Handling Constraints in the HP Model for Protein Structure Prediction by Multiobjective Optimization

Mario Garza-Fabre, Gregorio Toscano-Pulido and Eduardo Rodriguez-Tello
Information Technology Laboratory, CINVESTAV-Tamaulipas
Parque Científico y Tecnológico TECNOTAM
Km. 5.5 carretera Ciudad Victoria-Soto La Marina
Ciudad Victoria, Tamaulipas, 87130, MÉXICO
{mgarza, gtoscano, ertello}@tamps.cinvestav.mx

***Abstract*—The hydrophobic-polar (HP) model is an abstract representation of the protein structure prediction problem, where hydrophobic interactions are assumed to be the major determinant of the folded state of proteins. This paper inquires into the constraint-handling design issue of metaheuristics, which is crucial when dealing with such a challenging and highly constrained combinatorial optimization problem. A new constraint-handling strategy for the HP model, based on multiobjective optimization concepts, is here proposed. The multiobjective approach for handling constraints in this particular problem is explored for the first time in this study, to the authors' knowledge. Using a basic genetic algorithm and a large set of test instances for the two-dimensional HP model (based on the square lattice), the proposed multiobjective strategy was evaluated and compared with respect to commonly adopted techniques from the literature. On the one hand, through such a comparative analysis it was possible to demonstrate the suitability of the proposed multiobjective strategy. On the other hand, the results of this study provide further insight into whether infeasible protein conformations should be allowed or prevented during the metaheuristic search process, which has been a subject of concern in the specialized literature.**

I. Introduction

Proteins are composed from a set of 20 different building blocks called amino acids, carrying out most of the key processes associated with life. The specific configuration of amino acids in a protein determines how it folds into a unique and compact three-dimensional structure which defines its biological function [1]. The *protein structure prediction* problem, PSP, is the problem of finding the native (energy-minimizing) conformation for a protein given only its amino acid sequence.

The *hydrophobic-polar* (HP) *model* abstracts PSP by taking hydrophobicity as the main driving force in the protein folding process [12, 21]. Despite being an abstract formulation, the HP model of the PSP still represents a challenging problem in combinatorial optimization [2, 7]. Evolutionary algorithms and other metaheuristics are often implemented for searching the huge conformational space of this problem [23, 33].

An important decision when designing metaheuristics is how to deal with the constraints that the problem at hand involves. In the HP model of the PSP, a feasible protein conformation is defined as an embedding of the protein chain on a given lattice, such that this embedding presents *connectivity* and *self-avoidance*. While the connectivity property is implicitly satisfied by using an internal coordinates representation, see Section II-A, an explicit mechanism is required to be implemented in order to address the self-avoidance constraint.

In the literature, two main approaches have been adopted to cope with this issue. On the one hand, the search can be limited to the space of only feasible, self-avoiding protein conformations. This is usually accomplished either (i) by adapting the variation operators to iterate until new feasible conformations are generated, *i.e.*, infeasible conformations are always rejected [4, 5, 8, 9, 13, 32]; (ii) by using specialized operators which are closed on the feasible space, *i.e.*, always transforming feasible conformations into other feasible conformations [6, 22, 31]; or (iii) by implementing repairing procedures in order to map infeasible conformations into feasible ones [3, 6, 18, 29]. These strategies can be referred to as *reject strategies*, *preserving strategies* and *repairing strategies*, respectively [30]. On the other hand, infeasible protein conformations can also be taken into consideration, which is commonly done by implementing a *penalty strategy*. Using a penalty strategy, the energy value of a candidate conformation is negatively affected according to the number of collisions (overlaps) it presents [10, 19, 20, 24, 26].

It is not clear from the literature, however, whether it can be better to allow or to prevent infeasible protein conformations from being considered during the search process. Rather, very different and, to some extent, conflicting results have been reported in this respect [6, 10, 13, 20, 29]. It has been argued that the path from one compact feasible conformation to another, can be significantly shorter if the search is allowed to proceed through the space of infeasible conformations [20]. This has been, perhaps, the main motivation for applying penalty strategies when solving the HP model of the PSP. In spite of its simplicity, an inherent drawback of such an approach lies in the need for defining the severity of the penalties to be applied. Finding the right balance between objective function and penalty values has been regarded to be a difficult optimization problem itself; it is highly problem/instance-dependent and even different stages of the search process may require a different adjustment [25, 27].

In this paper, the use of multiobjective optimization concepts as a constraint-handling strategy for the HP model

of the PSP is proposed. The originally single-objective HP model is restated in multiobjective form by incorporating an additional objective function which measures the total number of collisions in a candidate conformation. In this way, infeasible protein conformations may compete against feasible ones at the selection process, being potentially exploited during the metaheuristic search. The multiobjective approach to constraint-handling has been previously applied with success to different problems in the literature. For a recent review on this topic the reader is referred to [25]. Similarly, multiobjective formulations of the HP model have been recently reported as effective mechanisms to escape from local optima [15]–[17]. Note, however, that the use of multiobjective optimization methods for handling the constraints of this particular problem is studied for the first time in the present paper, to the best of the authors' knowledge.

In order to investigate the suitability of the proposed multiobjective constraint-handling strategy, a comparative study is conducted in this paper. The proposed multibjective approach is evaluated with respect to conventional mechanisms commonly found in the specialized literature, namely, the use of reject and penalty strategies. A basic genetic algorithm and a large set of test sequences for the two-dimensional HP model, based on the square lattice, have been adopted for this sake.

The remainder of this paper is structured as follows. Background concepts and notation are covered in Section II. The studied constraint-handling methods, including the multiobjective approach proposed in this paper, are described in Section III. Section IV details the conducted experiments, the implemented genetic algorithm and the performance assessment methodology. The obtained results are discussed in Section V. Finally, Section VI provides the conclusions of this study.

II. Background and Notation

A. The hydrophobic-polar model

Amino acids can be classified as *hydrophobic* (H) or *polar* (P) on the basis of their affinity for water. While the H amino acids tend to clump together on the inside of proteins, the so-called *hydrophobic collapse*, the P ones are usually found at the outer surface interacting with the aqueous environment. The hydrophobicity of the amino acids represents, therefore, one of the major stabilizing forces responsible for the final three-dimensional conformation of proteins.

In the HP model [12, 21], proteins are abstracted as chains of H- and P-type beads. Protein sequences are thus of the form $S = (s_1, s_2, \ldots, s_L)$, where $s_i \in \{H, P\}$ denotes the i-th amino acid and L is the length of the sequence. A feasible conformation is modeled as an embedding of the protein chain on a given lattice such that two properties are satisfied: (i) *self-avoidance*, two different amino acids can not be mapped to the same lattice position; and (ii) *connectivity*, consecutive amino acids in S are to be also adjacent in the lattice.

With the aim of emulating the hydrophobic collapse, the goal in the HP model is to maximize the interaction among H amino acids in the lattice. Such interactions are to be referred to as *topological contacts*. Two H amino acids s_i and s_j are said to form a topological contact if they are nonconsecutive in S (*i.e.*, $|j - i| \geq 2$) but adjacent in the lattice. The objective is thus to find a feasible protein conformation where the number of H-H topological contacts ($HHtc$) is maximized. Adhering to the notation of the field, an energy function, to be minimized, is defined as the negative of $HHtc$; maximizing $HHtc$ is equivalent to minimizing such an energy function.

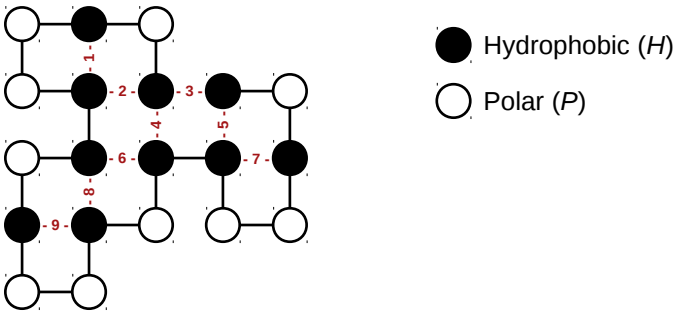

Fig. 1. Optimal conformation for sequence HPHPPHHPHPPHPHHPPHPH of length $L = 20$ on the two-dimensional square lattice. Black and white balls denote H and P residues, respectively. H-H topological contacts have been numbered. The energy of this conformation is $E(c) = -9$, since $HHtc = 9$.

Let $\mathcal{C}$ be the set of all potential protein conformations, and let $\mathcal{C}_\mathcal{F} \subseteq \mathcal{C}$ be the subset of all the feasible states. PSP under the HP model can be formally defined as the problem of finding $c^* \in \mathcal{C}_\mathcal{F}$ such that $E(c^*) = \min\{E(c) \mid c \in \mathcal{C}_\mathcal{F}\}$. $E : \mathcal{C} \to \mathbb{R}$ denotes the energy function which maps protein conformations to energy values. $E(c)$, the energy of a conformation $c \in \mathcal{C}$, is defined as follows:

$$E(c) \quad = \quad \sum_{s_i, s_j} e(s_i, s_j) \ , \tag{1}$$

where

$$e(s_i, s_j) \quad = \quad \begin{cases} -1 \ , & \text{if } s_i \text{ and } s_j \text{ are both } H \text{ and} \\ & \text{they form a topological contact} \ ; \\ 0 \ , & \text{otherwise} \ . \end{cases}$$

As an example, the optimal conformation for an HP protein sequence of length $L = 20$ on the two-dimensional square lattice is presented in Fig. 1. This example corresponds to sequence 2d4, one of the benchmark sequences considered for this study, see Section IV-C.

In this study, an *internal coordinates representation* based on *relative moves* has been adopted [26]. Protein conformations are represented as sequences in $\{F, L, R\}^{L-2}$, specifying the (Forward, Left and Right) lattice position for each amino acid with respect to the preceding one, see Fig. 2. Note that no backward moves are allowed, ensuring that the encoded conformations will always be *one-step self-avoiding*.

B. Single- and multiobjective optimization

A *single-objective optimization problem* can be stated as the problem of minimizing an objective function $f : \mathcal{F} \to \mathbb{R}$, where $\mathcal{F}$ denotes the set of all feasible solutions. The aim is to find those $x^* \in \mathcal{F}$ such that $f(x^*) = \min\{f(x) \mid x \in \mathcal{F}\}$.

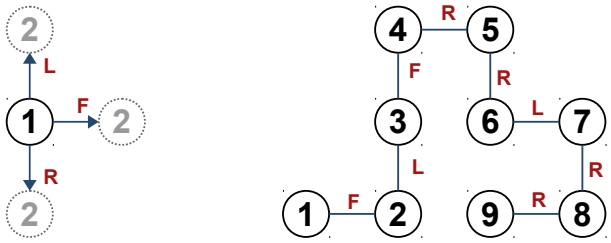

Fig. 2. Internal coordinates representation based on relative moves. Encoding scheme (left). An example conformation encoded as FLFRRLRR (rigth).

Similarly, a *multiobjective optimization problem* can be defined as the problem of minimizing an *objective vector* $\mathbf{f}(x) = [f_1(x), f_2(x), \ldots, f_k(x)]^T$, where $f_i : \mathcal{F} \to \mathbb{R}$ is the i-th objective function, $i \in \{1, 2, \ldots, k\}$. 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 defined as follows:

$$x \prec y \quad \Leftrightarrow \quad \forall i \in \{1, 2, \ldots, k\} : f_i(x) \leq f_i(y) \quad \wedge \quad (2)$$

$$\exists j \in \{1, 2, \ldots, k\} : f_j(x) < f_j(y)$$

If $x \prec y$, then x is said to *dominate* y. Otherwise ($x \not\prec y$), y is said to be *nondominated* with respect to x. The image of $\mathcal{P}^*$ in the objective space is called the *Pareto-optimal front*.

III. Studied Constraint-Handling Strategies

This section describes the constraint-handling techniques for the HP model of the PSP which have been included in this study. First, the new proposed strategy based on multiobjective optimization concepts is introduced in Section III-A. Two additional strategies, the first based on the rejection of infeasible conformations and the other based on applying penalties, were also considered as representatives of the approaches commonly used in the specialized literature. These strategies are detailed in Sections III-B and III-C.

A. Proposed multiobjective constraint-handling strategy

Based on the belief that allowing infeasible protein conformations may significantly contribute to the design of more efficient search metaheuristics, the use of multiobjective optimization is here proposed as an alternative constraint-handling strategy for the HP model of the PSP.

More formally, a two-objective formulation of the problem, $\mathbf{f}(c) = [f_1(c), f_2(c)]^T$, is defined over the set of all potential protein conformations $c \in \mathcal{C}$:

$$f_1(c) \quad = \quad E(c) \ , \qquad (3)$$

$$f_2(c) \quad = \quad Collisions(c) \ , \qquad (4)$$

where $f_1(c)$ and $f_2(c)$ are both to be minimized. $E(c)$ represents the conventional energy function of the HP model, as defined in Section II-A. $Collisions(c)$ denotes the total number of colliding amino acid pairs (s_i, s_j) in c, where both s_i and s_j were assigned to the same lattice coordinates.

Using the proposed multiobjective formulation, all feasible conformations $c \in \mathcal{C}_\mathcal{F}$ ($\mathcal{C}_\mathcal{F} \subsetneq \mathcal{C}$) will feature a value of $f_2(c) = 0$. Note, however, that an infeasible conformation c_1 ($f_2(c_1) > 0$) may become incomparable, mutually nondominated in terms of the Pareto-dominance relation, with respect to a feasible conformation c_2. This depends upon how c_1 and c_2 compare to each other with regard to the objective function f_1. Therefore, the proposed multiobjective strategy can be useful as a means of accepting infeasible protein conformations along the evolutionary process.

B. Reject strategy

A basic reject strategy was considered, where the variation operators iterate until new feasible conformations are obtained. A genetic algorithm is used in this study (see Section IV-A), whose genetic operators were adapted as follows. In the implemented one-point crossover operator, all possible crossover points are explored in random order until feasible children are generated; otherwise, either one or both of the parents are copied unchanged. Similarly, once mutation is to be applied to a given encoding position, all possible perturbations to the position are evaluated in random order until a feasible conformation is produced; otherwise, the original value is restored. Note that such a persistent application of the genetic operators involves an additional computational effort. This approach is similar to the one analyzed in [13].

C. Penalty function

A constraint-handling strategy based on the use of a penalty function was implemented according to the guidelines provided in [20]. Formally, the following objective function $f(c)$, to be minimized, is defined for every potential protein conformation $c \in \mathcal{C}$:

$$f(c) \quad = \quad E(c) + W \times Collisions(c) \ , \qquad (5)$$

where $E(c)$ denotes the conventional energy function of the HP model defined in Section II-A. $Collisions(c)$ refers to the total number of amino acid pairs (s_i, s_j) in c such that s_i and s_j overlap at the same lattice position. Finally, W is to be large enough so that $f(c_i) \leq 0, \forall c_i \in \mathcal{C}_\mathcal{F}$, while $f(c_j) > 0, \forall c_j \in \mathcal{C} \setminus \mathcal{C}_\mathcal{F}$.[1] In this study, $W = L_H^2$ was adopted, where L_H is the total number of H amino acids in the protein sequence.

IV. Experimental Setup

Using a genetic algorithm (GA), the new proposed multiobjective constraint-handling (MOCH) strategy for the HP model is evaluated and compared with respect to the reject (RJ) and penalty function (PF) approaches. The implemented GA is described in Section IV-A. Section IV-B defines the performance assessment methodology. Finally, the adopted test sequences for the HP model are detailed in Section IV-C.

[1] This ensures that the optimal (feasible) conformation is strictly better than the best penalized conformation [20].

A. The genetic algorithm

The general structure of the implemented genetic algorithm (GA) is provided in Algorithm 1. First, an initial parent population $\mathcal{P}$ of size N is randomly generated. At each generation, the fittest individuals in $\mathcal{P}$ are selected for mating ($selection\text{-}for\text{-}variation$). Then, a children population $\mathcal{P}'$ is created by applying the genetic operators to the selected parents. Finally, the parent and children populations are combined and the best individuals are selected to survive and to become the new parent population ($selection\text{-}for\text{-}survival$).

Algorithm 1 Genetic algorithm.

1: $choose\ \mathcal{P} \subset \mathcal{C} : |\mathcal{P}| = N\ uniformly\ at\ random$
2: **while** $< stop\ condition >$ **do**
3: $\quad \hat{\mathcal{P}} \leftarrow selection\text{-}for\text{-}variation(\mathcal{P})$
4: $\quad \mathcal{P}' \leftarrow variation(\hat{\mathcal{P}})$
5: $\quad \mathcal{P} \leftarrow selection\text{-}for\text{-}survival(\mathcal{P} \cup \mathcal{P}')$
6: **end while**

A decisive component of the GA is the selection process; that is, how the discrimination among the individuals is performed. Such a discrimination will depend upon the constraint-handling technique to be applied. On the one hand, it will be based on the single-objective energy value of the candidate conformations when using the RJ and PF approaches. On the other hand, if applying the proposed MOCH strategy, selection will be driven by means of *nondominated sorting* as in the *Nondominated Sorting Genetic Algorithm II*, NSGA-II [11].

Roughly, the functioning of the nondominated sorting procedure is as follows. The nondominated individuals are initially identified and isolated into the first nondominated layer, $\mathcal{L}_1$. From the remainder of the population, the new nondominated solutions are identified and assigned to the second nondominated layer, $\mathcal{L}_2$. The process repeats until each individual in the population is classified. At the $selection\text{-}for\text{-}survival$ stage, individuals are selected layer by layer, starting from $\mathcal{L}_1$, until completing the required number of individuals. Whenever the number of individuals in the current layer exceeds the available capacity, the conventional NSGA-II discriminates by using the *crowding distance* operator as a secondary criterion [11]. Rather than using the crowding distance operator, however, the implemented GA enables discrimination by using the degree of infeasibility of the individuals as the secondary criterion. This allows introducing a search bias, which is assumed essential when handling constraints by multiobjective optimization [28].

In all the cases, an internal coordinates representation based on relative moves was used, see Section II-A. Binary tournament selection was employed as mating strategy. The implemented genetic operators are as follows. One-point crossover is applied with a given probability p_c. In mutation, each encoding position is randomly and independently perturbed with probability p_m. Different recombination and mutation probabilities are explored in Section V-A. The RJ strategy requires initial feasible individuals to be generated. The backtracking algorithm proposed in [6] was adopted for this sake. In addition, in the RJ approach the genetic operators iterate until feasible children are generated, otherwise parents are copied unchanged, see Section III-B. Finally, preliminary testing has been conducted in order to explore the effects of preventing duplicate individuals (clones) from the population. As a result, the performance of the three analyzed constraint-handling methods was significantly improved in all the cases when duplicate individuals were removed from the population; this mechanism was enabled for the reported experiments.

B. Performance assessment

For all the experiments, 100 independent executions were performed and the GA was run until a maximum number of 10^6 solution evaluations was reached. The results are evaluated in terms of the best (lowest) obtained energy, the number of times this solution quality was reached, and the arithmetic mean. Moreover, given that the RJ approach involves an additional computational effort, see Section III-B, the CPU time (in seconds) required by the algorithm when using the three analyzed constraint-handling strategies is also presented.

Additionally, the *overall average performance* (OAP) measure was adopted in order to assess the overall behavior of the studied constraint-handling techniques [14]. The OAP measure is defined as follows:

$$\text{OAP} \quad = \quad \frac{100\%}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left(\frac{mean(t)}{E^*(t)} \right) , \tag{6}$$

where $\mathcal{T}$ is the set of all test cases, $mean(t)$ denotes the arithmetic mean of the energy values obtained when solving a particular test instance t, computed over multiple repetitions of the experiment, and $E^*(t)$ is the optimal energy value for instance t. Thus, OAP expresses the performance of the evaluated approaches in a 0% to 100% scale, being OAP $= 100\%$ the preferred value for this measure.

Finally, a statistical significance analysis was conducted as follows. 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.

C. Test instances

A total of 15 test sequences for the two-dimensional HP model based on the square lattice have been considered. Table I presents the full sequences, their length (L) and the optimal or best known energy value (E^*), to the authors' knowledge.

V. Results

In this section, the results for the implemented genetic algorithm (GA) are analyzed. Three different constraint-handling techniques for the HP model of the PSP are evaluated and compared: the reject (RJ) and penalty function (PF) strategies, and the new multiobjective constraint-handling (MOCH) approach which is proposed as part of this study.

TABLE I
HP MODEL INSTANCES FOR THE TWO-DIMENSIONAL SQUARE LATTICE. LENGTH OF THE PROTEIN SEQUENCE (L). BEST KNOWN ENERGY (E^*).

	Sequence	L	E^*
2d1	$H_2P_5H_2P_3HP_3HP$	18	-4
2d2	$HPHPH_3P_3H_4P_2H_2$	18	-8
2d3	$PHP_2HPH_3PH_2PH_5$	18	-9
2d4	$HPHP_2H_2PHP_2HPH_2P_2HPH$	20	-9
2d5	$H_3P_2HPHPHP_2HPHPHP_2H$	20	-10
2d6	$H_2P_2HP_2HP_2HP_2HP_2HP_2HP_2H_2$	24	-9
2d7	$P_2HP_2H_2P_4H_2P_4H_2P_4H_2$	25	-8
2d8	$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$	36	-14
2d9	$P_2HP_2H_2P_2H_2P_5H_{10}P_6H_2P_2H_2P_2HP_2H_5$	48	-23
2d10	$H_2(PH)_4H_3P(HP_3)_3(P_3H)_3PH_4(PH)_4H$	50	-21
2d11	$P_2H_3PH_8P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$	60	-36
2d12	$H_{12}PHPH(P_2H_2P_2H_2P_2H)_3PHPH_{12}$	64	-42
2d13	$H_4P_4H_{12}P_6(H_{12}P_3)_3HP_2H_2P_2H_2P_2HPH$	85	-53
2d14	$P_6HPH_2P_5H_3PH_5PH_2P_4H_2P_2H_2PH_5PH_{10}$ $PH_2PH_7P_{11}H_7P_2HPH_3P_6HPH_2$	100	-48
2d15	$P_3H_2P_2H_4P_2H_3PH_2PH_2PH_4P_8H_6P_2H_6P_9$ $HPH_2PH_{11}P_2H_3PH_2PHP_2HPH_3P_6H_3$	100	-50

This section is organized as follows. In Section V-A, different parameter settings for the GA are first evaluated with the aim of identifying the most appropriate conditions for each of the three studied constraint-handling approaches. Then, a detailed comparative analysis is presented in Section V-B.

A. Settings for the genetic algorithm

The RJ, PF and MOCH strategies were evaluated under different settings for the implemented GA. Three different recombination and mutation probabilities were considered: $p_c \in \{0.8, 0.9, 1.0\}$, $p_m \in \{\frac{1}{L-2}, \frac{2}{L-2}, \frac{3}{L-2}\}$. Thus, a total of 9 parameter configurations of the GA are investigated. The population size was fixed to $N = 100$ in all the cases. Figure 3 plots the OAP measure obtained by the studied approaches when using the different GA settings.

As it can be seen from this figure, the proposed MOCH strategy reached the highest OAP values for all the evaluated parameter configurations of the GA. From the results in Fig. 3, no strong conclusions can be made regarding the superiority of the RJ and PF approaches with respect to each other. In most of the cases, however, higher OAP values were obtained by PF. In general, the behavior of the GA seemed not to be seriously affected when varying the recombination probability, while it responded positively to the increased mutation rate. For the detailed analysis presented in Section V-B, the settings for the GA which allowed each of the compared approaches to reach the highest OAP value have been selected ($p_c = 1.0$ was selected for RJ and $p_c = 0.8$ was selected for both PF and MOCH, $p_m = \frac{3}{L-2}$ holds for all the three compared methods).

B. Comparative analysis

A detailed comparative analysis among the studied RJ, PF and MOCH strategies is presented in this section. The reported

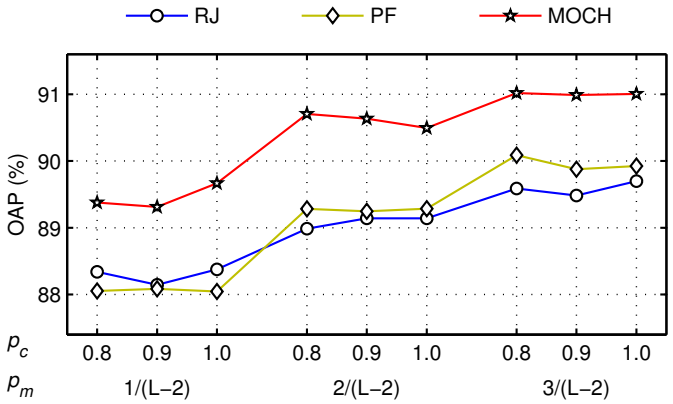

Fig. 3. Evaluating the RJ, PF and MOCH constraint-handling strategies under different parameter settings for the implemented GA.

results are based on the best performing GA settings for each of the three compared approaches, as derived in Section V-A.

Figure 4 shows the OAP measure obtained using the RJ, PF and MOCH strategies as the search process of the GA progressed (at different numbers of solution evaluations). This figure is quite revealing in several respects. First, it is evident from the plot that the best results at the end of the search process were obtained by using the proposed MOCH strategy. Nevertheless, this approach exhibited the poorest overall performance at the first stages of the search. A similar behavior can be observed for the PF strategy, but PF required a significantly higher number of solution evaluations to improve the results with respect to the RJ approach. The fact that the RJ method achieved higher OAP values at the first stages of the search, suggests that both the PF and MOCH approaches invest a considerable amount of effort in the exploration of infeasible protein conformations. It is important to note, however, that PF and particularly the proposed MOCH strategy presented a greater tendency to improve (the slope of the corresponding curves is more pronounced).

To further compare the three studied constraint-handling strategies, Table II details the results obtained by the GA after 10^6 solution evaluations. For each of the adopted test cases,

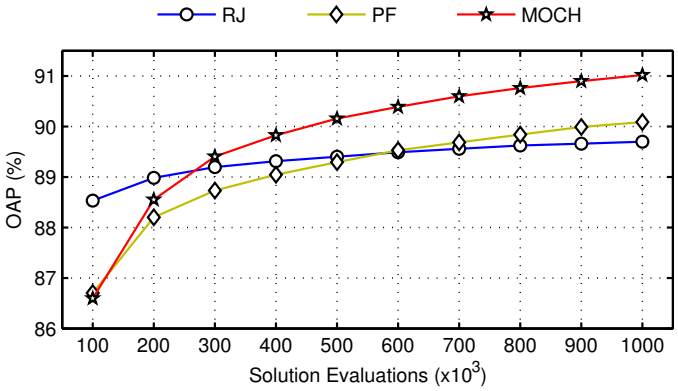

Fig. 4. Overall average performance (OAP) obtained by the studied constraint-handling techniques as the search process of the GA progresses.

TABLE II
THIS TABLE DETAILS THE RESULTS OBTAINED BY THE GA WHEN USING THE ANALYZED RJ, PF AND MOCH STRATEGIES.

			RJ			PF			MOCH		
Seq.	***L***	***E****	**Best (freq)**	**Mean**	**Time (s)**	**Best (freq)**	**Mean**	**Time (s)**	**Best (freq)**	**Mean**	**Time (s)**
2d1	**18**	**-4**	-4 (99)	-3.99	30.32	-4 (100)	**-4.00**	**13.88**	-4 (100)	**-4.00**	26.02
2d2	**18**	**-8**	-8 (100)	**-8.00**	32.05	-8 (100)	**-8.00**	**14.51**	-8 (100)	**-8.00**	27.86
2d3	**18**	**-9**	-9 (100)	**-9.00**	33.36	-9 (100)	**-9.00**	**15.03**	-9 (99)	-8.99	26.96
2d4	**20**	**-9**	-9 (100)	**-9.00**	36.84	-9 (100)	**-9.00**	**15.63**	-9 (100)	**-9.00**	28.75
2d5	**20**	**-10**	-10 (99)	-9.98	36.59	-10 (100)	**-10.00**	**16.29**	-10 (100)	**-10.00**	28.27
2d6	**24**	**-9**	-9 (93)	-8.93	46.81	-9 (89)	-8.89	**18.65**	-9 (94)	**-8.94**	31.27
2d7	**25**	**-8**	-8 (52)	-7.51	46.40	-8 (77)	-7.77	**18.06**	-8 (95)	**-7.95**	31.17
2d8	**36**	**-14**	-13 (24)	-11.87	81.01	-14 (4)	-11.93	**29.87**	-14 (4)	**-11.95**	42.77
2d9	**48**	**-23**	-22 (1)	-18.82	128.49	-22 (3)	-19.10	**46.62**	-22 (7)	**-19.67**	59.77
2d10	**50**	**-21**	-21 (20)	-19.40	139.49	-21 (17)	-19.24	**52.21**	-21 (33)	**-20.14**	63.78
2d11	**60**	**-36**	-33 (6)	-30.28	199.74	-34 (1)	-30.41	**81.23**	-35 (1)	**-31.37**	93.36
2d12	**64**	**-42**	-38 (1)	**-33.03**	223.40	-38 (3)	-32.86	**87.86**	-37 (1)	-31.49	97.15
2d13	**85**	**-53**	-48 (1)	-43.06	357.99	-48 (3)	-43.02	**146.69**	-49 (2)	**-43.20**	156.04
2d14	**100**	**-48**	-41 (1)	-35.91	436.12	-41 (1)	-36.21	**155.69**	-43 (1)	**-37.12**	168.57
2d15	**100**	**-50**	-42 (2)	-37.51	442.27	-43 (2)	-38.10	**160.54**	-44 (2)	**-39.47**	173.40
	OAP			89.70%			90.09%			**91.02%**	

this table includes the best obtained energy value, the number of GA runs where this solution was found (freq), the arithmetic mean, and the average CPU time (in seconds) consumed by the algorithm. The OAP measure is also provided at the bottom of the table for each of the analyzed approaches. In addition, the lowest average energy for each of the instances, as well as the highest OAP values, appear **shaded** and the best scored CPU times were highlighted in **bold** in this table.

The results in Table II indicate that the proposed MOCH strategy achieved the lowest average energy in 13 out of the 15 considered test cases. MOCH improved the OAP measure by $(91.02 - 89.70) = 1.32\%$ and by 0.93% over the OAP values obtained using RJ and PF, respectively. In most of the cases, better results were obtained by PF when compared to RJ.

As shown in Table II, the RJ strategy involves a significant amount of additional computational effort, which becomes more evident as the length of the protein sequence (L) increases. As expected, the best results in terms of CPU time were obtained in all the cases by using the PF approach. PF required only about 40% (on average) of the CPU time consumed by the RJ strategy on solving the adopted test instances. Due to the use of the nondominated sorting procedure, see Section IV-A, the proposed MOCH method induces an increase of roughly 13 seconds in all the cases when compared to the CPU time scored by the PF approach.[2] This implies that MOCH nearly doubled the time required by PF for the smallest test cases, while such additional 13 seconds represent an increase of less than 10% of the PF time for the largest instances. The proposed MOCH strategy takes only about 60% (on average) of the CPU time required by the RJ technique.

Finally, Table III outlines how the three studied constrain-handling strategies compare statistically with respect to each other in all the test cases. Each row in this table compares two approaches, say A and B, which is denoted by "A/B". If a statistically significant difference exists between the performance of A and B, the corresponding cells are marked either $+$ or $-$ depending on whether such a difference favors A or not. Unmarked cells indicate that there was not a significant difference between the A and B approaches. The rightmost column presents the overall results of this analysis.

TABLE III
STATISTICAL SIGNIFICANCE ANALYSIS FOR COMPARING THE PERFORMANCE OF THE GA WHEN USING THE DIFFERENT STUDIED CONSTRAINT-HANDLING APPROACHES.

	Protein sequence														
	2d1	**2d2**	**2d3**	**2d4**	**2d5**	**2d6**	**2d7**	**2d8**	**2d9**	**2d10**	**2d11**	**2d12**	**2d13**	**2d14**	**2d15**	**Overall**
PF/RJ							+									**1+ 0−**
MOCH/RJ							+		+	+	+	−		+	+	**6+ 1−**
MOCH/PF							+		+	+	+	−		+	+	**6+ 1−**

As Table III indicates, only in one of the instances (2d7) there was a statistically significant difference between RJ and

[2]The time increase presented by MOCH with respect to PF (about 13 seconds) remains almost invariant for the different instances. This is due to the fact that the computational effort required by the nondominated sorting procedure relates only to the number of objective functions and the population size, and not to the length of the protein sequence.

PF, a difference which favors PF. It can be observed from the table that the proposed MOCH strategy significantly improved the performance of the GA in 6 of the adopted test cases (2d7, 2d9, 2d10, 2d11, 2d14 and 2d15) when compared with respect to both RJ and PF. Notice, however, that MOCH presented a significantly inferior performance with respect to the RJ and PF approaches at solving one of the instances (2d12).

VI. Conclusions and future work

The HP model for protein structure prediction represents a highly constrained optimization problem. Therefore, explicit mechanisms are required to be implemented within metaheuristics in order to ensure the feasibility of the generated protein conformations. The efforts of the research community on this issue can be divided into two broad classes: approaches where only feasible conformations are considered, and those where the infeasible conformations are also allowed to participate during the search process. Nevertheless, there is no clear consensus in the literature on which of such directions could lead to more efficient metaheuristic algorithms; even contradictory results have been reported in this regard. The aim of the present study was to provide further insight into this matter, as well as to introduce a new constraint-handling strategy for the HP model which is based on multiobjective optimization.

A comparative study was performed where three different constraint-handling strategies for the HP model were considered: (i) a reject strategy, RJ, where the search was restricted to the space of only feasible protein conformations; (ii) a penalty function, PF, where infeasible solutions were penalized according to the number of collisions they present; and (iii) the new multiobjective constraint-handling strategy proposed in this paper, MOCH. Rather than penalizing, in MOCH an additional objective function accounts for the degree of infeasibility of the candidate conformations. Using such a strategy, infeasible conformations may become incomparable (nondominated) with respect to feasible ones, being thus potentially considered and exploited during the search process.

RJ, PF and MOCH were evaluated and compared in terms of how the use of these approaches impacted on the performance of a basic genetic algorithm (GA). On the one hand, the proposed MOCH strategy significantly increased the search performance of the implemented GA in most of the adopted test cases when compared with respect to the RJ and PF methods. In this way, the suitability of this proposal has been demonstrated. On the other hand, both MOCH and PF performed better in most of the conducted experiments when compared to the RJ strategy. It was also found that the RJ method involves a considerable amount of additional computational effort. Hence, these findings give further support to the belief that considering infeasible protein conformations may have beneficial effects on the search performance of metaheuristics for solving the HP model.

To the best of the authors' knowledge, this is the first study on the use of multiobjective optimization methods to face the constraint-handling requirement which arises when dealing with the HP model. Although quite promising results were obtained by using the proposed MOCH strategy, it has been argued that the multiobjective approach to constraint-handling could be rather ineffective if a bias towards the feasible region is not properly introduced [28]. As detailed in Section IV-A, a search bias was incorporated by using the feasibility of the individuals as a supplementary discrimination criterion. Future work will concentrate on evaluating how the incorporation of such a (naive) mechanism has contributed to the effectiveness of the proposed MOCH strategy. Also, alternative and more sophisticated mechanisms for biasing the search process are to be investigated. Finally, it is important to extend this research to other lattice configurations, such as the three-dimensional cubic lattice, and to consider a wider set of constraint-handling strategies for the HP model from the literature (other than RJ and PF) in order generalize the conclusions of this study.

Acknowledgment

The first author acknowledges support from CONACyT through a scholarship to pursue graduate studies at the Information Technology Laboratory, CINVESTAV-Tamaulipas. Also, the authors would like to acknowledge support from CONACyT through projects 105060 and 99276.

References

[1] C. Anfinsen, "Principles that Govern the Folding of Protein Chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[2] B. Berger and T. Leighton, "Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-complete," in *International Conference on Research in Computational Molecular Biology*. New York, NY, USA: ACM, 1998, pp. 30–39.

[3] B. Chen, L. Li, and J. Hu, "A Novel EDAs Based Method for HP Model Protein Folding," in *IEEE Congress on Evolutionary Computation*, Trondheim, Norway, 2009, pp. 309–315.

[4] C. Chira, "A Hybrid Evolutionary Approach to Protein Structure Prediction with Lattice Models," in *IEEE Congress on Evolutionary Computation*, New Orleans, LA, USA, 2011, pp. 2300–2306.

[5] C. Chira, D. Horvath, and D. Dumitrescu, "An Evolutionary Model Based on Hill-Climbing Search Operators for Protein Structure Prediction," in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, vol. 6023, pp. 38–49.

[6] C. Cotta, "Protein Structure Prediction Using Evolutionary Algorithms Hybridized with Backtracking," in *Artificial Neural Nets Problem Solving Methods*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2003, vol. 2687, pp. 321–328.

[7] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis, "On the Complexity of Protein Folding," in *ACM Symposium on Theory of Computing*. Dallas, TX, USA: ACM, 1998, pp. 597–603.

[8] V. Cutello, G. Morelli, G. Nicosia, M. Pavone, and G. Scollo, "On Discrete Models and Immunological Algorithms for Protein Structure Prediction," *Natural Computing*, vol. 10, no. 1, pp. 91–102, 2011.

[9] V. Cutello, G. Nicosia, M. Pavone, and J. Timmis, "An Immune Algorithm for Protein Structure Prediction on Lattice Models," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 1, pp. 101–117, 2007.

[10] C. de Almeida, R. Gonçalves, and M. Delgado, "A Hybrid Immune-Based System for the Protein Folding Problem," in *Evolutionary Computation in Combinatorial Optimization*, ser. Lecture Notes in Computer Science. Valencia, Spain: Springer Berlin / Heidelberg, 2007, vol. 4446, pp. 13–24.

[11] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA–II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, April 2002.

[12] K. Dill, "Theory for the Folding and Stability of Globular Proteins," *Biochemistry*, vol. 24, no. 6, pp. 1501–9, 1985.

[13] S. Duarte-Flores and J. Smith, "Study of Fitness Landscapes for the HP model of Protein Structure Srediction," in *IEEE Congress on Evolutionary Computation*, Canberra, Australia, 2003, vol. 4, pp. 2338–2345.

[14] M. Garza-Fabre, E. Rodriguez-Tello, and G. Toscano-Pulido, "Comparing Alternative Energy Functions for the HP Model of Protein Structure Prediction," in *IEEE Congress on Evolutionary Computation*, New Orleans, LA, USA, 2011, pp. 2307–2314.

[15] ——, "An Improved Multiobjectivization Strategy for HP Model-Based Protein Structure Prediction," in *Parallel Problem Solving from Nature*, ser. Lecture Notes in Computer Science. Taormina, Italy: Springer Berlin / Heidelberg, 2012, vol. 7492, pp. 82–92.

[16] ——, "Multiobjectivizing the HP Model for Protein Structure Prediction," in *Evolutionary Computation in Combinatorial Optimization*, ser. Lecture Notes in Computer Science. Málaga, Spain: Springer Berlin / Heidelberg, 2012, vol. 7245, pp. 182–193.

[17] M. Garza-Fabre, G. Toscano-Pulido, and E. Rodriguez-Tello, "Locality-based Multiobjectivization for the HP Model of Protein Structure Prediction," in *Genetic and Evolutionary Computation Conference*. Philadelphia, PA, USA: ACM, 2012, pp. 473–480.

[18] C. M. Johnson and A. Katikireddy, "A Genetic Algorithm with Backtracking for Protein Structure Prediction," in *Genetic and Evolutionary Computation Conference*. Seattle, WA, USA: ACM, 2006, pp. 299–300.

[19] M. Khimasia and P. Coveney, "Protein Structure Prediction as a Hard Optimization Problem: The Genetic Algorithm Approach," *Molecular Simulation*, vol. 19, no. 4, pp. 205–226, 1997.

[20] N. Krasnogor, W. Hart, J. Smith, and D. Pelta, "Protein Structure Prediction With Evolutionary Algorithms," in *Genetic and Evolutionary Computation Conference*. Orlando, FL, USA: Morgan Kaufman, 1999.

[21] K. Lau and K. Dill, "A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins," *Macromolecules*, vol. 22, no. 10, pp. 3986–3997, 1989.

[22] N. Lesh, M. Mitzenmacher, and S. Whitesides, "A Complete and Effective Move Set for Simplified Protein Folding," in *International Conference on Research in Computational Molecular Biology*. Berlin, Germany: ACM, 2003, pp. 188–195.

[23] H. Lopes, "Evolutionary Algorithms for the Protein Folding Problem: A Review and Current Trends," in *Computational Intelligence in Biomedicine and Bioinformatics*, ser. Studies in Computational Intelligence. Springer Berlin / Heidelberg, 2008, vol. 151, pp. 297–315.

[24] H. Lopes and M. Scapin, "An Enhanced Genetic Algorithm for Protein Structure Prediction Using the 2D Hydrophobic-Polar Model," in *Artificial Evolution*, ser. Lecture Notes in Computer Science. Lille, France: Springer Berlin / Heidelberg, 2006, vol. 3871, pp. 238–246.

[25] E. Mezura-Montes and C. A. Coello Coello, "Constraint-handling in Nature-inspired Numerical Optimization: Past, Present and Future," *Swarm and Evolutionary Computation*, vol. 1, no. 4, pp. 173–194, 2011.

[26] A. Patton, W. Punch III, and E. Goodman, "A Standard GA Approach to Native Protein Conformation Prediction," in *International Conference on Genetic Algorithms*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 574–581.

[27] T. Runarsson and X. Yao, "Stochastic Ranking for Constrained Evolutionary Optimization," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 3, pp. 284–294, 2000.

[28] ——, "Search Biases in Constrained Evolutionary Optimization," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 35, no. 2, pp. 233–243, 2005.

[29] J. Santos and M. Diéguez, "Differential Evolution for Protein Structure Prediction Using the HP Model," in *Foundations on Natural and Artificial Computation*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 6686, pp. 323–333.

[30] E. Talbi, *Metaheuristics: From Design to Implementation*. Wiley Publishing, 2009.

[31] C. Thachuk, A. Shmygelska, and H. Hoos, "A Replica Exchange Monte Carlo Algorithm for Protein Folding in the HP Model," *BMC Bioinformatics*, vol. 8, no. 1, p. 342, 2007.

[32] R. Unger and J. Moult, "Genetic Algorithms for Protein Folding Simulations," *Journal of Molecular Biology*, vol. 231, no. 1, pp. 75–81, 1993.

[33] X. Zhao, "Advances on Protein Folding Simulations Based on the Lattice HP models with Natural Computing," *Applied Soft Computing*, vol. 8, no. 2, pp. 1029–1040, 2008.