Automated Design of Hyper-Heuristics Components Applied to the PSP Problem

Vidal D. Fontoura
Federal University of Parana
(DInf-UFPR), Curitiba - PR, Brazil
Email: vdfontoura@inf.ufpr.br

Aurora T. R. Pozo
Federal University of Parana
(DInf-UFPR), Curitiba - PR, Brazil
Email: aurora@inf.ufpr.br

Roberto Santana Hermida
University of the Basque Country
Barrio Sarriena s/n 48940 Leioa Bizkaia
Email: roberto.santana@ehu.es

***Abstract*—The Protein Structure Prediction (PSP) problem is one the modern most challenging problems from science. Hence, many heuristics strategies have been applied in order to find protein structures that minimizes its free energy. However, these strategies have difficulties on finding the optimal solutions to the longer sequences of amino-acids, due to, the complexity of the problem and the huge amount of local optima. The hyper-heuristics framework are usually useful in this kind of context since they try to combine different heuristics strengths into a single framework. Moreover, there is lack of works whose aim the automated design of hyper-heuristics components. This paper proposes the GEHyPSP which aims the generation, through grammatical evolution, of selection mechanisms and acceptance criteria for a hyper-heuristic framework applied to PSP problem.**

I. Introduction

Proteins plays a fundamental role in nature, they are involved in many important functions for the living cells. Proteins are structures, composed by amino-acids that guarantees the correct operation of wide range of biological entities. These structures are result from a process so-called protein folding, in which initially an unfolded chain of amino-acids will be transformed into its final/native structure.

The protein structure prediction (PSP) has a wide range of biotechnology and medical applications. For instance: synthesis of new proteins and folds [28], structure based synthesis of new drugs [8], refinement of theoretical models obtained by comparative modeling [20], [15], and obtaining experimental structures from incomplete nuclear magnetic resonance data [24].

The determination of the native structure of a protein is a hard task even for the modern super computers. The difficulty is due to the huge search space to test all possible configurations that a given protein can adopt. Many models to represent the proteins structures exists and can be utilized to simulated the folding process. Although there exists extremely detailed models, these representations are computational very expensive. Hence, many authors [6], [12], [17], [27], [23], [7], [11] used simplified models to represent the protein structures. A useful model for this purpose is the Hydrophobic-Polar (HP) model presented by Lau and Dill [16]. This model generalizes the amino-acids that composes the proteins into just two types H or P. And, a grid (2D) or cube (3D) can be used to represent the conformations that a protein can adopt. Each conformation has an energy associated [27].

Many heuristics strategies were developed to find conformations of minimum energy, using the HP model. These approaches use genetic algorithms [27], ant colony optimization [25], [26], estimation distribution algorithms [**?**], among others[9]

Although, there are different strategies already proposed for the PSP, all of them face difficulties to reach the optimal configurations when the length of amino-acids sequences increases. This motivate the study here presented. The goal is to use a pool of heuristics proposed in the literature managed by a hyper-heuristic strategy [3]. Hence, different heuristics have different strengths and weakness. It makes sense to merge them into one framework and let a high-level strategy to manage the application of them.

Burke et al. [5] recently defined hyper-heuristics as ”an automated methodology for selecting or generating heuristics to solve hard computational search problems”. Over the years, these methodologies have demonstrated success in solving a wide range of real world problems. However, there are a lack of works that use this approach to solve the PSP problem.

The paper describes the GEHyPSP: an automated mechanism for generation of high level heuristics to a hyper-heuristic framework to solve the PSP problem. The high level heuristic includes two main components: a selection mechanisms and acceptance criteria. Both components are generated using grammatical evolution (GE), which is a kind of genetic programming (GP). A set of experiment was designed to evaluate the approach using eleven PSP instances. The results are compared with the previous works of the literature.

This paper is organized as follows: in the next section we briefly introduce the Protein Structure Prediction problem and a review of the related works is also presented. Section III presents a background of hyper-heuristics, genetic programming and the grammatical evolution algorithm. Thereafter, in Section IV the GEHyPSP is presented. In Section V, the experimental benchmark and numerical results of the conducted experiments are presented. Finally, in Section VI, the conclusions of the research are given, and further work is discussed.

II. Protein Structure Prediction

Proteins are macromolecules composed by an alphabet of twenty different amino acids, also referred to as residues. An

amino acid is formed by a peptide backbone and a distinctive side chain group. The peptide bond is defined by an amino group and a carboxyl group connected to an alpha carbon to which a hydrogen and side chain group are attached.

Under specific conditions, the protein sequence folds into a unique native 3-D structure. Each possible protein fold has an associated energy value. The *thermodynamic hypothesis* states that native structure of a protein is the one for which the free energy achieves the global minimum. Based on this hypothesis, many methods [6], [12], [13], [17], [27] that search for the protein native structure define an approximation of the protein energy and use optimization algorithms that look for the protein fold that minimizes the energy. These approaches mainly differ in the type of energy approximation employed and in the characteristics of the protein modeling.

A. The HP Model

The protein structures are very complex. Detailed representations of proteins exists and can be used to model the protein folding, these representations are computationally very costly. Having this in mind, Lau and Dill [16] created a model called *Hydrophobic-Hydrophilic* Model (HP Model), to represent the proteins using simplifications. The model can be used either to represent proteins in a 2D space or 3D space.

The HP model considers two types of residues: hydrophobic (H) residues and hydrophilic or polar (P) residues. A protein is considered a sequence of these two types of residues, which are located in regular lattice models forming self-avoided paths. Given a pair of residues, they are considered neighbors if they are adjacent either in the chain (connected neighbors) or in the lattice but not connected in the chain (topological neighbors).

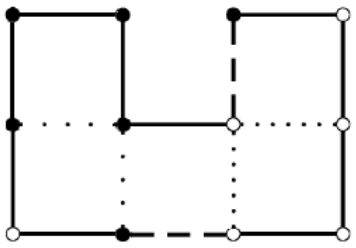

Fig. 1. One possible configuration of sequence $PPPPHPHHHHPH$ in the HP model. There is one HH (represented by a dotted line with wide spaces), one HP (represented by a dashed line) and two PP (represented by dotted lines) contacts.

For the HP model, an energy function that measures the interaction between topological neighbor residues is defined as $\epsilon_{HH} = -1$ and $\epsilon_{HP} = \epsilon_{PP} = 0$. The HP problem consists of finding the solution that minimizes the total energy. In the linear representation of the sequence, hydrophobic residues are represented with the letter H and polar ones, with P. In the graphical representation, hydrophobic proteins are represented by black beads and polar proteins, by white beads. Figure 1 shows the graphical representation of a possible configuration for the sequence $PPPPHPHHHHPH$ in a 2D space. The energy that the HP model associates with this configuration is -2 TODO: fixme because there is only two HH contact, arisen between the second and fifth residues.

Among many works that explores the PSP problem, here are some examples of the approaches that have been used to solve it.

A very influential paper from Unger and Moult [27], introduces a evolutionary algorithm (EA) which uses heuristic-based crossover and mutation operators for the HP model. The algorithm outperformed many variants of Monte Carlo methods for different instances. However, good results were obtained, the EA was unable to find the global optimal for the longest instances considered.

A multimeme algorithm (MMA) presented in [13] is a EA combined with a group of local search methods. For each individual in the population the MMA, selects a local search method that best fits. Used first to find solutions for the functional model protein. The strategy was later improved with fuzzy-logic-based local searches, leading the algorithm to achieve improved results in the PSP problem.

In [12], the author uses pruned-enriched Rosenbluth method (PERM) also know as Chain growth algorithm, that is based on growing the sequence conformation by adding particle by particle, aiming to increase good configurations and eliminating bad ones.

The ant colony optimization (ACO) was applied to the PSP problem using the HP-2D model in [25], [26]. This strategy, uses artificial ants to build conformations for a given HP instance. A local search method is then applied to further improve the solutions and also maintain the quality of the solutions.

The work of [23] utilizes Estimation of distribution algorithms (EDAs) as an efficient evolutionary algorithm that learns and exploits the search space in the form of probabilistic dependencies. New ideas were introduced for the application of EDAs to 2D and 3D simplified protein folding problem. The obtained results showed that EDAs can achieve superior solutions compared with other well-known population based optimization algorithms.

The present paper proposes the use of a grammatical evolution GE to generate high level heuristics for a hyper heuristic framework that will be applied to the PSP problem.

III. Background

This section will present the background context in order to supply the readers with the necessary information, for a good comprehension of the concepts and techniques used in this paper.

A. Hyper-Heuristics

Burke et al. [5] recently defined hyper-heuristics as "an automated methodology for selecting or generating heuristics to solve hard computational search problems". Mainly, a generic hyper-heuristic framework is composed of two main components known as high-level and low-level heuristics.

Figure III-A shows a general scheme of hyper-heuristic frameworks. The high and low level are separated by a domain barrier which means that the low-level component is problem dependent while the high-level component does not require any knowledge of the problem domain. Ideally, to apply a hyper-heuristic framework to a different problem domain: it is possible to do it only by replacing the low-level component no changes should be required at the high-level. It is responsibility from the high-level component to manage the selection or generation of which heuristic should be applied at each decision point. The low-level component corresponds to a pool of heuristics or heuristic components [22].

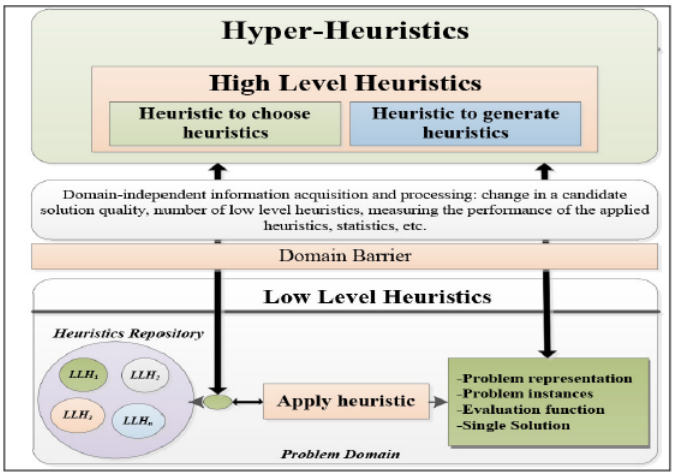

Fig. 2. General scheme of hyper-heuristic framework

Recently Burke et al. [5] classified the hyper-heuristics framework based on the nature of the search space: can be either selecting or generating heuristics for the underlying problem. Next will be analyzed both kinds of high-level heuristics.

- Selecting Heuristics: The majority of hyper-heuristic frameworks published define high level heuristics to select low level heuristics. In general, these frameworks define a selection mechanism and also an acceptance criterion. These strategies uses information from the statistical history of low level heuristics applications. For instance, a greedy selection considers only the number of improvements that a low level heuristic obtained [3]. In contrast, the choice function [3] is a reward-based strategy, which considers both the number of improvements and the elapsed time since the last application.
- Generating Heuristics: In this case, the hyper-heuristic framework starts with a set subcomponents from low level heuristics and have the goal of building new low level heuristics with it. Genetic programming is reported as a good strategy to combine and generated new heuristics for the SAT, scheduling and bin-packing problems [22].

A study presented by Sabar et al. [22], proposes an on-line approach to generate selection mechanisms and acceptance criteria (high level heuristics) to a hyper-heuristic framework, using GEP (gene expression programming a kind of genetic programming). The experiments presented by Sabar et al. [22], using the 6 problem domains provided by the HyFlex hyper-heuristic framework [19], shown a impressive results comparing with other human-designed hyper-heuristic strategies from the state-of-art. This inspired the present study which has the objective of off-line generate selection mechanisms and acceptance criteria, through a grammatical evolution process, for a hyper-heuristic framework to solve PSP problem.

B. Genetic Programming (GP)

GP is a sub-field from the program synthesis which uses ideas from the evolution theory to produce programs [3].

Grammatical Evolution (GE) is a relatively new technique from the evolutionary computing, introduced by Ryan el al. [21] and it is a type of GP where programs are evolved using a genetic algorithm. The chromossome encode production rules of the grammar. The GE uses a mapping mechanism between the genotype (coded individuals by integer vectors) and phenotype (generated programs to resolve a problem). The BNF notation is used to represent a grammar from a language in form of production rules. A BNF grammar consists in a set of terminals, which are items that are allowed to the language, for instance: +, -, *, / etc and non-terminals, which can be expanded into one or more terminals. The Grammar can be expressed as a tuple N, T, P, where N is a set of non-terminals presented in the Grammar 1 as $\langle expr \rangle$, $\langle op \rangle$, $\langle pre-op \rangle$ and $\langle var \rangle$, T a set of terminals and is presented in the Grammar 1 as $+$, $-$, $/$, $*$, Sin, Cos and Tan. Finally, P is the set of production rules that maps the elements N to T and is presented in Table I.

$\langle expr \rangle$	::=	$\langle expr \rangle\ \langle op \rangle\ \langle expr \rangle$	(0)
	|	$(\langle expr \rangle\ \langle op \rangle\ \langle expr \rangle)$	(1)
	|	$\langle pre\text{-}op \rangle\ (\langle expr \rangle)$	(2)
	|	$\langle var \rangle$	(3)
$\langle op \rangle$	::=	+	(0)
	|	-	(1)
	|	/	(2)
	|	*	(3)
$\langle pre\text{-}op \rangle$	::=	Sin	(0)
	|	Cos	(1)
	|	Tan	(2)
$\langle var \rangle$	::=	X	(0)

Grammar 1. Sample grammar to demonstrate how to decode integer vectors in computer programs

Ryan et al. [21] proposed the use of genetic algorithm (GA) to control which choices should be made, in this sense allowing the GA to select which production rules should be utilized. An individual (chromosome) consists in variable-length integer vector and it is called genotype. Algorithm 1

TABLE I
PRODUCTION RULES AND THE NUMBER OF CHOICES ALLOWED

Production Rules	Number of choices
$\langle expr \rangle$	4
$\langle op \rangle$	4
$\langle pre-op \rangle$	3
$\langle var \rangle$	1

presents the pseudo code of a grammatical evolution program. Lines 3 to 6 represents the initialization of the algorithm: a randomly generate population is created and then mapped to programs using the GF (Grammar File) those programs are evaluated and a fitness value is assigned for each individual in the population. Line 7 represent the main loop of the algorithm and the evolution process occur within this loop. Line 8,9,10,11 and 12 occurs: the parent selection; the crossover between the parents, the application of the prune and duplication operators and finally, the mutation operator is applied. Line 13 represents the mapping between the offspring (integer vectors) to programs using the GF. Line 14 and 15 represents the fitness assignment to the individuals from the population. Line 16 represents the population replacement with the fittest individuals. Finally, line 18 return the program with the best fitness found.

Algorithm 1: Pseudo code from the Grammatical Evolution

```
1  Input: GF – Grammar File
2  begin
3      population ← Create population;
4      programas ← Maps the population to programs using GF;
5      Execute the program;
6      Assign fitness value to the solutions of population according with
       the output obtained by the respective decoded program;
7      while Stop condition not reached do
8          parents ← Select individuals to crossover;
9          offspring ← Crossover(pais);
10         Apply the Prune operator to the offspring;
11         Apply the Duplicate operator to the offspring;
12         Apply the Mutation operator to the offspring;
13         programs ← Maps offspring to programs using GF;
14         Execute programs;
15         Assign fitness value to solutions offspring according with
           the output obtained by the respective decoded program;
16         population ← Replacement;
17     end
18     return Best program from the population;
19 end
```

For the sake of comprehension the chromosome mapping process will be demonstrated using the Grammar 1. The Algorithm 2 presents the general template of the generated programs. The expression $\langle expr \rangle$ presented on line 2 is replaced by mathematical expressions coded by the chromosomes (integer vectors).

Now suppose the following integer vector:

$$[220, 203, 17, 6, 108, 215, 104, 30]$$

Algorithm 2: General template for the generated algorithms

```
1 float symb(float x)
2 a = ⟨expr⟩;
3 return a;
```

This vector will be utilized to decode the chromosome (genotype) into a piece of code (phenotype) using the Grammar 1. Table I summarizes the number of choices associated within each production rule from Grammar 1. There are 4 options of production rules that can be selected for the expression $\langle expr \rangle$. In order to select which option, the first value from the vector should be used. The value is 220 and its mod divided by 4 (four options that can be selected for the expression $\langle expr \rangle$) results in 0, which means that the first option should be selected $\langle expr \rangle \langle op \rangle \langle expr \rangle$. Note that the first expression is also $\langle expr \rangle$ and following the same logic we should expand using the next value from the integer vector and applies its mod by the amount of options. The mod from the next value from the vector is 203 mod 4 = 3, which indicates that we should select the fourth option: $\langle var \rangle$ which is a terminal. The $\langle var \rangle$ has just one option associated with it and is the value X. Placing this selection in the original expression we have $X \langle op \rangle \langle expr \rangle$. Next it is necessary to decode the non-terminal expression $\langle op \rangle$. The next value from the integer vector is 17 and again we have 4 options $(+|-|/|*)$. The result is equal to 1, and indicates to select the: $-$. Re-writing the expression we got: $X - \langle expr \rangle$. This process should continue until all the non-terminals be expanded to terminals. In this example the resulting expression (phenotype) is: $X - Sin(X)$. Note that not all of the genes were necessary to obtain the phenotype. In this cases, the genes that were not used are discarded. Moreover, the opposite case can occur: if a chromosome does not contain the necessary number of genes to map to a program. In this scenario the strategy is re-utilize the genes starting from the first one.

IV. GEHYPSP

In this section will be presented the GEHyPSP: an off-line grammatical evolution application for generating high level heuristics of hyper-heuristics framework to the Protein Structure Prediction problem. The approach presented next is based on Sabar et al. [22] study.

The high level heuristics are composed by a selection mechanism and an acceptance criterion. These high level heuristics utilizes information related with the history of low level heuristics applications. Data about the improvements obtained by the low level heuristics; number of times since the last application of a low level heuristic; fitness difference between the current solution and the generated one are examples of information used by selection mechanisms and acceptance criteria. Furthermore, two terminal sets (one for the selection mechanisms and another for the acceptance criteria) were defined accordingly with the information that can be

extracted during the search progress. The selection terminals are presented next:

- RC (*Reward Credit*): The reward that a given heuristic should receive based on its performance. The improvement is calculated, for the i_{th} heuristic, using $M(i) = (|f1 - f2|/f1) * 100$ if $f2 < f1$, where $f1$ is the current fitness and $f2$ is the fitness of the generated solution by the i_{th} heuristic.
- C_{best}: Number of times that the i_{th} heuristic updated the best known solution. This terminal is useful to systematically improve the current local minimum.
- $C_{current}$: Number of times that the i_{th} heuristic updated the current solution. This terminal is useful to keep the search near from the current solution.
- C_{accept}: Number of times that the generated solution by the i_{th} heuristic was accepted by the acceptance criterion. This terminal flavors heuristics that can escape from local minimum.
- C_{ava}: The average of previous improvements made by the i_{th} heuristic during the search progress. This terminal flavors heuristics that made big improvements in average.
- C_r: Number of times that the i_{th} heuristic was classified as the first.

A specific terminal set for generating acceptance criteria was also defined.

- Delta: The difference between the quality of the current solution and the generated solution.
- PF: The quality of the previous solution.
- CF: The quality of the current solution.
- CI: Current iteration.
- TI: Total number of iterations.

Using these statistic data as terminals and a function set containing the following arithmetic operations: addition, subtraction, multiplication and division, a grammar was designed to support the generation of the high level heuristics. The designed grammar to generate selection mechanisms and acceptance criteria is presented in the Grammar 2.

For the initialization of the terminal set data: all heuristic were executed once and the data was calculated for each terminal. Every consecutive iteration will update the terminal set data and this information will be used during the search progress.

A. Fitness Function

In order to evaluate the generated individuals during the search progress, a fitness function was designed. The fitness function consisted on running the generated high level heuristics, within a hyper-heuristic framework, against three random selected instances from a set of eleven instances. Each run will be executed for one minute and will return the best HP solution found. The fitness value associated with the returned solution is then normalized between 0 and 1. The fitness of an individual, of the GE, it is the sum of the three outcomes from each execution of the three randomly selected instances. Hence, the best possible fitness value is 3 and the worst is 0. The motivation behind executing the high level heuristic (individual) against three HP instances is that executing with only one it might not be sufficient to train a high level heuristic to obtain good results in various instances of the HP model.

⟨*hh-selection*⟩ ::= ⟨*selection-mechanism*⟩ ⟨*acceptance-criterion*⟩

⟨*selection-mechanism*⟩ :== ⟨*selection-terminal*⟩
| ⟨*selection-mechanism*⟩ ⟨*math-function*⟩ ⟨*selection-mechanism*⟩
| (⟨*selection-mechanism*⟩ ⟨*math-function*⟩ ⟨*selection-mechanism*⟩)

⟨*selection-terminal*⟩ :== RC | Cbest | Ccurrent | Caccept | Cava | Cr

⟨*math-function*⟩ :== + | - | * | %

⟨*acceptance-criterion*⟩ ::== ⟨*acceptance-terminal*⟩
| ⟨*acceptance-criterion*⟩ ⟨*math-function*⟩ ⟨*acceptance-criterion*⟩
| (⟨*acceptance-criterion*⟩ ⟨*math-function*⟩ ⟨*acceptance-criterion*⟩)

⟨*acceptance-terminal*⟩ :== PF | CF | CI | TI

Grammar 2. Designed grammar to generate high level heuristics

B. Stopping Criterion

To stop the GE process a maximum number of evaluations was setup to 60000. This value was defined based on the work [21] where the grammatical evolution general process is first introduced.

C. Low Level Heuristics

1) Representation of the problem: There are different types of representing the protein structures within the HP-2d model. According to Krasnogor et al. [14] the relative representation have a better potential to achieve superior results. The relative representation stats that: the movements in the grid for each amino-acid are represented always based on the previous, that is why it is named relative. There are 3 possible movements within the HP-2d model: forward (F), left (L) and right (R). Hence, the following integer codification was used $F \rightarrow 0$, $E \rightarrow 1$ and $D \rightarrow 2$. Thereby, the allowed alphabet can be represented as $\{0, 1, 2\}$.

2) Low Level Heuristics Set: The low level heuristics set of was selected from previous studies [1], [7], [6], [10] that explore the PSP.

- *Two Points Crossover* (2X): This operator selects, randomly, two crossing points splitting the individuals in 3 pieces. The genes between the selected positions are exchanged between the parents in order to generate the offspring. [1].
- *Multi Points Crossover* (MPX): Similar with 2X although with c points of crossing. The c is calculated using $c =$

$int(n * 0.1)$, where n is the sequence length. The MPX is useful to promote a wide structural diversity [22].

- *Segment Mutation* (SMUT): It changes a random number (5 to 7) of consecutive genes to distinct directions. This heuristic introduces large changes in the conformation, and it has a great probability of creating collisions.
- *Exhaustive Search Mutation* (EMUT): This heuristic selects a random gene and changes to all possible directions and it will keep the change that achieve improvement. This operator has a trade-off because it demands four fitness evaluations instead of just one. However, this heuristic has great potential of improving the input solution.
- *Local Move Operator* (LM): This heuristic exchanges directions between two consecutive random selected genes. There are some conditions in order to execute this heuristic, for instance, the new directions can not create redundant movements.
- *Loop Move Operator* (LPM): Similar with LM, this heuristic exchanges directions between two genes that are five genes of distance between each other, creating a loop movement.
- *Opposite Mutation* (OM): This heuristic exchanges directions, for the opposite, within a sequence of genes between two genes (i, j) randomly selected.

3) Backtrack Repair: Since the low level heuristics have a great probability of generating solutions with collisions [1]. Whenever it is possible, to repair a solution using a backtrack strategy the solution is maintained otherwise it is penalized.

4) Memory Mechanism: Sabar et al. [22] suggested that the use of a memory of solutions to the problem would be more effective than relying on a single solution and may restrict the ability of dealing with large and heavily constrained search spaces, as it is widely known that single solution based methods are not well suited to cope with large search spaces and heavily constrained problems [2].

V. Experiments

In this section, we will discuss the conducted experiments in order to evaluate the GEHyPSP proposed by this paper. Three groups of experiments were designed/executed and will be described next. The first group was only concerned on generating selection mechanisms. The second group was designed only to generate acceptance criteria. The third group was developed to generate both selection and acceptance mechanisms. All experiments were executed 30 times because of the stochastic behavior of the GEHyPSP. The results obtained by the groups were compared each other. Also a comparison with a state-of-art hyper-heuristic will be presented. The results were compared with the Generic Intelligent Hyper-heuristic (GIHH) presented by Misir el al [18].

The first group of experiments GEHyPSP-1 was executed consisting on only generating selection mechanisms. The acceptance criterion was fixed with a "better or equal" acceptance [4]. The goal of this group is: evaluate the ability to generate only selection mechanisms using a fixed acceptance criterion.

TABLE II
Instances and the respective size of each one

Inst	1	2	3	4	5	6	7	8	9	10	11
Size	20	24	25	36	48	50	60	64	85	100	100

TABLE III
Results from the best individual found in GEHyPSP-1

Inst	1	2	3	4	5	6	7	8	9	10	11
Avg	8.1	7.6	6.7	11.9	17.4	16	30	28.3	40.1	35.6	35.9
St Dv	0.3	0.5	0.5	0.7	1	1.4	1.7	2	2.7	2.1	2.9
Min	8	7	5	11	15	13	25	23	34	32	27
Max	9	9	7	13	19	20	33	32	46	40	41
O(x*)	9	9	8	14	23	21	36	42	53	48	50

The second group of experiments, GEHyPSP-2, consisted on generating only acceptance criteria. The selection mechanism was fixed using the best selection mechanism found in the first group. Consequently, this group of experiments depends on the output from the first group. The goal of this experiment was to evaluate the generation of acceptance criteria separately from the selection mechanism using a fixed one.

The third group of experiments GEHyPSP-3 was designed to generate both selection mechanisms and acceptance criteria. The goal of this group was evaluate the ability of generating selecting mechanisms along with acceptance criteria. Differently from GEHyPSP-2, this group of experiment does not depend on of any output from previous experiments since it generate both mechanisms without fixing any piece.

For each group of experiments it was utilized eleven instances (amino-acid sequences from the HP model) selected from the previous studies with the PSP problem [6], [12], [17], [27], [23], [7], [11]. The instances used in the experiments are presented in table II. For the sake of space, the instances formula were suppressed please refer to [23] for the sequences formula. Also, for each group of experiments the training phase consisted on executing the GE process with tree ramdonly selected instances from the eleven available instances. In the validation phase the best generated selection mechanisms and acceptance criteria were executed together against all eleven instances.

A. Results from GEHyPSP-1

The best individual found in the GEHyPSP-1 was the following selection mechanism: $RC * Ccurrent * Cava - Cr$ and it was executed 30 times with a time limit of 10 minutes against the eleven instances. Table III presents the average, standard deviation, minimum and maximum of 30 executions of the best individual found in the GEHyPSP-1 experiment. The last row denoted with O(x*) is the best known value for each sequence. It is possible to note analyzing the last two lines that only for the smaller instances the generated selection mechanism achieved good results. For the sequences 8,9,10 and 11 the results obtained are very far from the best known results.

TABLE IV
RESULTS FROM THE BEST INDIVIDUAL FOUND IN GEHYPSP-2

Inst	1	2	3	4	5	6	7	8	9	10	11
Avg	8	7.6	6.7	10.1	14.8	14.7	27	25.7	38.3	32.8	30.
St Dv	0.6	0.4	0.6	0.7	1.5	1.4	2.0	2.5	3.3	3.7	3.4
Min	7	7	5	8	12	12	23	22	31	26	24
Max	9	8	8	11	17	18	31	31	44	40	37
O(x*)	9	9	8	14	23	21	36	42	53	48	50

TABLE V
RESULTS FROM THE BEST INDIVIDUAL FOUND IN GEHYPSP-3

Inst	1	2	3	4	5	6	7	8	9	10	11
Avg	7.6	7.0	5.7	9.7	13.8	12.7	24	24.2	31.6	27	26.4
St Dv	0.6	0.7	0.8	0.9	1.3	0.9	1.1	1.6	1.7	1.7	2
Min	7	6	4	7	12	11	21	21	29	24	24
Max	9	8	8	11	18	15	26	28	37	31	31
O(x*)	9	9	8	14	23	21	36	42	53	48	50

B. Results from GEHyPSP-2

The best individual found in the GEHyPSP-2 was the following acceptance criterion: $((TI/Delta)/((Delta * ((TI/Delta)/CI) * Delta/Delta * TI) - CI))$ and combined with the best individual from the GEHyPSP-1 the hyper-heuristic framework was executed 30 times for each one of the eleven instances with a time limit of 10 minutes. Table IV presents the results for the best generated acceptance criterion in GEHyPSP-2 and also using the best selection mechanism generated in GEHyPSP-1. Again, looking to the two last lines it is possible to notice that the hyper-heuristic framework, using the best generated selection mechanisms and acceptance criterion, was not able to reach the best known results for the larger sequences. And comparing Tables III and IV it is possible to visualize that there is a slight difference between the results, favoring the generated selection mechanism using the "better or equal" acceptance criterion.

C. Results from GEHyPSP-3

The best individual generated by the experiment GEHyPSP-3 was both a selection mechanism and an acceptance criterion. And they are presented below:
Selection Mechanism: $(((((Caccept/RC) * Cr/Caccept)/RC * Cr)/Caccept/RC) * Cr)/Caccept$
Acceptance Criterion: $(((((CI/PF) * Delta/CI)/PF * Delta)/CI/PF) * Delta)/CI$

Table V presents the results, of the best individual generated by the experiment GEHyPSP-3, of 30 executions with a time limit of 10 minutes. Once again the values presented did not achieve good results when analyzing the bigger instances. Furthermore, when comparing Table V with Tables III and IV it is possible to see that generating both selection mechanisms and acceptance criteria it is worst than generating them separately.

D. Comparison with a state-of-art hyper-heuristic

In order to compare the generated high level heuristics with a already proposed hyper-heuristic strategy it was selected a state-of-art human-designed hyper-heuristic framework [18] which presented the best results applied to the 6 domain problems from HyFlex framework [19]. Misir et al. [18] provided us the source code from the GIHH and then it was executed against the PSP problem.

TABLE VI
BEST RESULTS FOUND BY GEHYPSP, BEST RESULTS FOUND BY GIHH AND THE BEST KNOW RESULTS O(X*)

Inst	1	2	3	4	5	6	7	8	9	10	11
GEHyPSP	9	9	7	13	19	20	33	32	46	40	41
GIHH (max)	9	9	8	13	22	21	35	37	49	43	45
O(x*)	9	9	8	14	23	21	36	42	53	48	50

Table VI presents the best results, for each instance, found by the best generated high level heuristic with GEHyPSP, the best results found by the GIHH [18] and finally the best know results (O_x(*)). The GEHyPSP and the GIHH were executed 10 minutes. It is possible to note that the GEHyPSP found worst results than the GIHH as the sequences length increase. However, the GIHH is a human-designed hyper-heuristic which demanded several years of research in order to be completed and achieve good results in the 6 domains problems provided by HyFlex. Although, the good performance obtained by the GIHH in the domains from HyFlex, it is possible to notice that the GIHH was also unable to achieve the best known results in many instances.

E. Discussion

In order to better investigate the relationship between the generated selection mechanisms and acceptance criteria another experiment was designed. From the 30 executions of the GEHyPSP-2 ten random individuals were selected to be further analyzed. These individuals were re-executed in debugging mode in order to check its behavior. Note that only the acceptance criterion was different between them because the selection mechanism was fixed, using the best generated in the previous experiment GEHyPSP-1. From 10 individuals 7 were accepting only better or equal solutions just like the fixed acceptance criterion that was used in the GEHyPSP-1. The difference between the individuals and a fixed acceptance criterion was that individuals were slower than the fixed, because it is required to execute arithmetical functions and in the other hand only a simple if was evaluated. But thinking in the behavior they were exact the same. It was also noticed that 2 individuals, the worst of the group, were always accepting worst solutions just like the "all moves" described by Burke et al. [3]. Finally, 1 individual was never accepting any solution.

These experiments showed that the GE managed, several times, to find different acceptance criteria with the same behavior of human-designed strategies.

VI. CONCLUSION

In this work the GEHyPSP, an automatic way of generating high level heuristics to a hyper-heuristic framework to the PSP problem, was presented and evaluated. The PSP is very challenging problem with lots of local optima and a very complex landscape. Many authors explored the PSP problem with heuristic methods. However, the majority have difficult to achieve the best known results when executing against

longer sequences. Usually, the hyper-heuristic frameworks fits well in this kind of complex scenario. Hence, the goal of this paper was to generate, using a grammatical evolution strategy, selection mechanisms and acceptance criteria to a hyper-heuristic framework and evaluate its performance and behavior with a set of eleven HP instances. Three groups of experiments were executed, using three randomly selected HP instances, in order to generate the high level heuristics and later the best individuals found in the experiments were executed using all the eleven HP instances.

Three groups of experiments were executed: first generating only selection mechanisms within a fixed acceptance criterion; second generating only acceptance criteria using the best selection mechanism found in the first; finally both high level heuristics were generated in parallel. The results showed that better high level heuristics were found when generating them separately. Unfortunately when analyzing the behavior of the generated high level heuristics against the eleven instances it was possible to see that they were not able to achieve the best known results for the longer sequences. However, when comparing with a good state-of-art human-designed hyper-heuristic framework (GIHH) [18] the results are slightly near. This fact shows that it is possible to automate the creation of high level heuristics and obtain results close to the state-of-art hyper-heuristics frameworks.

Another finding of this work was the behavior of the best generated acceptance criteria. It was noticed that it behave just like "a better or equal" human-designed move acceptance strategy. Also some of the generated acceptance criterion were always accepting worst solution and this fact impacted in the individual fitness. This demonstrates that the GEHyPSP was able to generate acceptance criteria with the same behavior of simple human-designed move acceptance strategies. However, in order to get better results: it might be necessary to improve the GEHyPSP to generate more complex selection mechanisms and acceptance criteria to couple with the landscape complexity of the PSP problem.

REFERENCES

[1] César Manuel Vargas Benítez. Um algoritmo genético paralelo para o problema de dobramento de proteínas utilizando o modelo 3dhp com cadeia lateral. 2010.

[2] Christian Blum, Jakob Puchinger, Günther R Raidl, and Andrea Roli. Hybrid metaheuristics in combinatorial optimization: A survey. *Applied Soft Computing*, 11(6):4135–4151, 2011.

[3] Edmund K Burke, Michel Gendreau, Matthew Hyde, Graham Kendall, Gabriela Ochoa, Ender Özcan, and Rong Qu. Hyper-heuristics: A survey of the state of the art. *Journal of the Operational Research Society*, 64(12):1695–1724, 2013.

[4] Edmund K Burke, Mathew R Hyde, Graham Kendall, Gabriela Ochoa, Ender Ozcan, and John R Woodward. Exploring hyper-heuristic methodologies with genetic programming. In *Computational intelligence*, pages 177–201. Springer, 2009.

[5] Edmund K Burke, Matthew Hyde, Graham Kendall, Gabriela Ochoa, Ender Özcan, and John R Woodward. A classification of hyper-heuristic approaches. In *Handbook of metaheuristics*, pages 449–468. Springer, 2010.

[6] Fábio L Custódio, Hélio JC Barbosa, and Laurent E Dardenne. Investigation of the three-dimensional lattice HP protein folding model using a genetic algorithm. *Genetics and Molecular Biology*, 27(4):611–615, 2004.

[7] Fábio Lima Custódio, Helio JC Barbosa, and Laurent Emmanuel Dardenne. A multiple minima genetic algorithm for protein structure prediction. *Applied Soft Computing*, 15:88–99, 2014.

[8] Ian W Davis and David Baker. Rosettaligand docking with full ligand and receptor flexibility. *Journal of molecular biology*, 385(2):381–392, 2009.

[9] Paulo HR Gabriel, Vinícius V de Melo, and Alexandre CB Delbem. Algoritmos evolutivos e modelo hp para predição de estruturas de proteínas. *Revista de Controle e Automação*, 23(1):25–37, 2012.

[10] Mario Garza-Fabre, Eduardo Rodriguez-Tello, and Gregorio Toscano-Pulido. Multiobjectivizing the hp model for protein structure prediction. In *European Conference on Evolutionary Computation in Combinatorial Optimization*, pages 182–193. Springer, 2012.

[11] Mario Garza-Fabre, Gregorio Toscano-Pulido, and Eduardo Rodriguez-Tello. Locality-based multiobjectivization for the hp model of protein structure prediction. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation*, pages 473–480. ACM, 2012.

[12] Hsiao-Ping Hsu, Vishal Mehra, Walter Nadler, and Peter Grassberger. Growth algorithms for lattice heteropolymers at low temperatures. *The Journal of chemical physics*, 118(1):444–451, 2003.

[13] Natalio Krasnogor, BP Blackburne, Edmund K Burke, and Jonathan D Hirst. Multimeme algorithms for protein structure prediction. In *Parallel Problem Solving from NaturePPSN VII*, pages 769–778. Springer, 2002.

[14] Natalio Krasnogor, W Hart, Jim Smith, and D Pelta. Protein structure prediction with evolutionary algorithms. 1999.

[15] Elmar Krieger, Keehyoung Joo, Jinwoo Lee, Jooyoung Lee, Srivatsan Raman, James Thompson, Mike Tyka, David Baker, and Kevin Karplus. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: four approaches that performed well in casp8. *Proteins: Structure, Function, and Bioinformatics*, 77(S9):114–122, 2009.

[16] Kit Fun Lau and Ken A Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997, 1989.

[17] Cheng-Jian Lin and Shih-Chieh Su. Protein 3 d hp model folding simulation using a hybrid of genetic algorithm and particle swarm optimization. *International Journal of Fuzzy Systems*, 13(2):140–147, 2011.

[18] Mustafa Misir. Intelligent hyper-heuristics: a tool for solving generic optimisation problems. 2012.

[19] Gabriela Ochoa, Matthew Hyde, Tim Curtois, Jose A Vazquez-Rodriguez, James Walker, Michel Gendreau, Graham Kendall, Barry McCollum, Andrew J Parkes, Sanja Petrovic, et al. Hyflex: A benchmark framework for cross-domain heuristic search. In *Evolutionary computation in combinatorial optimization*, pages 136–147. Springer, 2012.

[20] Bin Qian, Angel R Ortiz, and David Baker. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(43):15346–15351, 2004.

[21] Conor Ryan, JJ Collins, and Michael O Neill. Grammatical evolution: Evolving programs for an arbitrary language. In *Genetic Programming*, pages 83–96. Springer, 1998.

[22] Nasser R Sabar, Masri Ayob, Graham Kendall, and Rong Qu. Automatic design of a hyper-heuristic framework with gene expression programming for combinatorial optimization problems. *IEEE Transactions on Evolutionary Computation*, 19(3):309–325, 2015.

[23] Roberto Santana, Pedro Larrañaga, Jose Lozano, et al. Protein folding in simplified models with estimation of distribution algorithms. *Evolutionary Computation, IEEE Transactions on*, 12(4):418–438, 2008.

[24] Yang Shen, Robert Vernon, David Baker, and Ad Bax. De novo protein structure generation from incomplete chemical shift assignments. *Journal of biomolecular NMR*, 43(2):63–78, 2009.

[25] Alena Shmygelska, Rosalia Aguirre-Hernandez, and Holger H Hoos. An ant colony optimization algorithm for the 2d hp protein folding problem. In *Ant Algorithms*, pages 40–52. Springer, 2002.

[26] Alena Shmygelska and Holger H Hoos. An improved ant colony optimisation algorithm for the 2d hp protein folding problem. In *Advances in Artificial Intelligence*, pages 400–417. Springer, 2003.

[27] Ron Unger and John Moult. Genetic algorithms for protein folding simulations. *Journal of molecular biology*, 231(1):75–81, 1993.

[28] Ling Wang, Eric A Althoff, Jill Bolduc, Lin Jiang, James Moody, Jonathan K Lassila, Lars Giger, Donald Hilvert, Barry Stoddard, and David Baker. Structural analyses of covalent enzyme–substrate analog complexes reveal strengths and limitations of de novo enzyme design. *Journal of molecular biology*, 415(3):615–625, 2012.