Comparing Alternative Energy Functions for the HP Model of 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*—Protein structure prediction is the problem of finding the functional conformation of a protein given only its amino acid sequence. The HP lattice model is an abstract formulation of this problem, which captures the fact that hydrophobicity is one of the major driving forces in the protein folding process. This model represents a hard combinatorial optimization problem and has been widely addressed through metaheuristics such as evolutionary algorithms. However, the conventional energy (evaluation) function of the HP model does not provide an adequate discrimination among potential solutions, which is an essential requirement for metaheuristics in order to perform an effective search. Therefore, alternative energy functions have been proposed in the literature to cope with this issue. In this study, we inquire into the effectiveness of several of such alternative approaches. We analyzed the degree of discrimination provided by each of the studied functions as well as their impact on the behavior of a basic memetic algorithm. The obtained results support the relevance of following this research direction. To our knowledge, this is the first work reported in this regard.**

I. Introduction

Proteins are at the heart of cellular function, making possible most of the key processes associated with life. It is widely accepted that the specific functionality of a protein is dictated by its three-dimensional conformation. To fully understand the biological roles of a protein it is imperative, therefore, to first determine its structure. However, given the limitations of the experimental methods, computational approaches to determine the structure of proteins have become increasingly necessary for the understanding of such important biological macromolecules.

The *Protein Structure Prediction* (PSP) problem is concerned with finding the native conformation of proteins. Such a structure is assumed to be encoded in the amino acid sequence and to correspond to the thermodynamically most stable state [1]. Nevertheless, exploring the huge conformational space to find the native structure of a protein represents a very computationally-intensive task, which makes studies at atomic resolution prohibitive even for relatively small proteins. On the other hand, simplified protein models have emerged as valuable tools for studying the most general and essential principles governing the protein folding process [2]–[5].

One of such simplified formulations of PSP is the HP model [6, 7]. This model captures the fact that hydrophobicity is one of the main driving forces determining the functional conformation of proteins. In spite of its apparent conceptual simplicity, the task of finding the optimal structure of a protein in the HP model represents a hard combinatorial optimization problem which has been proved to be $\mathcal{NP}$-complete [8, 9]. Such a complexity has motivated the use of evolutionary algorithms (EAs) and a variety of other metaheuristics to address this problem [10].

EAs rely on an effective evaluation scheme in order to guide the search towards promising regions of the solutions space. However, the conventional energy (evaluation) function of the HP model provides a very poor discrimination among potential conformations. As a consequence, no preferences can be set among individuals for selection purposes, leading the search process to be driven practically at random. This problem is expected to impact largely the performance of local search algorithms. A weak discrimination will produce large plateaus in the energy landscape, on which local search strategies could fail to detect a promising search direction [11].

In the literature, alternative HP energy functions have been proposed to improve the performance of search algorithms [11]–[14]. Nevertheless, there are no reported results on the advantages of using most of such approaches. In this paper, a comparative study is presented where five different formulations of the HP energy function are considered. The discrimination potential of each of these functions is first analyzed. This factor is assumed to be decisive for the behavior of search algorithms. Finally, the effectiveness of the studied approaches to guide the search process is investigated by using a basic memetic algorithm. To the best of our knowledge, this is the first work that has been reported in this direction.

The remainder of this paper is organized as follows. Background concepts are provided in Section II. In Section III, the alternative HP energy functions adopted for this study are described. Our experimental results are discussed in Section IV. Finally, Section V provides our conclusions as well as some possible directions for future research.

II. Background

A. Proteins

Proteins are the working molecules of the cell. Amino acids, the building blocks of proteins, are all of them consistent with the general structure presented in Figure 1.

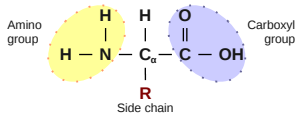

Fig. 1. General structure of amino acids.

Each amino acid has a central carbon atom (C_α) which is covalently bonded to a carboxyl group (COOH), to an amino group (NH_2), to a hydrogen atom (H) and to a radical (R) group or side chain. There are 20 amino acids commonly found in proteins, each of which has a distinctive R group that is responsible for its particular chemical properties.

In proteins, amino acids are held together by *peptide bonds*. Hence, protein chains are also referred to as *polypeptides*. The peptide bond is formed when the carboxyl group of an amino acid reacts with the amino group of another, releasing a water molecule. The elements of a polypeptide chain are, therefore, amino acid *residues*. This process is illustrated in Figure 2.

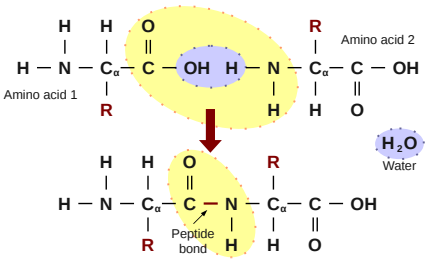

Fig. 2. Peptide bond formation.

Protein structure is commonly described in terms of four hierarchical levels of organization. The amino acid sequence constitutes the *primary structure* of a protein. While the *secondary structure* describes the arrangement of amino acids within certain areas of a polypeptide chain, *the tertiary structure* defines its overall folding in three-dimensional space. It is the three-dimensional conformation which is essential to the function of the molecule. Finally, some proteins are composed of more than one polypeptide chain. The spatial arrangement of these subunits is known as the *quaternary structure*.

B. Protein Structure Prediction

Anfinsen's theory of protein folding states that the native structure of a protein is given by the chemistry of its amino acid sequence. Such a structure minimizes the overall free energy; *i.e.*, the thermodynamically most stable conformation. This is the so-called *thermodynamic hypothesis* [1].

The *Protein Structure Prediction* (PSP) problem aims to determine the native conformation of proteins given only their linear chain of amino acids. In PSP, one considers a fixed energy model $E : C \rightarrow \mathbb{R}$, where C is the set of all possible conformations for a given protein. The native structure is assumed to be the one with the lowest energy value according to the adopted model. That is, the conformation $c^* \in C$ such that $E(c^*) = \min\{E(c) \mid c \in C\}$.

Thus, we could simply enumerate and evaluate all possible conformations to identify the one with minimal energy. Nevertheless, proteins are very flexible and, consequently, the space of potential conformations is huge. This makes studies at atomic resolution to some extent prohibitive even for relatively small proteins. In this context, simplified models have emerged as important tools for theoretical studies of protein structure, dynamics and thermodynamics. These models provide a valuable insight to advance the understanding of the most general and essential principles governing the protein folding process [2]–[5]. This study focuses on one of such simplified protein models: the so-called HP model [6, 7].

C. The HP model

Amino acids can be classified on the basis of their affinity for water. *Hydrophilic* or *polar* amino acids (P) are usually found at the outer surface of proteins. By interacting with the aqueous environment, these residues contribute to the solubility of the molecule. In contrast, *hydrophobic* or *nonpolar* residues (H) tend to pack on the inside of proteins, where they interact with one another to form a water-insoluble core. These properties of the amino acids represent, therefore, one of the major driving forces responsible for the final three-dimensional structure of proteins.

In the Hydrophobic-Polar (HP) model, proposed by Dill in 1985 [6, 7], proteins are represented as sequences of the form $S \in \{H, P\}^L$, where L denotes the number of amino acids. The subsets of H and P residues in S are here referred to as S_H and S_P, respectively. Valid conformations are modeled as *Self-Avoiding Walks* of the HP sequence S on a lattice. That is, 1) lattice nodes are labeled by the amino acids, 2) a lattice node can be assigned to at most one residue and 3) adjacent residues in S are also adjacent in the lattice. This study focuses on the two-dimensional square lattice.

By emulating hydrophobic interactions, the goal in the HP model is to find a valid conformation where the number of *H-H topological contacts* (HHtc) is maximized. Two residues $s_i, s_j \in S$ are said to form a topological contact, denoted by $tc(s_i, s_j)$, if they are nonconsecutive in S (*i.e.*, $|i-j| \geq 2$) but adjacent in the lattice. To be consistent with the notation of the field, the free-energy function in the HP model is defined as the negative of HHtc; maximizing HHtc is equivalent to minimize such an energy function.

Formally, PSP in the HP model is defined as the problem of finding the conformation $c^* \in C(S)$ such that $E_{D85}(c^*) = \min\{E_{D85}(c) \mid c \in C(S)\}$, where $C(S)$ is the set of all valid conformations of S. $E_{D85}(c)$ denotes the free energy of conformation c, which is given by:[1]

$$E_{D85}(c) = \sum_{s_i, s_j \in S_H} e(s_i, s_j) \quad (1)$$

where

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

[1]The acronym D85 is used to distinguish this conventional function from the other approaches considered in this study.

An example of the optimal conformation for an HP protein of length $L = 20$ on the square lattice is shown in Figure 3.

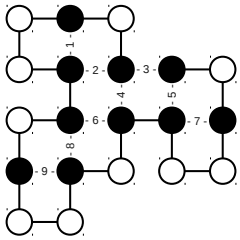

Fig. 3. Optimal conformation for sequence HPHPPHHPHPPHPHHPPHPH of length $L = 20$. Black and white balls denote H and P residues, respectively. H-H topological contacts (HHtc) have been numbered. The free energy of this conformation is $E_{D85}(c) = -9$, since HHtc $= 9$.

Despite its apparent simplicity, finding the optimal conformation for a protein in the HP model represents a hard combinatorial optimization problem. This problem has been proved to be $\mathcal{NP}$-complete [8, 9], which justifies the diversity of metaheuristic-based approaches that have been adopted to address it [10].

III. Alternative HP energy functions

This section describes the alternative HP energy functions considered for this study. A three-letter acronym has been assigned to each of the studied approaches. The acronyms are based on first author's initial and publication year.

A. Krasnogor et al., 1999 (K99)

In the conventional HP free-energy function, only H-H topological contacts (HHtc) contribute to the quality assessment of conformations. Given two conformations with the same HHtc value, it is possible, however, that one of them has better characteristics than the other. An example of this scenario is shown in Figure 4.

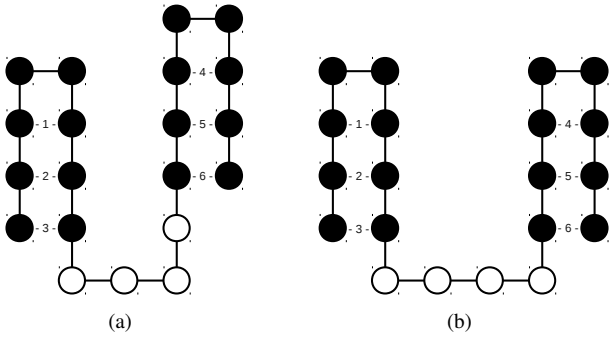

Fig. 4. Two conformations with the same HHtc value (HHtc $= 6$). However, the structure in (a) is more compact and, therefore, closer to the optimal conformation. This example was taken from [11].

Based on this observation, Krasnogor *et al.* [11] proposed the following distance-dependent energy function:

$$E_{K99}(c) \quad = \quad \sum_{s_i, s_j \in S_H} e(s_i, s_j) \qquad (2)$$

where

$$e(s_i, s_j) \quad = \quad \begin{cases} -1 & \text{if } tc(s_i, s_j) \\ -1/(d(s_i, s_j)^k |S_H|) & \text{otherwise} \end{cases}$$

where $d(s_i, s_j)$ denotes the distance between residues s_i and s_j. In [11], the value of k was set to $k = 4$ for the square lattice and $k = 5$ for the cubic and triangular lattices.

This approach preserves the rank order of the conformations as imposed by the conventional function, at the same time it enables a finer level of distinction among conformations with the same HHtc value [11]. In [11], no significant improvements in performance were achieved when using the modified energy function. However, as the authors pointed out, the superiority of this function is expected to become more evident for larger instances and, particularly, when local search strategies are implemented. The relevance of using this proposal needs to be further investigated.

B. Lopes and Scapin, 2006 (L06)

Lopes and Scapin [12] proposed an energy function which is based on the concept of *radius of gyration*. The radius of gyration is a measure of the compactness of conformations; the more compact the conformation, the smaller the value for this measure. The proposed function is given by:

$$E_{L06}(c) \quad = \quad HnLB \cdot RadiusH \cdot RadiusP \qquad (3)$$

The $HnLB$ term comprises the number of H-H topological contacts (HHtc) and a penalty factor which accounts for the violation of the self-avoiding constraint. Formally:

$$HnLB \quad = \quad \text{HHtc} - (NC \cdot PW) \qquad (4)$$

where NC is the number of collisions (*i.e.*, lattice nodes assigned to more than one residue) and PW is the penalty weight. The value of PW depends on the chain length, L, and can be computed as $PW = (0.033 \cdot L) + 1.33$ [15].

Before defining the $RadiusH$ and $RadiusP$ terms, let us first define RgH as the radius of gyration for H residues:

$$RgH \quad = \quad \sqrt{\frac{\sum_{s \in S_H} \left[(x_s - \bar{X})^2 + (y_s - \bar{Y})^2 \right]}{|S_H|}} \qquad (5)$$

where x_s and y_s are the lattice coordinates of residue s while $\bar{X}$ and $\bar{Y}$ denote the arithmetic mean of the coordinates of all H residues. Analogously, we can compute RgP, the radius of gyration for P residues, by considering only P rather than H residues in (5).

Once RgH has been defined, the $RadiusH$ term measures how compact the hydrophobic core of the conformation is, as given by:

$$RadiusH \quad = \quad MaxRgH - RgH \qquad (6)$$

where $MaxRgH$ is the radius of gyration of a totally unfolded conformation; *i.e.*, the maximum possible RgH value.

Finally, the $RadiusP$ term aims to push P residues away from the hydrophobic core. Given the previously defined RgH and RgP measures, the $RadiusP$ term is computed as:

$$RadiusP = \begin{cases} 1 & \text{if } (RgP - RgH) \geq 0 \\ \frac{1}{1-(RgP-RgH)} & \text{otherwise} \end{cases} \quad (7)$$

The $RadiusP$ term will always lie in the range $[0, 1]$. A value of $(RgP - RgH) > 0$ means that P residues are more exposed than H residues. This is a convenient scenario, so the $RadiusP$ term has no contribution to the final energy value ($RadiusP = 1$). Otherwise, $(RgP - RgH) < 0$ suggests H residues to be more spread than the P ones, so the energy value of the conformation is decreased. Note that (3) is to be maximized.

In [12, 15], the authors argue that the above described function provides an adequate discrimination among conformations with the same HHtc value, which makes the search landscape smoother. However, no results are provided on the impact of using this function rather than the conventional approach.

C. Berenboym and Avigal, 2008 (B08)

Berenboym and Avigal [13] proposed an alternative energy function, called by them the *global energy*. In this function, each pair of nonconsecutive H residues contributes to the energy value, even if they are not topological neighbors. The global energy for a given conformation c is defined as:

$$E_{B08}(c) = \sum_{s_i, s_j \in S_H} e(s_i, s_j) \quad (8)$$

where

$$e(s_i, s_j) = \begin{cases} \frac{-1}{(x_{s_i} - x_{s_j})^2 + (y_{s_i} - y_{s_j})^2} & \text{if } |i - j| \geq 2 \\ 0 & \text{otherwise} \end{cases}$$

In [13], the effects of using a local search operator within a genetic algorithm were investigated for both, the conventional and the proposed energy functions. However, an explicit comparison to demonstrate the advantages of using a particular energy function was not reported.

D. Islam and Chetty, 2009 (I09)

Islam and Chetty [14, 16] implemented into their memetic algorithm a modified energy function which incorporates two additional measures: *H-compliance* and *P-compliance*.

H-compliance measures the proximity of H residues to the center of a hypothetical rectangle enclosing all H residues, which is denoted by the reference point (x_r, y_r). Formally, this measure is given by:

$$H\text{-}comp(c) = \frac{\sum_{s \in S_H} (x_r - x_s)^2 + (y_r - y_s)^2}{|S_H|} \quad (9)$$

where x_s and y_s denote the coordinates of the s residue.

P-compliance is a measure of how close P residues are to the boundaries of a hypothetical rectangle enclosing all P residues. Such a rectangle is defined by x_{min}, x_{max}, y_{min} and y_{max}. The *P-compliance* measure is formally given by:

$$P\text{-}comp(c) = \frac{\sum_{s \in S_P} \min \left\{ \begin{array}{l} |x_{min} - x_s|, |x_{max} - x_s|, \\ |y_{min} - y_s|, |y_{max} - y_s| \end{array} \right\}}{|S_P|} \quad (10)$$

Finally, the energy of a given conformation c is defined as:

$$E_{I09}(c) = \alpha E_{D85}(c) + H\text{-}comp(c) + P\text{-}comp(c) \quad (11)$$

where $E_{D85}(c)$ is the conventional HP energy function (as defined in Section II-C) and α is a high value integer constant to ensure this will be the dominant term in (11). The value of $\alpha = 10,000$ was used in this study.

In [14], the advantages of using the proposed energy function were demonstrated for a 85-length HP benchmark sequence. However, the impact of using this function should be carefully investigated for a larger set of test cases.

IV. Experimental Results

In Section III, several alternative approaches to evaluate the quality of conformations in the HP model have been described. The aim of this experimental phase is to investigate the effectiveness of such approaches. It is important to note, however, that even when an alternative evaluation function is used, the goal of the optimization process remains to maximize HHtc, which is the singular objective in the HP model (see Section II-C). In this study, the exclusive purpose for using alternative formulations of the energy function is to guide the search process in a more effective manner.

Table I presents the 15 HP benchmark sequences adopted for this study. These benchmarks are commonly used in studies focusing on the two-dimensional square lattice.

TABLE I
HP benchmark sequences for the 2-dimensional square lattice. Sequence length (L). Optimal HHtc value (HHtc*).

	Sequence	L	HHtc*
S1	$H_2P_5H_2P_3HP_3HP$	18	4
S2	$HPHPH_3P_3H_4P_2H_2$	18	8
S3	$PHP_2HPH_3PH_2PH_5$	18	9
S4	$HPHP_2H_2PHP_2HPH_2P_2HPH$	20	9
S5	$H_3P_2HPHPHP_2HPHPHP_2H$	20	10
S6	$H_2P_2HP_2HP_2HP_2HP_2HP_2HP_2H_2$	24	9
S7	$P_2HP_2H_2P_4H_2P_4H_2P_4H_2$	25	8
S8	$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$	36	14
S9	$P_2HP_2H_2P_2H_2P_5H_{10}P_6H_2P_2H_2P_2HP_2H_5$	48	23
S10	$H_2(PH)_4H_3P(HP_3)_3(P_3H)_3PH_4(PH)_4H$	50	21
S11	$P_2H_3PH_8P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$	60	36
S12	$H_{12}PHPH(P_2H_2P_2H_2P_2H)_3PHPH_{12}$	64	42
S13	$H_4P_4H_{12}P_6(H_{12}P_3)_3HP_2H_2P_2H_2P_2HPH$	85	53
S14	$P_6HPH_2P_5H_3PH_5PH_2P_4H_2P_2H_2PH_5P$ $H_{10}PH_2PH_7P_{11}H_7P_2HPH_3P_6HPH_2$	100	48
S15	$P_3H_2P_2H_4P_2H_3PH_2PH_2PH_4P_8H_6P_2H_6$ $P_9HPH_2PH_{11}P_2H_3PH_2PHP_2HPH_3P_6H_3$	100	50

A. Degree of discrimination

The discrimination strategy is a very important component which directly impacts the performance of search algorithms. That is, if it is not possible to discriminate among solutions, then the search process will be guided practically at random.

In this section, the degree of discrimination that each of the studied approaches provides is investigated. This is done by analyzing the distribution of ranks that these functions induce on a set of solutions. A ranking expresses the relationship among a set of items according to a given property. In the context of this study, potential conformations are to be ranked and the property to set such a relationship corresponds to their quality. Given a set of solutions, we can get a ranking by assigning the first rank to the best solution, the next rank to the second best solution, and so on. If two or more solutions present the same quality, they will share the same rank.

The *relative entropy* (RE) measure proposed by Corne and Knowles [17] was adopted. Given a set of n ranked solutions (there are at most n ranks, and at least 1), the relative entropy of the distribution of ranks D is defined as:

$$\mathrm{RE}(D) = \frac{\sum_{r} \frac{D(r)}{n} \log(\frac{D(r)}{n})}{\log(1/n)} \quad (12)$$

where $D(r)$ denotes the number of solutions with rank r. $\mathrm{RE}(D)$ tends to 1 as approaching to the ideal situation where each solution has a different rank (*i.e.*, a total ordering of the solutions). On the other hand, when all the solutions share the same ranking position (*i.e.*, the poorest discrimination), $\mathrm{RE}(D)$ takes a value of zero.

In this experiment, $1,000$ different valid structures were generated at random. For each of the studied energy functions, these solutions were evaluated and ranked to finally compute the RE measure. A total of 100 repetitions of this experiment were performed for all the adopted benchmarks. The box plots in Figure 5 present the overall statistics of this experiment.

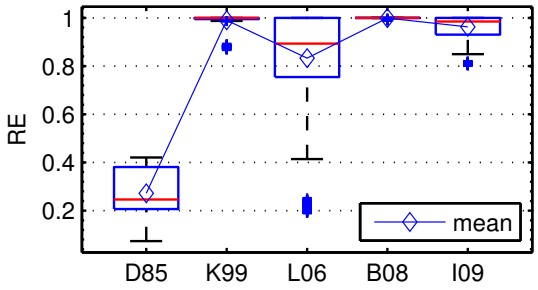

Fig. 5. Relative entropy (RE) of the distribution of ranks obtained by the studied energy functions. Overall statistics for all test cases.

From Figure 5, it can be seen that some of the studied approaches discriminate stronger than others. It is possible to note that the conventional HP energy function, D85, achieved the lowest RE values. These results confirm the poor discrimination capabilities of this function, which has been the main factor motivating the exploration of alternative approaches. Function L06 reached high RE values most of the time. However, the outliers indicate a low performance of this function for some of the adopted benchmarks. Regarding I09, note that the RE values obtained by this function were almost always above 0.9, which is a strong discrimination. Finally, it is important to remark the high degree of discrimination provided by functions B08 and K99. B08 achieved the best RE values, followed by function K99 whose outliers indicate some slight decreases.

The above results can be better understood by analyzing the histograms with the distribution of ranks achieved by each function. Figure 6 presents such histograms for the first repetition of this experiment regarding sequence S4.

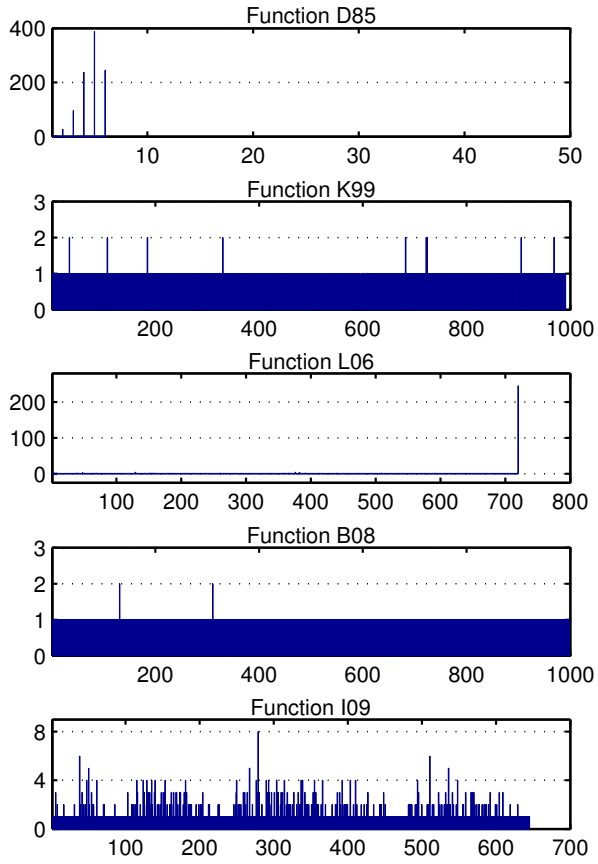

Fig. 6. Histograms showing the density of the distribution of ranks achieved by the studied energy functions. Sequence S4, run 1.

From Figure 6, it is possible to note how poor the distribution of ranks achieved by function D85 is. Only five different ranking positions were enough to classify the $1,000$ generated solutions. It can be seen that almost 400 solutions are sharing the same ranking position. In fact, no matter the amount of generated solutions, the maximum number of ranks which can be assigned through function D85 is 9, since $\mathrm{HHtc}^* = 9$ for this benchmark sequence (S4).

Functions L06 and I09 enabled an increased discrimination, since roughly 720 and 650 different ranking positions were occupied to classify the totality of solutions, respectively. In the case of function I09, a maximum of eight solutions were assigned to the same rank. On the other hand, the histogram for L06 indicates that there are about 250 equally ranked conformations. Function L06 is defined as the product of three terms, out of which one corresponds to HHtc (see Section III-B). All solutions for which $\mathrm{HHtc} = 0$ will have the same energy value, 0. To some extent, this can be seen as

a drawback. Function L06 will not be able to discriminate among these solutions even if some of them have better chances than others to further improve.

Finally, the histograms for B08 and K99 confirm the high discrimination potential of these approaches. B08 exhibited the strongest discrimination among all the studied functions. The corresponding histograms for these functions reveal that almost all solutions were mapped to a different ranking position. Only a few ranks were assigned to at most two solutions.

B. Search performance

A basic Memetic Algorithm (MA) was used in order to evaluate the effectiveness of the studied energy functions at guiding the search process. MAs were first introduced by Moscato in 1989 [18]. The general idea behind MAs is the use of a local search heuristic within a population-based technique. The implemented MA is based on a Genetic Algorithm. Local search is applied to the new individuals generated from the standard genetic operators. Algorithm 1 describes such an MA.

Algorithm 1 The implemented Memetic Algorithm.

BEGIN MemeticAlgorithm()
1: $P \leftarrow GenerateInitialPopulation()$
2: **while** $(stop\ condition)$ **do**
3: *Selection-for-variation(*P*)*
4: $P' \leftarrow \begin{cases} Recombination() \\ Mutation() \\ LocalSearch() \end{cases}$
5: $P \leftarrow$ *Selection-for-survival(*$P \cup P'$*)*
6: **end while**
END

At the beginning, an initial parent population P of different valid structures is generated at random. At each generation, the fittest individuals in P are selected for mating (*selection-for-variation*). A children population P' is generated by applying the variation operators to the selected parents. Then, local search is applied to a randomly selected individual from P'. Finally, parents and children compete for a place in the new population (*selection-for-survival*). This process repeats until a stop condition is met. Implementation details are as follows:

- **Encoding.** An internal coordinates representation with absolute moves was adopted [19]. That is, candidate conformations are encoded as sequences in $\{U, D, L, R\}^{L-1}$, denoting the up, down, left and right possible locations for a residue with regard to the preceding one. Solutions are decoded to Cartesian coordinates for evaluation.
- **Selection.** *Selection-for-variation* is performed by means of binary tournament. In the *selection-for-survival* process, the best individuals from the union of parents and children are selected (duplicates are not allowed). It is through these processes that the different energy functions will play a decisive role in the behavior of the MA.
- **Recombination.** One-point crossover is applied with a probability of 0.8. If invalid children are generated, then parents are copied unchanged.
- **Mutation.** Uniform mutation is applied to all individuals in P'. Each encoding position is randomly mutated with a probability of $1/(L-1)$. If mutation of a position leads to an invalid conformation, the original value is restored.
- **Local search.** The implemented local search operator is illustrated in Algorithm 2. This operator is applied to a randomly selected individual from P', denoted by c. At each iteration, a new conformation c' is generated through a single random change in the encoding of c. If c' has a better energy value than c ($E(c') < E(c)$), then c is replaced by c'. Only valid moves are accepted.

Algorithm 2 The local search operator.

BEGIN LocalSeach()
1: $c \leftarrow SelectRandomSolution(P')$
2: **while** $(stop\ condition)$ **do**
3: $c' \leftarrow RandomChange(c)$
4: **if** $valid(c')$ **and** $E(c') < E(c)$ **then**
5: $c \leftarrow c'$
6: **end if**
7: **end while**
END

It is important to remark that the aim of using the above described MA is not to improve the state-of-the-art results for this problem. In this study, the MA serves only as a tool to measure the impact of using each of the studied functions.

The population size was set to $|P| = 60$ individuals. The stopping condition for the local search operator was fixed to 25 iterations. The MA was allowed to run until a maximum number of $200,000$ function evaluations was reached. All the parameters were arbitrarily adjusted, so that the same values were used for all the approaches here compared. For each benchmark, 30 independent executions were performed.

The results of this experiment are presented in Table II. For each benchmark sequence, this table shows the best obtained HHtc value (β) as well as the number of times this solution was found (fx). Also, the mean (μ) and standard deviation (σ) are presented. Three additional measures have been defined in order to assess the overall performance of the studied approaches. First, the *Overall Average Proximity* (OAP) measure is defined as the average ratio of the obtained mean values (μ) to the optimum (HHtc*). Formally:

$$\text{OAP} = \frac{100}{|B|}\left(\sum_{S_i \in B} \frac{\mu(S_i)}{\text{HHtc}^*(S_i)}\right) \quad (13)$$

where B denotes the set of all benchmark sequences. Note that the average ratio in (13) is multiplied by 100 in order to express OAP as a percentage. Thus, a value of OAP = 100% would suggest the ideal situation where the optimum value was found in each of the 30 runs for all the benchmarks. The second measure is given by replacing the mean μ in (13) with the best found value β. Such a measure is to be referred to as the *Overall Best Proximity* (OBP). A value of OBP = 100% indicates that the optimum value for each benchmark

TABLE II

PERFORMANCE OF THE MEMETIC ALGORITHM WHEN USING THE DIFFERENT ENERGY FUNCTIONS. BEST FOUND VALUE (β) AND FREQUENCY (fx). MEAN (μ). STD. DEVIATION (σ). OVERALL AVERAGE PROXIMITY (OAP). OVERALL BEST PROXIMITY (OBP). OVERALL SUCCESS COUNTER (OSC).

		D85			K99			L06			B08			I09		
Seq.	**HHtc***	β (fx)	μ	σ	β (fx)	μ	σ	β (fx)	μ	σ	β (fx)	μ	σ	β (fx)	μ	σ
S1	**4**	4 (10x)	3.3	0.5	4 (22x)	3.7	0.4	4 (16x)	3.5	0.5	4 (11x)	3.4	0.5	4 (22x)	3.7	0.4
S2	**8**	8 (20x)	7.7	0.5	8 (25x)	7.8	0.4	8 (20x)	7.7	0.5	8 (24x)	7.8	0.4	8 (24x)	7.8	0.4
S3	**9**	9 (6x)	8.2	0.5	9 (4x)	8.1	0.3	9 (10x)	8.3	0.5	9 (1x)	8.0	0.2	9 (5x)	8.2	0.4
S4	**9**	9 (24x)	8.8	0.4	9 (27x)	8.9	0.3	9 (25x)	8.8	0.4	9 (23x)	8.7	0.6	9 (25x)	8.8	0.4
S5	**10**	10 (15x)	9.0	1.0	10 (18x)	9.2	1.0	10 (19x)	9.3	0.9	10 (20x)	9.3	0.9	10 (24x)	9.6	0.8
S6	**9**	9 (3x)	7.7	0.7	9 (14x)	8.4	0.7	9 (21x)	8.7	0.5	9 (21x)	8.7	0.5	9 (20x)	8.6	0.5
S7	**8**	8 (8x)	6.7	1.0	8 (14x)	7.3	0.7	8 (12x)	7.1	0.9	8 (14x)	7.3	0.8	8 (16x)	7.5	0.6
S8	**14**	12 (4x)	10.4	1.1	14 (1x)	10.4	1.3	13 (1x)	10.2	1.1	14 (1x)	10.7	1.4	14 (1x)	10.5	1.4
S9	**23**	19 (2x)	15.5	1.9	19 (1x)	15.8	1.6	19 (1x)	16.0	1.3	21 (1x)	16.4	1.8	20 (1x)	16.2	1.9
S10	**21**	19 (1x)	15.6	1.3	20 (1x)	16.2	1.4	18 (2x)	15.5	1.1	17 (3x)	14.4	1.9	18 (4x)	16.3	1.1
S11	**36**	31 (1x)	26.8	2.2	33 (1x)	27.6	2.5	32 (1x)	27.8	2.0	31 (1x)	26.1	2.4	31 (3x)	27.6	2.0
S12	**42**	30 (3x)	26.9	1.6	30 (2x)	27.4	1.6	32 (1x)	27.8	1.9	30 (1x)	24.7	1.9	30 (3x)	27.2	1.9
S13	**53**	43 (1x)	37.5	2.3	43 (3x)	38.7	2.7	42 (3x)	37.9	2.5	41 (2x)	35.5	2.8	44 (1x)	38.7	2.4
S14	**48**	35 (3x)	31.3	2.4	35 (2x)	31.0	1.9	36 (1x)	31.6	2.5	35 (2x)	30.0	2.7	36 (1x)	31.8	2.0
S15	**50**	36 (1x)	31.7	2.1	38 (1x)	32.8	2.3	37 (1x)	32.0	2.5	35 (1x)	28.6	3.1	36 (2x)	32.3	2.1
OAP		78.72%			81.51%			80.91%			78.82%			**82.27%**		
OBP		89.49%			**91.40%**			90.30%			90.00%			90.68%		
OSC		86			125			123			115			**137**		

was reached at least once. Finally, the third defined measure is the *Overall Success Counter* (OSC), which refers to the total number of times where the optimal HHtc value was found considering all the test cases.

From Table II, it can be seen that the use of alternative energy functions significantly impacted the performance of the adopted MA. In most cases, the alternative functions allowed better solutions to be reached or, at least, to increase the success rate regarding the best found values.

The conventional function D85 exhibited the worst average performance according to the OAP measure. However, the OAP value for function B08 is almost as low as that for D85. In spite of the outstanding behavior of function B08 at solving benchmarks S2, S6, S8 and particularly S9, this approach obtained the lowest mean value (μ) for 8 out of the 15 test cases (S3, S4, S10-S15). These results suggest a slightly inconsistent behavior of function B08. On the other hand, the advantages of using functions I09, K99 and L06 are more evident. These approaches allowed to increase the average performance of the MA in most cases. Function I09 achieved the highest μ value for 7 out of the benchmarks, thus being the best performer with regard to the OAP measure.

The results are slightly different when focusing on the OBP measure. This measure suggests that function K99 performed the best, reaching the best found HHtc value (β) for 11 out of the 15 test cases (S1-S8, S10, S11 and S15). The I09 and L06 approaches achieved the highest β value for 10 and 9 of the benchmarks, being the second and third best performers in terms of OBP, respectively. Although B08 obtained the best β value for sequences S1 to S9, its β values were even worse than those of function D85 for some of the remaining test cases (S10, S13 and S15). The worst behavior regarding the OBP measure was shown by the conventional function D85.

Finally, the OSC measure makes the benefits of using alternative energy functions more evident. Even B08, the worst performer among the alternative approaches, reached the optimum HHtc value $(115 - 86) = 29$ more times than the conventional function D85. Functions K99 and L06 achieved the optimum value in 39 and 37 more runs than function D85, respectively. The best performer in terms of the OSC measure is function I09, allowing the optimum solution to be found in $(137 - 85) = 51$ more cases than the conventional approach.

V. CONCLUSIONS AND FUTURE WORK

The HP model for protein structure prediction captures the fact that hydrophobicity is the dominant factor which determines the functional conformation of proteins. Despite its level of abstraction, this problem has been proved to be $\mathcal{NP}$-complete and constitutes a hard combinatorial optimization task. Such a complexity represents the main motivation for the use of metaheuristic algorithms to address this problem.

The conventional energy function of the HP model provides a very poor discrimination among potential conformations. Nevertheless, an effective evaluation scheme is an essential requirement for metaheuristics in order to drive the search process towards promising regions of the solutions space. Alternative HP energy functions have been proposed in the literature to enhance the performance of search algorithms. However, for most of these approaches there are not reported results where the benefits of their usage are demonstrated.

This paper presented the results of a comparative study where five different formulations of the HP energy function were considered. Our first experiment focused on the degree of discrimination that each of the studied functions provides. The obtained results confirmed the poor discrimination capabilities of the conventional function D85, which has been

the main motivation for exploring alternative approaches. All the alternative functions demonstrated to provide a more fine-grained discrimination than the conventional approach. The most discriminative function according to our results is B08, followed by the K99 and I09 approaches, in this order.

In our second experiment, we evaluated the impact of using the different functions on the behavior of a basic memetic algorithm (MA). In general, the use of alternative energy functions improved the performance of the implemented MA. These approaches allowed better solutions to be reached or to increase the frequency of the times where the best found values were achieved. Three measures were defined in order to assess the overall performance of the approaches. As expected, the conventional function D85 presented the worst overall behavior. However, the average performance of function B08 was almost as poor as that of D85. Functions I09 and K99 exhibited the best performance according to the adopted measures.

From this study, it is possible to conclude that intensity of discrimination does not imply effectiveness at guiding the search process. Even when B08 provides the strongest discrimination, this function presented a limited search performance. That function D85 consistently exposed the worst behavior confirmed, however, that a tighter evaluation scheme is important to increase the performance of search algorithms. These results support the relevance of exploring the use of alternative approaches. Functions I09 and K99 were the best performing approaches according to all the adopted criteria.

To the best of our knowledge, this is the first study that has been reported in this direction. Nevertheless, this research is in progress. The preliminary results presented in this paper suggest that functions I09 and K99 are very promising approaches for studies on the HP model. However, the impact of using these functions needs to be further investigated for different search metaheuristics. In this study, all the parameters of the adopted MA were arbitrarily adjusted. A fairer approach would be to tune the algorithm independently for each function before the comparison. Also, it is important to extend this study to other lattice configurations in order to generalize our conclusions. Finally, there are additional HP energy functions reported in the literature (see, for example, [20, 21]) to be considered in an extended version 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. We would like to acknowledge support from CONACyT through projects 105060 and 99276. Also, this research was partially funded by project number 51623 from "Fondo Mixto Conacyt-Gobierno del Estado de Tamaulipas". Finally, we would like to thank to "Fondo Mixto de Fomento a la Investigación científica y Tecnológica CONACyT - Gobierno del Estado de Tamaulipas" for their support to publish this paper.

REFERENCES

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

[2] A. Kolinski and J. Skolnick, "Reduced Models of Proteins and their Applications," *Polymer*, vol. 45, no. 2, pp. 511–524, 2004.

[3] W. E. Hart and A. Newman, "Protein Structure Prediction with Lattice Models," in *Handbook on Computational Molecular Biology*, S. Aluru, Ed. Chapman and Hall/CRC Computer and Information Science Series, 2005.

[4] C. Clementi, "Coarse-grained Models of Protein Folding: Toy Models or Predictive Tools?" *Current Opinion in Structural Biology*, vol. 18, no. 1, pp. 10–15, 2008.

[5] C. L. Pierri, A. De Grassi, and A. Turi, "Lattices for Ab Initio Protein Structure Prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 73, no. 2, pp. 351–361, 2008.

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

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

[8] B. Berger and T. Leighton, "Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-complete," in *RECOMB '98: Proceedings of the second annual international conference on Computational molecular biology*. New York, NY, USA: ACM, 1998, pp. 30–39.

[9] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis, "On the Complexity of Protein Folding," in *STOC '98: Proceedings of the thirtieth annual ACM symposium on Theory of computing*. New York, NY, USA: ACM, 1998, pp. 597–603.

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

[11] N. Krasnogor, W. E. Hart, J. Smith, and D. A. Pelta, "Protein Structure Prediction With Evolutionary Algorithms," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 1999)*, W. Banzhaf, J. M. Daida, A. E. Eiben, M. H. Garzon, V. Honavar, M. J. Jakiela, and R. E. Smith, Eds. Orlando, Florida, USA: Morgan Kaufman, July 1999.

[12] 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, E.-G. Talbi, P. Liardet, P. Collet, E. Lutton, and M. Schoenauer, Eds. Springer Berlin / Heidelberg, 2006, vol. 3871, pp. 238–246.

[13] I. Berenboym and M. Avigal, "Genetic Algorithms with Local Search Optimization for Protein Structure Prediction Problem," in *GECCO '08: Proceedings of the 10th annual conference on Genetic and evolutionary computation*. New York, NY, USA: ACM, 2008, pp. 1097–1098.

[14] K. Islam and M. Chetty, "Novel Memetic Algorithm for Protein Structure Prediction," in *AI 2009: Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, A. Nicholson and X. Li, Eds. Springer Berlin / Heidelberg, 2009, vol. 5866, pp. 412–421.

[15] H. Lopes and M. Scapin, "A Hybrid Genetic Algorithm for the Protein Folding Problem Using the 2D-HP Lattice Model," in *Success in Evolutionary Computation*, ser. Studies in Computational Intelligence, A. Yang, Y. Shan, and L. Bui, Eds. Springer Berlin / Heidelberg, 2008, vol. 92, pp. 121–140.

[16] K. Islam and M. Chetty, "Clustered Memetic Algorithm for Protein Structure Prediction," in *IEEE Congress on Evolutionary Computation, CEC 2010*, Barcelona, Spain, July 2010.

[17] D. Corne and J. Knowles, "Techniques for Highly Multiobjective Optimisation: Some Nondominated Points are Better than Others," in *2007 Genetic and Evolutionary Computation Conference (GECCO'2007)*, D. Thierens, Ed., vol. 1. London, UK: ACM Press, July 2007, pp. 773–780.

[18] P. Moscato, "On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms," Caltech Concurrent Computation Program, Pasadena, CA, Tech. Rep. C3P Report 826, 1989.

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

[20] 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*, vol. 27, pp. 611–615, 2004.

[21] M. Cebrián, I. Dotú, P. Van Hentenryck, and P. Clote, "Protein Structure Prediction on the Face Centered Cubic Lattice by Local Search," in *Proceedings of the 23rd national conference on Artificial intelligence - Volume 1*. AAAI Press, 2008, pp. 241–246.