# Comparative Study of Alternative Energy Functions for the HP Model of Protein Structure Prediction

Mario Garza-Fabre, Gregorio Toscano-Pulido and Eduardo Rodriguez-Tello
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, gtoscano, ertello}@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 which has been widely addressed through metaheuristics. However, the conventional energy function of the HP model does not provide an adequate discrimination among candidate conformations, which is an essential requirement for metaheuristics in order to perform an effective search. Therefore, alternative HP 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. The discrimination potential of each of the studied functions is analyzed as well as their impact on the behavior of a basic local search algorithm. The obtained results support the relevance of exploring this research direction.**

## I. Introduction

Proteins are at the heart of cellular function, carrying most of the key processes associated with life. The functional properties of a protein are dictated by its three-dimensional conformation. To fully understand the biological roles of a protein it is imperative, therefore, to determine its structure.

The *Protein Structure Prediction* (PSP) problem aims to determine the native conformation of proteins given only their linear chain of amino acids. Such a structure is assumed to be the one minimizing the overall free energy; *i.e.*, the thermodynamically most stable state [1]. Solving PSP at atomic resolution requires a prohibitive computational effort even for relatively small proteins. Thus, simplified protein models have emerged as valuable tools for studying the most general and essential principles governing the protein folding process.

One of such simplified formulations of PSP is the HP model [2, 3]. 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 [4, 5]. Such a complexity has motivated the use of a variety of metaheuristic-based approaches to address this problem [6].

Metaheuristics rely on an effective evaluation scheme to guide the search towards promising regions of the solutions space. However, the conventional energy function of the HP model features a very poor discrimination ability. Thus, no preferences can be set among potential conformations, leading the search process to be oriented almost at random. This problem is expected to have a major impact on the performance of local search-based algorithms. The low discrimination of the conventional HP energy function produces large plateaus in the energy landscape, on which local search strategies could fail to detect a promising search direction [7].

Alternative HP energy functions have been proposed to improve the performance of search algorithms [7]–[12]. Nevertheless, there are no reported results on the advantages of using most of such approaches. In this paper, a comparative study is presented where seven different formulations of the HP energy function are considered. The discrimination potential of these approaches is first analyzed. Then, the effectiveness of each of the studied functions to guide the search process is evaluated. A basic local search algorithm was adopted for this sake.

The remainder of this paper is organized as follows. The problem statement is provided in Section II. In Section III, the alternative HP energy functions considered 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. The HP model

Amino acids, the building blocks of proteins, 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 folded state of proteins.

In the Hydrophobic-Polar (HP) model, proposed by Dill in 1985 [2, 3], 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

This research was partially funded by CONACyT project 99276.

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 HP model aims 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. 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

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

An example of the optimal conformation for an HP protein of length $L = 20$ on the square lattice is shown in Figure 1.

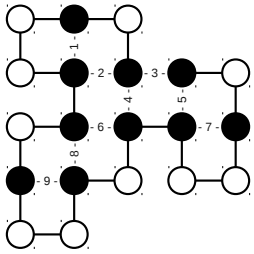

Fig. 1. Optimal conformation for sequence HPHPPHHPHPPHPHHPPHPH of length $L = 20$. Black and white balls denote $H$ and $P$ residues, respectively. $H$-$H$ topological contacts (HHtc) have been numbered. The free energy of this conformation is $E_{D85}(c) = -9$, since HHtc $= 9$.

Despite its apparent simplicity, finding the optimal conformation for a protein in the HP model is a hard combinatorial optimization problem, proved to be $\mathcal{NP}$-complete [4, 5].

## 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 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 (more compact) than the other. Krasnogor *et al.* [7] proposed the following distance-dependent energy function:

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

where

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

where $d(s_i, s_j)$ denotes the distance between residues $s_i$ and $s_j$. In [7], the value of $k = 4$ was used for the square lattice.

In [7], no significant improvements were achieved when using the modified energy function. As the authors pointed out, the superiority of the approach 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. Custódio et al., 2004 (C04)

Given that the aim of the conventional HP energy function is only to maximize $H$-$H$ interactions, the positioning of $P$ residues is not directly optimized. This may result in unnatural structures for sequences with long $P$ segments and, especially, when $P$ segments are located at the ends of the chain. An example is presented in Figure 2.

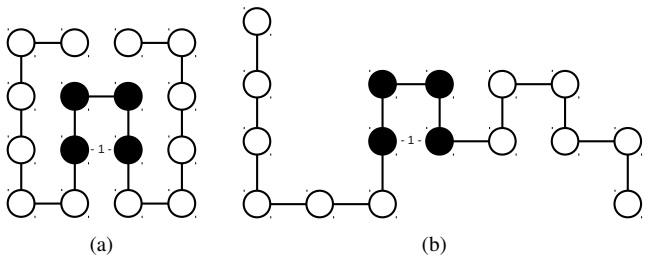

Fig. 2. Two conformations with the same HHtc value. However, the structure in (a) is more natural-like (globular) than the one in (b).

Custódio *et al.* [8] proposed a modified energy function based on the assumption that it may be preferable for an $H$ residue to have a $P$ neighbor than to be in contact with the aqueous solvent. In the proposed function, the energy of a conformation is computed as the weighted sum of the number of hydrophobic-hydrophobic ($HHc$), hydrophobic-polar ($HPc$) and hydrophobic-solvent contacts ($HSc$).[2] Formally:

$$E_{C04}(c) = \omega_1 HHc + \omega_2 HPc + \omega_3 HSc \quad (3)$$

where $\omega_1$, $\omega_2$ and $\omega_3$ denote the relative importance of $HHc$, $HPc$ and $HSc$. In [8], these weights were set to $\omega_1 = 0$, $\omega_2 = 10$ and $\omega_3 = 40$ for the reported experiments.

In [8], the proposed function allowed to improve the performance of a genetic algorithm for some of the adopted test cases. The reported results also suggest that this function presents a greater tendency to form more natural-like conformations (more compact and more segregated).

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

[2]A free lattice location is said to be occupied by the solvent.

### C. Lopes and Scapin, 2006 (L06)

Lopes and Scapin [9] 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:

$$E_{L06}(c) = HnLB \cdot RadiusH \cdot RadiusP \quad (4)$$

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:

$$HnLB = \text{HHtc} - (NC \cdot PW) \quad (5)$$

where $NC$ is the number of collisions and the penalty weight $PW$ can be computed as $PW = (0.033 \cdot L) + 1.33$ [13].

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

$$RgH = \sqrt{\frac{\sum_{s \in S_H} \left[(x_s - \bar{X})^2 + (y_s - \bar{Y})^2\right]}{|S_H|}} \quad (6)$$

where $x_s$ and $y_s$ are the coordinates of residue $s$ while $\bar{X}$ and $\bar{Y}$ denote the mean coordinates for $H$ residues. Analogously, we can compute $RgP$, the radius of gyration for $P$ residues, by considering only $P$ rather than $H$ residues in (6).

The $RadiusH$ term measures how compact the hydrophobic core of the conformation is. This term is given by:

$$RadiusH = MaxRgH - RgH \quad (7)$$

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:

$$RadiusP = \begin{cases} 1 & \text{if } (RgP - RgH) \geq 0 \\ \frac{1}{1-(RgP-RgH)} & \text{otherwise} \end{cases} \quad (8)$$

$RadiusP$ lies 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 (4) is to be maximized.

In [9, 13], no results are provided on the impact of using this function rather than the conventional approach.

### D. Berenboym and Avigal, 2008 (B08)

Berenboym and Avigal [10] 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:

$$E_{B08}(c) = \sum_{s_i, s_j \in S_H} e(s_i, s_j) \quad (9)$$

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 [10], 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.

### E. Cébrian et al., 2008 (C08)

Cébrian *et al.* [11] proposed an alternative formulation of the HP energy function which measures the deviation from the unit distance (*i.e.*, topological contact distance) for each pair of $H$ residues. Let $d(s_i, s_j)^2 = (x_{s_i} - x_{s_j})^2 + (y_{s_i} - y_{s_j})^2$ be the distance between residues $s_i$ and $s_j$, and let $dv(s_i, s_j) = d(s_i, s_j)^2 - 1$ denote its deviation from the unit distance. The energy value of a conformation $c$ is given by:

$$E_{C08}(c) = \sum_{s_i, s_j \in S_H} dv(s_i, s_j)^k \quad (10)$$

where $k \geq 1$ is a parameter of the function, whose larger values give more weight to unit distances. We used $k = 2$, since this value seems to provide the best behavior based on the results reported in [11]. $E_{C08}(c^*) = 0$ would refer to the ideal (potentially unrealistic) scenario where all pairs of $H$ residues are at a unit distance in conformation $c^*$. In [11], no experimental results were reported about the benefits of using the proposed energy function instead of the conventional one.

### F. Islam and Chetty, 2009 (I09)

Islam and Chetty [12] proposed a modified HP function based on two 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, denoted by the reference point $(x_r, y_r)$. Formally:

$$H\text{-}comp(c) = \frac{\sum_{s \in S_H} (x_r - x_s)^2 + (y_r - y_s)^2}{|S_H|} \quad (11)$$

where $x_s$ and $y_s$ denote the lattice 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, defined by $x_{min}$, $x_{max}$, $y_{min}$ and $y_{max}$. Formally:

$$P\text{-}comp(c) = \frac{\sum_{s \in S_P} \min \left\{ \begin{array}{l} |x_{min} - x_s|, |x_{max} - x_s|, \\ |y_{min} - y_s|, |y_{max} - y_s| \end{array} \right\}}{|S_P|} \quad (12)$$

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

$$E_{I09}(c) = \alpha E_{D85}(c) + H\text{-}comp(c) + P\text{-}comp(c) \quad (13)$$

where $E_{D85}(c)$ is the conventional HP energy function (see Section II) and $\alpha$ is a high value integer constant to ensure this will be the dominant term in (13). We used $\alpha = 10,000$.

In [12], 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, 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). 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 9 HP benchmark sequences adopted for this study. These benchmarks are commonly used in studies focusing on the two-dimensional square lattice.

TABLE I
HP benchmarks. ($L$): seq. length. (HHtc$^*$): optimal HHtc value.

| | Sequence | $L$ | HHtc$^*$ |
|---|---|---|---|
| **S1** | $HPHP_2H_2PHP_2HPH_2P_2HPH$ | 20 | 9 |
| **S2** | $P_2HP_2H_2P_4H_2P_4H_2P_4H_2$ | 25 | 8 |
| **S3** | $P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$ | 36 | 14 |
| **S4** | $P_2HP_2H_2P_2H_2P_5H_{10}P_6H_2P_2H_2P_2HP_2H_5$ | 48 | 23 |
| **S5** | $H_2(PH)_4H_3P(HP_3)_3(P_3H)_3PH_4(PH)_4H$ | 50 | 21 |
| **S6** | $P_2H_3PH_8P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$ | 60 | 36 |
| **S7** | $H_{12}PHPH(P_2H_2P_2H_2P_2H)_3PHPH_{12}$ | 64 | 42 |
| **S8** | $H_4P_4H_{12}P_6(H_{12}P_3)_3HP_2H_2P_2H_2P_2HPH$ | 85 | 53 |
| **S9** | $P_6HPH_2P_5H_3PH_5PH_2P_4H_2P_2H_2PH_5P$ $H_{10}PH_2PH_7P_{11}H_7P_2HPH_3P_6HPH_2$ | 100 | 48 |

### 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 set preferences among solutions the search process will be guided practically at random.

In this section, the degree of discrimination that each of the studied functions provides is investigated. We analyzed the distribution of ranks that these approaches induce on a set of candidate 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 ranked according to their quality. The first rank is assigned to the best solution, the next rank to the second best solution, and so on. Solutions with the same quality will share the same rank.

We adopted the *relative entropy* (RE) measure proposed by Corne and Knowles [14]. 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:

$$\mathrm{RE}(D) = \frac{\sum_r \frac{D(r)}{n} \log(\frac{D(r)}{n})}{\log(1/n)} \quad (14)$$

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.*, the maximum possible discrimination). 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. We performed $100$ repetitions of this experiment for all the adopted benchmarks. The box plots in Figure 3 present the overall statistics of this experiment.

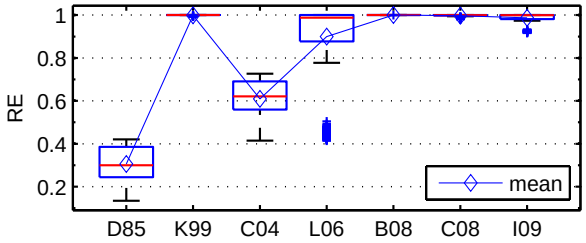


Fig. 3. Relative entropy (RE) of the distribution of ranks obtained by the studied energy functions. Overall statistics for all test cases.

From Figure 3, it is possible to note that the conventional HP energy function, D85, achieved the lowest RE values. This confirms the poor discrimination capabilities of this function, which has been the main factor motivating the exploration of alternative approaches. C04 showed the worst performance among the alternative functions. Function L06 achieved high RE values most of the time, but the outliers indicate a low performance of this function for some of the benchmarks. Finally, it is important to remark the high discrimination provided by functions B08, K99, C08 and I09.

The above results can be better understood by analyzing Figure 4. This figure presents the histograms with the distribution of ranks achieved by each function for the first repetition of this experiment regarding sequence S1. From this figure, 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 a peak where there are almost $400$ solutions sharing the same rank. In fact, no matter the amount of generated solutions, the maximum number of ranks which can be assigned through function D85 is 9, since HHtc$^* = 9$ for this benchmark sequence (S1). The second worst scenario is presented by function C04, where less than $40$ different ranking positions were required, out of which two were each assigned to at around 100 conformations.

Functions L06 and I09 showed an increased discrimination, since about $720$ and $650$ 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 presents a high peak indicating 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-C). All solutions for which 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, K99 and C08 confirm the high degree of discrimination that these approaches provide. We can see that function C08 allowed roughly $950$ different

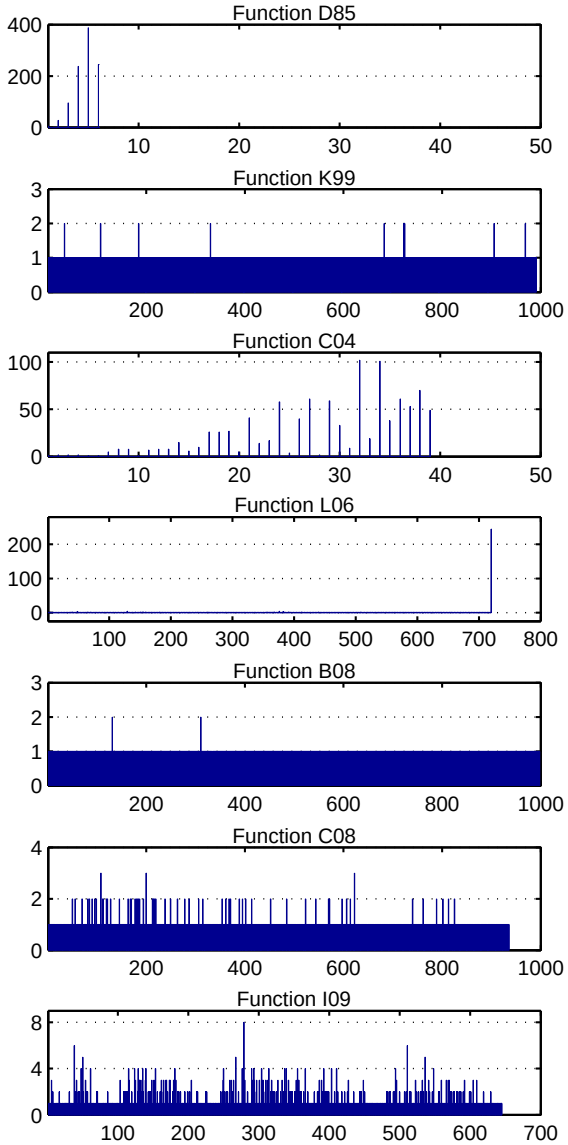


Fig. 4. Histograms showing the density of the distribution of ranks achieved by the studied evaluation functions. Sequence S1, run 1.

ranks to be assigned. B08 showed the strongest discrimination among all the studied functions, followed by K99. The corresponding histograms for these functions reveal that almost all solutions were mapped to a different rank. Only a few ranks were assigned to at most two solutions.

### B. Search performance

We implemented a *Steepest Descent Hill Climbing* algorithm (SDHC) to evaluate the effectiveness of the studied energy functions at guiding the search process. SDHC is a parameter-free algorithm, whose motivation in this study is to avoid affecting (neither negatively nor positively) the performance of the approaches through parameter settings. Given that SDHC is a local search technique, functions providing a finer discrimination are expected to perform better. As pointed out by Krasnogor *et al.* [7], a poor discrimination will produce large plateaus in the energy landscape, on which local search strategies could fail to identify a descent direction.

**Algorithm 1** Steepest Descent Hill Climbing (SDHC).

**BEGIN** SDHC()
1: $c \leftarrow getRandomValidSolution()$
2: **loop**
3: $c' \leftarrow getBest(N(c))$
4: **if** $E(c') < E(c)$ **then**
5: $c \leftarrow c'$
6: **else**
7: $Stop()$
**END**

Algorithm 1 describes the implemented SDHC. The algorithm starts with a valid conformation generated at random, denoted by $c$. Once $c$ is generated, we identify $c'$, the best conformation among all defined in the neighborhood of $c$, $N(c)$. Then, solutions $c$ and $c'$ are compared with respect to their energy values. At this point is where the different energy functions come to play a decisive role in the behavior of the algorithm. If $c'$ has a better energy value than $c$ ($E(c') < E(c)$), then a replacement occurs and the process repeats. Otherwise, the process ends, since given the current solution and the adopted neighborhood it is not possible to achieve an improvement.

An internal coordinates representation with absolute moves was adopted [15]. 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). The implemented neighborhood structure $N(c)$ is defined by all solutions that can be reached through 1-variable perturbations of $c$. Given a sequence of length $L$, the size of such a neighborhood is $|N(c)| = 3(L-1)$. However, only valid conformations are considered.

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

The behavior of the SDHC algorithm was evaluated when using each of the studied functions. A total of 100 independent executions were performed for all the adopted benchmarks. The results of this experiment are presented in Figure 5. Each plot in this figure shows the average number of $H$-$H$ topological contacts (HHtc) achieved by the algorithm as the search progressed (iteration by iteration).

From Figure 5, it is possible to derive some general conclusions. The poorest performance for this experiment was presented by function C08, whose results were even worse than those of function D85 in most of the considered test cases. This behavior can be explained by the fact that function C08 is not consistent with the conventional objective of the HP model. As stated at the beginning of Section IV, even when alternative functions are used to guide the search process, the goal remains to maximize HHtc; or, which is equivalent, to minimize function D85. The alternative function should not contradict D85 when discriminating among potential confor-

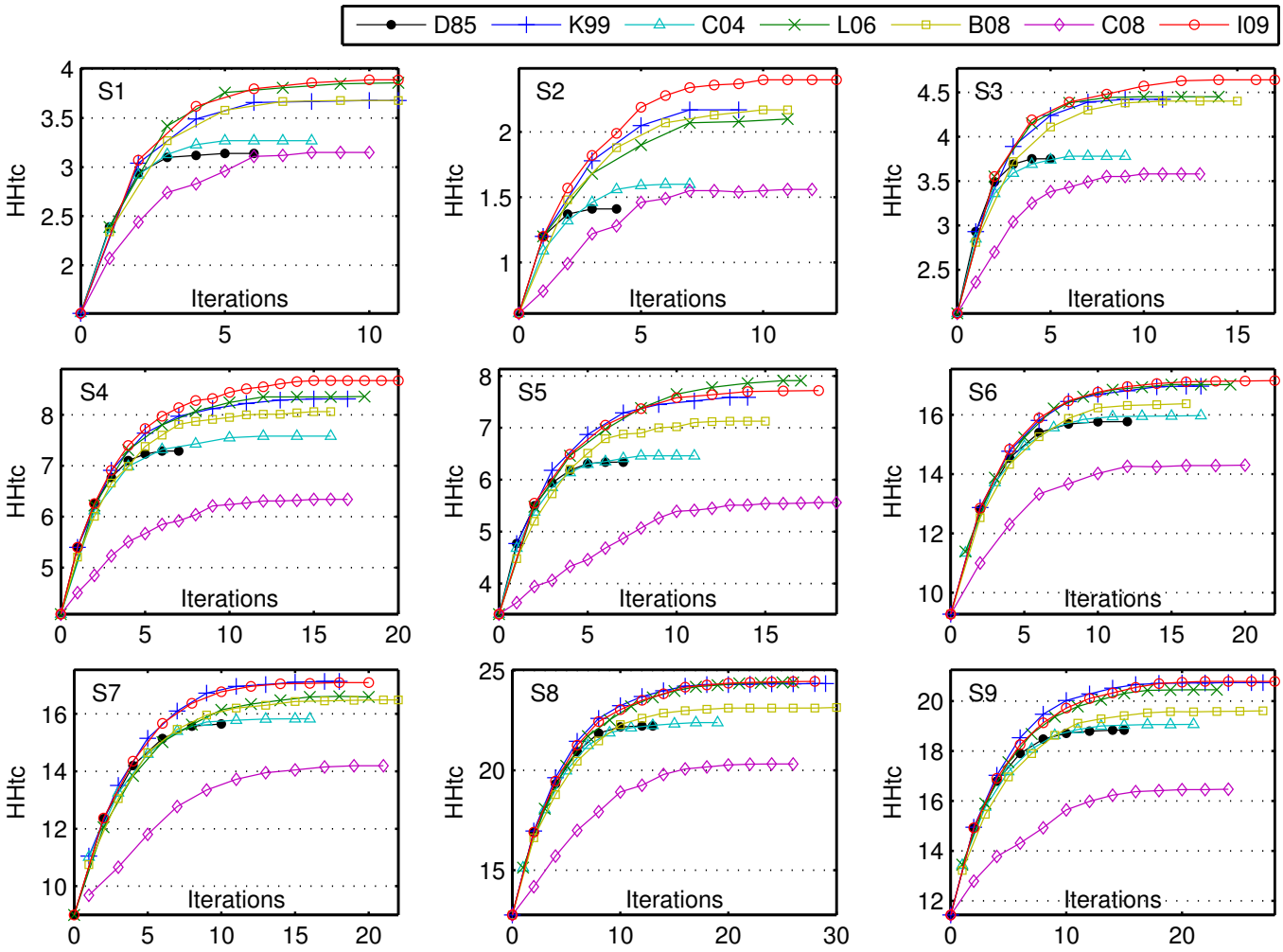


Fig. 5. Results of the SDHC algorithm. Achieved number of $H$-$H$ topological contacts (HHtc) at each iteration. Average of 100 independent executions.

mations, otherwise we will probably be pursuing a different optimum. Nevertheless, given two conformations $c_1$ and $c_2$, it is possible the case where $E_{D85}(c_1) < E_{D85}(c_2)$ but $E_{C08}(c_1) > E_{C08}(c_2)$, which is a contradiction.[3] An example of this scenario is presented in Figure 6. This can be seen as a drawback, so function C08 is not expected to steer the search in an effective manner. Such an important issue needs to be further explored for all the studied approaches.

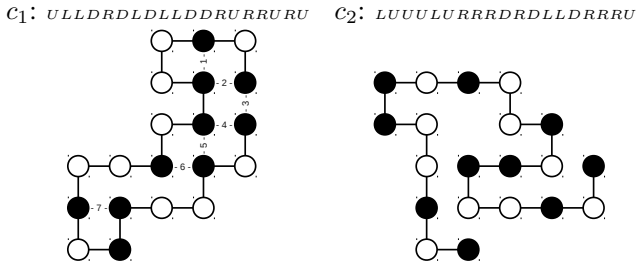


Fig. 6. C08 contradicts D85, since $E_{D85}(c_1) = -7 < E_{D85}(c_2) = 0$ but $E_{C08}(c_1) = 5548 > E_{C08}(c_2) = 5308$.

[3]Note that the case where $E_{D85}(c_1) = E_{D85}(c_2)$ but $E(c_1) \neq E(c_2)$ is not a contradiction. This is a convenient scenario, since the aim of using the alternative function $E$ is to enable a more fine-grained discrimination.

As expected, function D85 showed a low performance for this experiment. For all instances, the algorithm achieved the lowest number of iterations due to the poor discrimination this function provides. Function D85 exposed the second worst overall behavior. C04 reached slight improvements, but its limited performance was comparable with that of function D85 in some cases. Note that functions D85 and C04 were previously identified in Section IV-A because of their low discrimination capabilities. To some extent, this explains the poor performance presented by these approaches.

Functions K99 and B08 behaved similarly for the smallest benchmarks, but their performance curves diverged as the size of the problem was increased. The results of B08 deteriorated for the largest test cases, while the increasing performance of K99 allowed this function to compete at the top of the ranking. L06 obtained very competitive results most of the time. Finally, we can highlight the outstanding behavior that function I09 consistently showed for all the considered test cases. Our results indicate that the best performers for this experiment were I09, L06 and K99, in this order.

Functions I09, K99, B08 and C08 were all identified in Section IV-A to provide a strong discrimination. However, only K99 and I09 are among the best performers of this experiment.

That some equally discriminative functions performed better than others suggests that more important than the strength is the effectiveness of the discrimination (intensity does not imply effectiveness).

## 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 (D85) enables a very poor discrimination among potential conformations. Nevertheless, an effective evaluation scheme is an essential requirement for metaheuristics in order to guide the search process towards promising regions of the solutions space. Alternative HP energy functions have been proposed to enhance the performance of search algorithms. However, for most of these approaches there are not reported experimental results where the benefits of their usage are demonstrated.

This paper presented the results of a comparative study where seven different formulations of the HP energy function were considered. Our first experiment was concerned with the analysis of the degree of discrimination that each of these functions provides. The obtained results confirmed the poor discrimination capabilities of the conventional function, which has been the main motivation for exploring alternative approaches. All the alternative functions demonstrated to provide a more fine-grained discrimination. The most discriminative function according to our results is B08, followed by the K99, C08 and I09 approaches, in this order.

In our second experiment, we evaluated the impact of using the studied functions on the performance of a parameter-free local optimizer. The aim of using a parameter-free algorithm was to avoid influencing the behavior of the approaches through parameter settings. In general, most of the alternative functions allowed to increase the performance of the implemented algorithm. As expected, function D85 exhibited a low performance for this experiment. However, the C08 approach behaved even worse for most of the adopted test cases. On the other hand, functions I09, L06 and K99 consistently achieved very competitive results, being the best performers in this test.

From this study, it is possible to derive some general conclusions. First, intensity of discrimination does not necessarily imply effectiveness at guiding the search process. Even when functions I09, K99, B08 and C08 were all identified to provide a strong discrimination, only I09 and K99 behaved favorably. In contrast, B08 and particularly C08 presented a limited search performance. That the less discriminative approaches (D85 and C04) showed a low overall performance confirmed, however, that a tighter evaluation scheme is important to improve the behavior of search algorithms.

The fact that D85 consistently exposed a poor performance supports the relevance of exploring the use of alternative approaches. To the best of our knowledge, this research is producing the first results that have been reported in this direction. Nevertheless, this research is in progress. The preliminary results presented in this paper suggest that functions I09, L06 and K99 are very promising approaches for studies on the HP model. However, the impact of using these approaches needs to be further investigated for more sophisticated search algorithms. Also, it is important to extend this study to the three-dimensional cubic lattice, or to other lattice configurations (for example, the face-centered cubic lattice), in order to generalize our conclusions.

## References


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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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