNCC, Rio de Janeiro/ Brasil, 2011.

[18] J. Handl, D. B. Kell, and J. Knowles. Multiobjective optimization in bioinformatics and computational biology. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 4(2):279–292, 2007.

[19] J. H. Holland. Adaptation in Natural and Artificial Systems, volume 183. University of Michigan Press, 1975.

[20] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology, 292(2):195–202, 1999.

[21] M. Judy, K. Ravichandran, and K. Murugesan. A multi-objective evolutionary algorithm for protein structure prediction with immune operators. Computer Methods in Biomechanics and Biomedical Engineering, 12(4):407–413, 2009.

[22] G. Kappaun Rocha, F. Lima Custodio, H. J. C. Barbosa, and L. E. Dardenne. A multiobjective approach for protein structure prediction using a steady-state genetic algorithm with phenotypic crowding. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on, pages 1–8. IEEE, 2015.

[23] M. Karplus. The Levinthal paradox: yesterday and today. Folding and Design, 2:S69–S75, 1997.

[24] J. Knowles and D. Corne. Properties of an adaptive archiving algorithm for storing nondominated vectors. Evolutionary Computation, IEEE Transactions on, 7(2):100–116, 2003.

[25] J. Lee, S. Wu, and Y. Zhang. Ab initio protein structure prediction. In From protein structure to function with bioinformatics, pages 3–25. Springer, 2009.

[26] S. W. Mahfoud. Niching methods for genetic algorithms. Urbana, 51(95001):62–94, 1995.

[27] J. Maupetit, P. Derreumaux, and P. Tufféry. A fast method for large-scale de novo peptide and miniprotein structure prediction. Journal of computational chemistry, 31(4):726–738, 2010.

[28] G. K. Rocha. Desenvolvimento de Metodologias para Predição de Estruturas de Proteínas Independente de Moldes. PhD thesis, LNCC, Rio de Janeiro/ Brasil, 2015.

[29] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. Protein structure prediction using rosetta. Methods in Enzymology, 383:66–93, 2004.

[30] K. B. Santos, R. Trevizani, F. L. Custodio, and L. E. Dardenne. Profrager web server: Fragment libraries generation for protein structure prediction. In Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP), page 38. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2015.

[31] L. D. Schuler, X. Daura, and W. F. van Gunsteren. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. Journal of Computational Chemistry, 22(11):1205–1218, 2001.

[32] R. Trevizani. Desenvolvimento de Metodologias *de novo* para predição de estruturas de proteínas. PhD thesis, LNCC, Rio de Janeiro/ Brasil, 2014.

[33] W. F. van Gunsteren and H. J. C. Berendsen. Groningen molecular simulation (GROMOS) library manual. Biomos, Groningen, 1987.

[34] D. Xu and Y. Zhang. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins: Structure, Function, and Bioinformatics, 80(7):1715–1735, 2012.

[35] J. Yon. Protein folding: a perspective for biology, medicine and biotechnology. Brazilian Journal of Medical and Biological Research, 34(4):419–435, 2001.

[36] E. Zitzler, M. Laumanns, L. Thiele, E. Zitzler, E. Zitzler, L. Thiele, and L. Thiele. SPEA2: Improving the strength pareto evolutionary algorithm, 2001.