Multiobjectivizing the HP Model for Protein Structure Prediction

Mario Garza-Fabre, Eduardo Rodriguez-Tello, and Gregorio Toscano-Pulido

Information Technology Laboratory, CINVESTAV-Tamaulipas
Parque Científico y Tecnológico TECNOTAM
Km. 5.5 carretera Cd. Victoria-Soto La Marina
Cd. Victoria, Tamaulipas 87130, MÉXICO
{mgarza,ertello,gtoscano}@tamps.cinvestav.mx

Abstract. The hydrophobic-polar (HP) model for protein structure prediction abstracts the fact that hydrophobic interactions are a dominant force in the protein folding process. This model represents a hard combinatorial optimization problem, which has been widely addressed using evolutionary algorithms and other metaheuristics. In this paper, the multiobjectivization of the HP model is proposed. This originally single-objective problem is restated as a multiobjective one by decomposing the conventional objective function into two independent objectives. By using different evolutionary algorithms and a large set of test cases, the new alternative formulation was compared against the conventional single-objective problem formulation. As a result, the proposed formulation increased the search performance of the implemented algorithms in most of the cases. Both two- and three-dimensional lattices are considered. To the best of authors' knowledge, this is the first study where multiobjective optimization methods are used for solving the HP model.

Keywords: Multiobjectivization, protein structure prediction, HP model.

1 Introduction

Proteins, the working molecules of the cell, are linear chains composed from up to 20 different building blocks called amino acids. The specific sequence of amino acids determines how proteins fold into unique three-dimensional structures which allow them to carry out their biological functions [1]. The *protein structure prediction* problem (PSP) can be defined as the problem of finding the functional conformation for a protein given only its amino acid sequence.

The hydrophobic-polar (HP) model [12] is an abstraction of the PSP. This model captures the fact that hydrophobicity is one of the main driving forces in protein folding. The prediction of protein structures using the HP model is a hard combinatorial optimization problem which has been demonstrated to be $\mathcal{NP}$-complete [3, 7]. A variety of metaheuristic approaches have been applied to this problem, including genetic algorithms [16, 31], memetic and hybrid algorithms [6, 17], ant colony optimization [29], immune-based algorithms [9], particle swarm optimization [5], differential evolution [25] and estimation of distribution algorithms [24]. Some of the work in this regard is reviewed in [22, 33].

Multiobjectivization concerns the reformulation of single-objective optimization problems in terms of two or more objective functions [20]. This transformation introduces fundamental changes in the search landscape, potentially allowing algorithms to perform a more efficient exploration [4, 15]. Multiobjectivization has been successfully used to deal with difficult optimization problems. Among them, there can be mentioned well-known combinatorial problems such as the traveling salesman problem [18–20], shortest path and minimum spanning tree problems [23], job-shop scheduling [19, 21] and bin packing problems [28], as well as important problems in the fields of mobile communications [26, 27] and computer vision [32]. Multiobjectivization approaches have also been proposed for the PSP [2, 8, 10, 14, 30]. However, it was not until the present study that this concept is applied to the particular HP model of this problem.

In this paper, the multiobjectivization for the HP model is proposed. The conventional HP model's energy function is decomposed into two separate objectives based on the parity of amino acid positions in the protein sequence. The suitability of this approach is investigated by comparing it with respect to the conventional single-objective formulation. Different evolutionary algorithms (EAs) and a large set of test cases were adopted for this sake. Results are provided for both the two-dimensional square lattice and the three-dimensional cubic lattice.

This paper is organized as follows. Background concepts are given in Section 2. In Section 3, the proposed multiobjectivization is described. Section 4 details the implemented EAs and the performance assessment methodology. Results are presented in Section 5. Finally, Section 6 provides the conclusions of this study.

2 Background and notation

2.1 The HP model for protein structure prediction

Amino acids can be classified either as *hydrophobic* (H) or *polar* (P) on the basis of their affinity for water. In the hydrophobic-polar (HP) model [12], proteins are abstracted as chains of H and P beads. Protein sequences, originally defined over a 20-letters alphabet, are thus of the form $S \in \{H, P\}^L$, where L is the number of amino acids. Valid conformations are modeled as *Self-Avoiding Walks* of the HP chain on a lattice. That is, each lattice node can be assigned to at most one amino acid and consecutive amino acids in S are to be also adjacent in the lattice.

By emulating the hydrophobic effect, the HP model aims to maximize the interaction among H amino acids. Two H amino acids $s_i, s_j \in S$ are said to form a *hydrophobic topological contact*, denoted by $htc(s_i, s_j)$, if they are nonconsecutive in S (*i.e.*, $|j - i| \geq 2$) but adjacent in the lattice. Following the notation of the field, an energy minimization function $E : C \rightarrow \mathbb{R}$ is defined as the negative of the total number of hydrophobic topological contacts; C is the set of all valid protein conformations. Formally, the energy of a conformation $c \in C$ is given by:

$$E(c) = \sum_{s_i, s_j \in S | i < j} e(s_i, s_j) \tag{1}$$

where

$$e(s_i, s_j) = \begin{cases} -1 \text{ if } htc(s_i, s_j) \\ 0 \;\; \text{otherwise} \end{cases}$$

The protein structure prediction problem using the HP model can be formally stated as the problem of finding the conformation $c^* \in C$ such that $E(c^*) = \min\{E(c) \mid c \in C\}$. An example conformation for an HP chain of length $L = 20$ on the two-dimensional square lattice is shown in Figure 1.

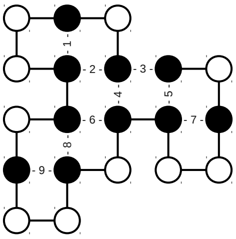

Fig. 1. Black and white beads denote H and P amino acids, respectively. Hydrophobic topological contacts have been numbered. The energy is $E(c) = -9$.

2.2 Single-objective and multiobjective optimization

Without loss of generality, a *single-objective optimization problem* can be stated as the problem of minimizing an *objective function* $f : \mathcal{F} \rightarrow \mathbb{R}$, where $\mathcal{F}$ denotes the set of all feasible solutions. The aim is to find the solution(s) $x^* \in \mathcal{F}$ yielding the optimum value for the objective function; that is, $f(x^*) = \min\{f(x) \mid x \in \mathcal{F}\}$.

Similarly, a *multiobjective optimization problem* is the problem of minimizing an *objective vector* $\mathbf{f}(x) = [f_1(x), f_2(x), \ldots, f_k(x)]^T$, where $f_i : \mathcal{F} \rightarrow \mathbb{R}$ is the i-th objective function, $i \in \{1, \ldots, k\}$. Rather than searching for a single optimal solution, the task in multiobjective optimization is to identify a set of trade-offs among the, usually conflicting, objectives. More formally, the goal is to find a set of *Pareto-optimal solutions* $\mathcal{P}^* \subset \mathcal{F}$, such that $\mathcal{P}^* = \{x^* \in \mathcal{F} \mid \nexists x \in \mathcal{F} : x \prec x^*\}$. The symbol "$\prec$" denotes the *Pareto-dominance* relation, which is given by:

$$\begin{aligned} x \prec y \Leftrightarrow \forall i \in \{1, \ldots, k\} : f_i(x) \leq f_i(y) \;\; \wedge \\ \exists j \in \{1, \ldots, k\} : f_j(x) < f_j(y) \end{aligned} \tag{2}$$

If $x \prec y$, x is said to *dominate* y. Otherwise ($x \nprec y$), y is said to be *nondominated* by x. The image of $\mathcal{P}^*$ in the objective space is called the *Pareto-optimal front*.

2.3 Multiobjectivization

Multiobjectivization refers to the process of reformulating a single-objective optimization problem as a multiobjective one [20]. Two different approaches are possible. On the one hand, additional information can be incorporated and used as *supplementary* (also called artificial or helper) objectives [4, 19]. On the other hand, in the *decomposition* approach the original objective is fragmented into several different components, each to be treated as an objective function under the new alternative formulation [15, 20]. In either approach, the idea is to alter the search landscape in order to enable a more efficient exploration, but the goal remains to solve the original problem. Therefore, the original optima are to be also Pareto-optimal with regard to the multiobjectivized version of the problem.

This work is based on the decomposition approach. More formally, a single-objective problem, with a given objective function $f : \mathcal{F} \rightarrow \mathbb{R}$, is restated in terms of $k \geq 2$ objectives $f_i : \mathcal{F} \rightarrow \mathbb{R}, i \in \{1, \ldots, k\}$ such that for all $x \in \mathcal{F}$ it holds that $f(x) = \sum_{i=1}^{k} f_i(x)$. As the only possible effect [15], plateaus may be introduced in the search landscape. That is, originally comparable solutions may become incomparable (mutually nondominated) with regard to the decomposed formulation. This can be seen as a potential strategy to escape from local optima [15, 20].

3 Multiobjectivization proposal: the parity decomposition

In the two-dimensional square and the three-dimensional cubic lattices, adjacencies (topological contacts) are only possible between amino acids whose sequence positions are of opposite parity. Based on this fact and following the multiobjectivization by decomposition approach (Section 2.3), a two-objective formulation $\mathbf{f}(c) = [f_1(c), f_2(c)]^T$ is defined over the set of feasible conformations $c \in C$:

$$f_1(c) = \sum_{s_i, s_j \in S | i<j} e_p(s_i, s_j, 0) \tag{3}$$

$$f_2(c) = \sum_{s_i, s_j \in S | i<j} e_p(s_i, s_j, 1) \tag{4}$$

where both $f_1(c)$ and $f_2(c)$ are to be minimized and

$$e_p(s_i, s_j, \rho) = \begin{cases} -1 & \text{if } htc(s_i, s_j) \wedge i \equiv \rho \pmod 2 \\ 0 & \text{otherwise} \end{cases}$$

That is, the objective function f_1 accounts only for hydrophobic topological contacts $htc(s_i, s_j)$ where i, the sequence position of amino acid s_i, is even. On the contrary, f_2 is defined for those cases where such the i-th sequence position is odd. Note that the sum of the two proposed objectives equals the conventional energy function defined in Section 2.1 (*i.e.*, $E(c) = f_1(c) + f_2(c)$ for all $c \in C$), which is in accordance with the decomposition approach for multiobjectivization.

4 Experimental setup

4.1 Algorithms

Several evolutionary algorithms (EAs) are used to investigate the suitability of the proposed multiobjectivization. The so-called (1+1) EA is described in Algorithm 1. First, an initial individual c is generated at random. At each generation, a new individual c' is created by means of mutation. If c' is at least as good as c, then c' is accepted as the starting point for the next generation. Depending on the problem formulation, this acceptance criterion is to be based either on the conventional energy evaluation or on the Pareto-dominance relation.

Algorithm 1 Basic (1+1) evolutionary algorithm.

```
choose $c \in C$ uniformly at random
repeat
    $c' \leftarrow mutate(c)$
    if $c'$ not worse than $c$ then
        $c \leftarrow c'$
    end if
until $< stop\ condition >$
```

A variant of the above described (1+1) EA is presented in Algorithm 2. An external archive stores the nondominated solutions found along the evolutionary process. The archive influences the behavior of the algorithm in such a way that the mutant c' is only accepted if it is not dominated by any archived individual. If accepted, c' is included in the archive and all individuals dominated by c', and those mapping to the same objective vector $\mathbf{f}(c')$, are removed. Note that the use of this external archive makes only sense for the multiobjectivized formulation.

Algorithm 2 Archiving (1+1) evolutionary algorithm.

```
choose $c \in C$ uniformly at random
$A \leftarrow \{c\}$
repeat
    $c' \leftarrow mutate(c)$
    if $\nexists \hat{c} \in A : \hat{c} \prec c'$ then
        $A \leftarrow \{\hat{c} \in A : c' \not\prec \hat{c} \wedge \mathbf{f}(\hat{c}) \neq \mathbf{f}(c')\} \cup \{c'\}$
        $c \leftarrow c'$
    end if
until $< stop\ condition >$
```

It was also considered a genetic algorithm (GA) whose general structure is given in Algorithm 3. First, an initial parent population P of size N is randomly generated. At each generation, the fittest individuals in P are selected for mating (*selection-for-variation*). Then, a children population P' is created by applying the genetic operators. Finally, parents and children compete for a place in the new population (*selection-for-survival*). When applied to the single-objective formulation, selection is driven by the conventional energy value of the candidate conformations. For the multiobjective formulation, the discrimination among individuals is to be based on *nondominated sorting* and *crowding distance* [11].

Algorithm 3 Genetic algorithm.

```
choose $P \subset C : |P| = N$ uniformly at random
while $< stop\ condition >$ do
    $\hat{P} \leftarrow selection\text{-}for\text{-}variation(P)$
    $P' \leftarrow variation(\hat{P})$
    $P \leftarrow selection\text{-}for\text{-}survival(P \cup P')$
end while
```

An internal coordinates representation with absolute moves was adopted in all cases. Conformations are encoded as sequences in $\{U, D, L, R, F, B\}^{L-1}$, denoting the up, down, left, right, forward and backward possible lattice locations for an amino acid with regard to the preceding one (the position of the first amino acid is fixed). Only directions $\{U, D, L, R\}$ hold for the two-dimensional lattice. The implemented genetic operators are as follows. One-point crossover (only for the GA) is applied with a given probability p_c. In mutation, each encoding position is randomly and independently perturbed with probability p_m. In all cases, only valid solutions are accepted during the search process.

4.2 Test cases and performance assessment

A total of 30 HP benchmark sequences were used (15 for the two-dimensional square lattice and 15 for the three-dimensional case). Due to space limitations, details of these instances are not provided here, but they are available online. [1]

For all the experiments, 100 independent executions were performed. Results are evaluated in terms of the best obtained energy value (β), the number of times this solution was found (f) and the arithmetic mean (μ). Additionally, the *overall average performance* (OAP) measure [13] was defined as the average ratio of the obtained mean values to the optimum (E^*). Formally, $\text{OAP} = \frac{100}{|T|}\left(\sum_{t \in T} \frac{\mu(t)}{E^*(t)}\right)$, where T denotes the set of all test cases. OAP is expressed as a percentage. Thus, a value of OAP $= 100\%$ suggests the ideal situation where the optimum solution for each benchmark was reached during all the performed executions.

Statistical significance analysis was performed for all the experiments. First, *D'Agostino-Pearson's omnibus* K^2 test was used to evaluate the normality of data distributions. For normally distributed data, either *ANOVA* or the *Welch's t* parametric tests were used depending on whether the variances across the samples were homogeneous (*homoskedasticity*) or not. This was investigated using the *Bartlett's* test. For non-normal data, the nonparametric *Kruskal-Wallis* test was adopted. A significance level of $\alpha = 0.05$ has been considered.

Most of the results are presented in tables, where values **marked ▲** highlight a statistically significant increase in performance achieved by the proposed formulation with regard to the conventional one. Conversely, values **marked ▼** indicate that a statistically significant performance decrease was obtained as a consequence of using the new alternative formulation. Additionally, the best average performance (μ) for each test instance has been **shaded** in these tables.

5 Results

5.1 Results for the (1+1) evolutionary algorithm

In this section, the (1+1) EA is used for comparing the conventional single-objective HP model formulation with respect to the proposed parity decomposition. Results are also provided for the archiving (1+1) EA, which applies only

[1] http://www.tamps.cinvestav.mx/~mgarza/HPmodel/

for the proposed formulation. A fixed mutation probability of $p_m = \frac{1}{L-1}$ and a stopping condition of 100,000 evaluations were adopted. Tables 1 and 2 present the obtained results for the two- and three-dimensional test cases, respectively.

Table 1. Results for the (1+1) EA on two-dimensional benchmarks.

			Single-objective		Parity decomposition		Parity dec. - archive	
Seq.	L	E^*	β (f)	μ	β (f)	μ	β (f)	μ
2d1	**18**	**-4**	-4 (4)	-2.70	-4 (6)	**-2.71**	-4 (5)	-2.69
2d2	**18**	**-8**	-8 (18)	-6.81	-8 (24)	**-7.04**	-8 (21)	-7.00
2d3	**18**	**-9**	-8 (11)	-7.00	-8 (48)	**-7.45 ▲**	-8 (24)	-7.12
2d4	**20**	**-9**	-9 (8)	-6.84	-9 (4)	**-6.95**	-9 (6)	-6.88
2d5	**20**	**-10**	-9 (3)	-6.92	-10 (2)	**-7.08**	-9 (1)	-6.99
2d6	**24**	**-9**	-8 (14)	-6.81	-9 (1)	-6.87	-9 (1)	**-6.89**
2d7	**25**	**-8**	-7 (26)	-5.79	-8 (6)	**-5.90**	-8 (5)	-5.80
2d8	**36**	**-14**	-13 (1)	-9.97	-13 (1)	**-10.23**	-13 (1)	-10.12
2d9	**48**	**-23**	-18 (5)	-14.23	-19 (2)	**-15.20 ▲**	-18 (5)	**-15.02 ▲**
2d10	**50**	**-21**	-18 (2)	-13.79	-18 (1)	**-14.06**	-17 (4)	-13.76
2d11	**60**	**-36**	-30 (2)	-24.39	-30 (7)	**-25.43 ▲**	-31 (1)	**-25.32 ▲**
2d12	**64**	**-42**	-29 (1)	-23.82	-30 (1)	**-25.12 ▲**	-30 (1)	**-24.63 ▲**
2d13	**85**	**-53**	-41 (1)	-33.81	-41 (1)	**-34.54**	-42 (1)	-34.18
2d14	**100**	**-48**	-41 (1)	-30.80	-39 (3)	**-32.18 ▲**	-41 (1)	**-31.72 ▲**
2d15	**100**	**-50**	-40 (1)	-31.71	-40 (3)	**-32.70 ▲**	-40 (1)	-32.57
OAP			69.22%		**71.39%**		70.47%	

Without using the archiving strategy, the parity decomposition improved the average performance of the algorithm in all the 15 two-dimensional test cases (see Table 1). For 6 out of them, such an improvement was statistically significant with regard to the conventional formulation, leading to an OAP increase of (71.39 − 69.22) = 2.17%. The use of the nondominated solutions archive seems not to be favorable for the proposed multiobjectivization. However, even in this case it was possible to score better results than the conventional single-objective formulation for most of the instances, with a statistically important difference in 4 of them. Also, an increase of 1.25% for the OAP measure has been obtained.

Table 2. Results for the (1+1) EA on three-dimensional benchmarks.

			Single-objective		Parity decomposition		Parity dec. - archive	
Seq.	L	E^*	β (f)	μ	β (f)	μ	β (f)	μ
3d1	**20**	**-11**	-11 (57)	-10.48	-11 (69)	**-10.64**	-11 (64)	-10.51
3d2	**24**	**-13**	-13 (23)	-11.30	-13 (34)	**-11.70 ▲**	-13 (27)	-11.59
3d3	**25**	**-9**	-9 (57)	-8.48	-9 (70)	**-8.65 ▲**	-9 (62)	-8.51
3d4	**36**	**-18**	-18 (10)	-15.19	-18 (13)	**-15.74 ▲**	-18 (8)	-15.30
3d5	**46**	**-32**	-30 (2)	-23.87	-30 (1)	**-25.38 ▲**	-30 (1)	-24.56
3d6	**48**	**-31**	-29 (1)	-22.79	-29 (2)	**-24.42 ▲**	-28 (3)	**-23.64 ▲**
3d7	**50**	**-32**	-25 (6)	-20.64	-27 (1)	**-22.07 ▲**	-27 (1)	-21.22
3d8	**58**	**-44**	-35 (1)	-27.34	-36 (1)	**-29.02 ▲**	-35 (1)	-27.96
3d9	**60**	**-52**	-46 (1)	-37.20	-47 (1)	**-40.03 ▲**	-47 (1)	**-38.81 ▲**
3d10	**64**	**-55**	-45 (1)	-35.59	-46 (1)	**-37.69 ▲**	-43 (2)	-36.51
3d11	**67**	**-56**	-38 (2)	-30.17	-39 (2)	**-32.65 ▲**	-38 (2)	-31.17
3d12	**88**	**-72**	-47 (1)	-36.22	-49 (1)	**-39.85 ▲**	-48 (1)	**-38.09 ▲**
3d13	**103**	**-56**	-40 (1)	-29.97	-41 (1)	**-31.31 ▲**	-38 (1)	-29.94
3d14	**124**	**-71**	-43 (4)	-34.51	-48 (1)	**-36.97 ▲**	-47 (1)	-35.04
3d15	**136**	**-80**	-51 (1)	-37.26	-52 (1)	**-42.11 ▲**	-50 (1)	**-40.43 ▲**
OAP			68.31%		**72.20%**		70.00%	

As shown in Table 2, the proposed decomposition reached the lowest average energy for all the three-dimensional instances when using the basic, non-archiving, (1+1) EA. Statistical analysis indicates a significant outperformance over the conventional single-objective formulation in all but one of the test cases. This was also reflected as an OAP increase of 3.89%. Again, the advantages of the multiobjective formulation were not as impressive when using the archiving (1+1) EA. Even so, the results were improved in most cases with regard to the conventional formulation. This performance increase was found to be statistically significant in 4 of the instances. The OAP measure was improved by 1.69%.

5.2 Results for the genetic algorithm

In this section, the obtained results regarding the implemented genetic algorithm (GA) are analyzed. The behavior of this algorithm is sensitive to several parameters. Therefore, different parameter settings have been considered in order to identify the most convenient adjustment for the compared approaches.

Three different recombination and mutation probabilities were considered: $p_c = \{0.8, 0.9, 1.0\}$ and $p_m = \{\frac{1}{L-1}, 0.01, 0.05\}$. Also, the effects of preventing duplicate individuals (clones) from the population are analyzed. This leads to a total of 18 different parameter configurations for the GA. The population size was fixed to $N = 100$ in all cases, and the algorithm was allowed to run until a maximum number of $100,000$ function evaluations was reached. Figure 2 presents the overall average performance (OAP) for both the conventional and the proposed formulations when using the different GA parameter settings.

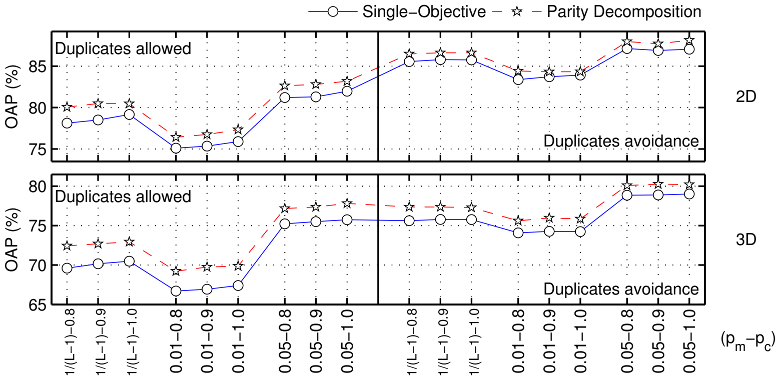

Fig. 2. Performance for all configurations of the GA.

From this figure, it is possible to note that there was a performance difference in favor of the proposed decomposition, in all the cases. On the one hand, the algorithm seemed not to be seriously affected when varying the recombination probability (p_c). On the other hand, it responded positively to the increased mutation rate, being $p_m = 0.05$ the fixed value which provided the best performance in all the cases. Finally, it can be seen that an important performance increase was achieved in all cases when duplicates avoidance was enabled.

In order to provide a more detailed analysis, the parameters adjustment which allowed each of the approaches to reach the highest OAP value has been selected. For the two-dimensional instances, a recombination probability of $p_c =$

0.8 was chosen for the conventional formulation and $p_c = 1.0$ for the proposed one. For the three-dimensional test cases, $p_c = 1.0$ and $p_c = 0.9$ were respectively selected. A mutation probability of $p_m = 0.05$ and enabled duplicates avoidance hold for all cases. The obtained results are presented in Tables 3 and 4.

Table 3. Results for the GA on two-dimensional benchmarks (best settings).

			Single-objective		Parity decomposition	
Seq.	L	E^*	β (f)	μ	β (f)	μ
2d1	**18**	**-4**	-4 (69)	-3.69	-4 (78)	**-3.78**
2d2	**18**	**-8**	-8 (92)	**-7.92**	-8 (91)	-7.91
2d3	**18**	**-9**	-9 (68)	-8.68	-9 (73)	**-8.73**
2d4	**20**	**-9**	-9 (99)	**-8.99**	-9 (93)	-8.93 ▼
2d5	**20**	**-10**	-10 (87)	-9.75	-10 (94)	**-9.89**
2d6	**24**	**-9**	-9 (62)	-8.60	-9 (69)	**-8.69**
2d7	**25**	**-8**	-8 (47)	-7.40	-8 (49)	**-7.47**
2d8	**36**	**-14**	-13 (12)	-11.45	-14 (2)	**-11.49**
2d9	**48**	**-23**	-21 (2)	-17.85	-23 (1)	**-18.30**
2d10	**50**	**-21**	-21 (4)	-18.27	-21 (1)	**-18.54**
2d11	**60**	**-36**	-34 (1)	-30.27	-34 (1)	**-30.54**
2d12	**64**	**-42**	-36 (2)	**-30.94**	-35 (3)	-30.75
2d13	**85**	**-53**	-49 (1)	-41.75	-48 (1)	**-42.57 ▲**
2d14	**100**	**-48**	-44 (1)	-36.74	-43 (1)	**-37.74 ▲**
2d15	**100**	**-50**	-43 (2)	-37.14	-43 (1)	**-38.28 ▲**
OAP			87.13%		**88.13%**	

As shown in Table 3, the parity decomposition increased the average performance of the algorithm for 12 out of the 15 two-dimensional test cases. Such an increase was statistically significant for the three largest sequences. The single-objective formulation performed best for the remaining three instances, with a statistically important difference in one of them. An increase of $(88.13 - 87.13) = 1\%$ in the OAP measure was obtained by using the proposed formulation.

Table 4. Results for the GA on three-dimensional benchmarks (best settings).

			Single-objective		Parity decomposition	
Seq.	L	E^*	β (f)	μ	β (f)	μ
3d1	**20**	**-11**	-11 (100)	**-11.00**	-11 (100)	**-11.00**
3d2	**24**	**-13**	-13 (95)	**-12.94**	-13 (97)	**-12.94**
3d3	**25**	**-9**	-9 (72)	-8.71	-9 (87)	**-8.87 ▲**
3d4	**36**	**-18**	-18 (12)	-15.91	-18 (31)	**-16.54 ▲**
3d5	**46**	**-32**	-32 (1)	-27.72	-32 (1)	**-28.12**
3d6	**48**	**-31**	-31 (1)	-26.59	-30 (3)	**-26.89**
3d7	**50**	**-32**	-30 (1)	-26.43	-29 (12)	**-26.70**
3d8	**58**	**-44**	-37 (1)	-32.39	-37 (3)	**-33.03 ▲**
3d9	**60**	**-52**	-50 (1)	-43.46	-50 (1)	**-44.56 ▲**
3d10	**64**	**-55**	-52 (1)	-46.12	-53 (1)	**-46.15**
3d11	**67**	**-56**	-41 (1)	-36.39	-43 (1)	**-37.36 ▲**
3d12	**88**	**-72**	-50 (5)	-44.02	-54 (1)	**-44.85 ▲**
3d13	**103**	**-56**	-41 (1)	-34.99	-43 (1)	**-35.78 ▲**
3d14	**124**	**-71**	-51 (1)	-41.83	-50 (1)	**-42.80 ▲**
3d15	**136**	**-80**	-52 (2)	-45.51	-56 (2)	**-46.43**
OAP			79.01%		**80.26%**	

Regarding the three-dimensional instances, it can be seen from Table 4 that the best average performance of the algorithm was obtained in all cases when using the proposed multiobjectivization. Statistical analysis has shown that for 8 out of the 15 test cases, the achieved improvement was significant with regard to the conventional formulation. The OAP measure was increased by 1.25%.

6 Conclusions and future work

The multiobjectivization of the HP model for protein structure prediction was proposed. An alternative two-objective formulation for this problem was defined by means of the decomposition of the original objective function. This approach, called the parity decomposition, is based on the fact that hydrophobic interactions in the lattice are only possible between amino acids of opposite parity.

Experiments were conducted using different evolutionary algorithms and a total of 30 HP instances. Both two- and three-dimensional lattices were explored. As the main finding, the proposed parity decomposition increased the average performance of the implemented algorithms in most of the cases. Thus, the suitability of this approach was demonstrated. The obtained results support previous evidence regarding the effectiveness of multiobjectivization to overcome search difficulties such as that of becoming trapped in local optima [15, 20].

Although still competitive, the proposed multiobjectivization was negatively affected by the use of the nondominated solutions archive within the (1+1) EA. This is contrary to what is expected in multiobjective optimization, where it is the goal to converge towards different trade-offs among the problem objectives. Nevertheless, the addressed problem of this study is actually a single-objective problem, so that maintaining an approximation set of nondominated solutions becomes not as important. In addition, the archiving strategy influences the acceptance criterion, restricting the exploration behavior of the algorithm.

Even when the performance of the GA was increased in most cases by using the proposed formulation, such an increase was not as remarkable as that achieved for the (1+1) EA. This can be explained by the fact that population-based approaches are inherently less susceptible to get stuck in local optima. On the other hand, the multiobjectivized formulation enabled diversity promotion in the objective space, thus enhancing the exploration capabilities of the algorithm.

To the best of authors' knowledge, this is the first study on the application of multiobjective optimization techniques to the HP model for protein structure prediction. It is important to remark that the aim was not to improve the state-of-the-art results, but rather to evaluate the impact of using the proposed multiobjectivization on the resolution of this problem. The findings of this study motivate further research in this direction. An important issue would be to investigate whether the proposed parity decomposition can be incorporated in order to improve the performance of established state-of-the-art algorithms. Also, the conflicting relationship between the objectives of the proposed formulation needs to be analyzed. Finally, the multiobjectivization of the HP model by means of the addition of supplementary objectives has not been addressed yet.

References

1. Anfinsen, C.: Principles that Govern the Folding of Protein Chains. Science 181(4096), 223–230 (1973)

2. Becerra, D., Sandoval, A., Restrepo-Montoya, D., Nino, L.: A Parallel Multi-Objective Ab Initio Approach for Protein Structure Prediction. In: IEEE International Conference on Bioinformatics and Biomedicine. pp. 137–141 (2010)
3. Berger, B., Leighton, T.: Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-complete. In: International Conference on Research in Computational Molecular Biology. pp. 30–39. ACM, New York, NY, USA (1998)
4. Brockhoff, D., Friedrich, T., Hebbinghaus, N., Klein, C., Neumann, F., Zitzler, E.: Do Additional Objectives Make a Problem Harder? In: Genetic and Evolutionary Computation Conference. pp. 765–772. ACM, London, England (2007)
5. Băutu, A., Luchian, H.: Protein structure prediction in lattice models with particle swarm optimization. In: Swarm Intelligence, Lecture Notes in Computer Science, vol. 6234, pp. 512–519. Springer Berlin / Heidelberg (2010)
6. Chira, C.: A Hybrid Evolutionary Approach to Protein Structure Prediction with Lattice Models. In: IEEE Congress on Evolutionary Computation. pp. 2300–2306. New Orleans, LA, USA (2011)
7. Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A., Yannakakis, M.: On the Complexity of Protein Folding. In: ACM Symposium on Theory of Computing. pp. 597–603. ACM, Dallas, TX, USA (1998)
8. Cutello, V., Narzisi, G., Nicosia, G.: A Multi-Objective Evolutionary Approach to the Protein Structure Prediction Problem. Journal of The Royal Society Interface 3(6), 139–151 (2006)
9. Cutello, V., Nicosia, G., Pavone, M., Timmis, J.: An Immune Algorithm for Protein Structure Prediction on Lattice Models. IEEE Transactions on Evolutionary Computation 11(1), 101–117 (2007)
10. Day, R., Zydallis, J., Lamont, G.: Solving the Protein structure Prediction Problem Through a Multi-Objective Genetic Algorithm. In: IEEE/DARPA International Conference on Computational Nanoscience. pp. 32–35 (2002)
11. Deb, K., Agrawal, S., Pratab, A., Meyarivan, T.: A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II. In: Parallel Problem Solving from Nature. pp. 849–858. Springer. Lecture Notes in Computer Science No. 1917, Paris, France (2000)
12. Dill, K.: Theory for the Folding and Stability of Globular Proteins. Biochemistry 24(6), 1501–9 (1985)
13. Garza-Fabre, M., Rodriguez-Tello, E., Toscano-Pulido, G.: Comparing Alternative Energy Functions for the HP Model of Protein Structure Prediction. In: IEEE Congress on Evolutionary Computation. pp. 2307–2314. New Orleans, LA, USA (2011)
14. Handl, J., Lovell, S., Knowles, J.: Investigations into the Effect of Multiobjectivization in Protein Structure Prediction. In: Parallel Problem Solving from Nature, Lecture Notes in Computer Science, vol. 5199, pp. 702–711. Springer Berlin / Heidelberg, Dortmund, Germany (2008)
15. Handl, J., Lovell, S., Knowles, J.: Multiobjectivization by Decomposition of Scalar Cost Functions. In: Parallel Problem Solving from Nature, Lecture Notes in Computer Science, vol. 5199, pp. 31–40. Springer Berlin / Heidelberg, Dortmund, Germany (2008)
16. Hoque, M., Chetty, M., Lewis, A., Sattar, A.: Twin Removal in Genetic Algorithms for Protein Structure Prediction Using Low-Resolution Model. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(1), 234–245 (2011)
17. Islam, M., Chetty, M., Murshed, M.: Novel Local Improvement Techniques in Clustered Memetic Algorithm for Protein Structure Prediction. In: IEEE Congress on Evolutionary Computation. pp. 1003–1011. New Orleans, LA, USA (2011)

18. Jähne, M., Li, X., Branke, J.: Evolutionary Algorithms and Multi-Objectivization for the Travelling Salesman Problem. In: Genetic and Evolutionary Computation Conference. pp. 595–602. ACM, Montreal, Canada (2009)
19. Jensen, M.: Helper-Objectives: Using Multi-Objective Evolutionary Algorithms for Single-Objective Optimisation. Journal of Mathematical Modelling and Algorithms 3, 323–347 (2004)
20. Knowles, J., Watson, R., Corne, D.: Reducing Local Optima in Single-Objective Problems by Multi-objectivization. In: Evolutionary Multi-Criterion Optimization. pp. 269–283. Springer-Verlag, London, UK (2001)
21. Lochtefeld, D., Ciarallo, F.: Helper-Objective Optimization Strategies for the Job-Shop Scheduling Problem. Applied Soft Computing 11(6), 4161–4174 (2011)
22. Lopes, H.: Evolutionary Algorithms for the Protein Folding Problem: A Review and Current Trends. In: Computational Intelligence in Biomedicine and Bioinformatics, Studies in Computational Intelligence, vol. 151, pp. 297–315. Springer Berlin / Heidelberg (2008)
23. Neumann, F., Wegener, I.: Can Single-Objective Optimization Profit from Multiobjective Optimization? In: Multiobjective Problem Solving from Nature, pp. 115–130. Natural Computing Series, Springer Berlin Heidelberg (2008)
24. Santana, R., Larranaga, P., Lozano, J.: Protein Folding in Simplified Models With Estimation of Distribution Algorithms. IEEE Transactions on Evolutionary Computation 12(4), 418–438 (2008)
25. Santos, J., Diéguez, M.: Differential Evolution for Protein Structure Prediction Using the HP Model. In: Foundations on Natural and Artificial Computation, Lecture Notes in Computer Science, vol. 6686, pp. 323–333. Springer Berlin / Heidelberg (2011)
26. Segredo, E., Segura, C., Leon, C.: A Multiobjectivised Memetic Algorithm for the Frequency Assignment Problem. In: IEEE Congress on Evolutionary Computation. pp. 1132–1139. New Orleans, LA, USA (2011)
27. Segura, C., Segredo, E., González, Y., León, C.: Multiobjectivisation of the Antenna Positioning Problem. In: International Symposium on Distributed Computing and Artificial Intelligence, Advances in Intelligent and Soft Computing, vol. 91, pp. 319–327. Springer Berlin / Heidelberg, Salamanca, Spain (2011)
28. Segura, C., Segredo, E., León, C.: Parallel Island-Based Multiobjectivised Memetic Algorithms for a 2D Packing Problem. In: Genetic and Evolutionary Computation Conference. pp. 1611–1618. ACM, Dublin, Ireland (2011)
29. Shmygelska, A., Hoos, H.: An Ant Colony Optimisation Algorithm for the 2D and 3D Hydrophobic Polar Protein Folding Problem. BMC Bioinformatics 6(1), 30 (2005)
30. Soares Brasil, C., Botazzo Delbem, A., Ferraz Bonetti, D.: Investigating Relevant Aspects of MOEAs for Protein Structures Prediction. In: Genetic and Evolutionary Computation Conference. pp. 705–712. ACM, Dublin, Ireland (2011)
31. Unger, R.: The Genetic Algorithm Approach to Protein Structure Prediction. In: Applications of Evolutionary Computation in Chemistry, Structure & Bonding, vol. 110, pp. 2697–2699. Springer Berlin / Heidelberg (2004)
32. Vite-Silva, I., Cruz-Cortés, N., Toscano-Pulido, G., de la Fraga, L.: Optimal Triangulation in 3D Computer Vision Using a Multi-objective Evolutionary Algorithm. In: Applications of Evolutionary Computing, Lecture Notes in Computer Science, vol. 4448, pp. 330–339. Springer Berlin / Heidelberg, Valencia, Spain (2007)
33. Zhao, X.: Advances on Protein Folding Simulations Based on the Lattice HP models with Natural Computing. Applied Soft Computing 8(2), 1029–1040 (2008)