Protein Folding in the HP-Model Solved With a Hybrid Population Based ACO Algorithm

Torsten Thalheim, Daniel Merkle, and Martin Middendorf *

Abstract—**A hybrid population based Ant Colony Optimization (ACO) algorithm PFold-P-ACO for protein folding in the HP model is proposed in this paper. This is the first population based ACO algorithm in the bioinformatics. It is shown experimentally that the algorithms achieves on nearly all test sequences at least comparable results to other state of the art algorithms. Compared to the state of the art ACO algorithm PFold-P-ACO obtains slightly better results and is faster on long sequences.**

Keywords: Population based ACO, Ant Colony Optimization, HP model, Protein Folding

1 Introduction

Proteins are one of the most important classes of biological molecules. Proteins for example have structural functions in the muscle and the cytoskeleton, have catalytic functions and coordinate motion and signal transduction. Chemically, a protein is a chain where each element is one of 20 different amino acids. Each amino acid consists of a central carbon atom bonded to an amino group (NH_2), a carboxyl group (COOH) and a side chain or residue (R). Hence, the amino acids differ only in the residue R. One of the most important differences between the residues is their hydrophobicity, i.e., how much they are repelled from a mass of water. The properties of the residues together with the environment are responsible that the protein chain folds into a complex conformation. This conformation is called the "native" conformation of the molecule. The native conformation is thermodynamically stable, i.e., it has small Gibbs free energy, and is very important for the function of the protein.

The structure of a protein can be described on different levels: the amino acid sequence is the primary structure, the secondary structure describes characteristic structures of the backbone of the molecule within local regions (e.g., alpha-helices or beta-sheets), the tertiary structure refers to the entire 3-dimensional structure. Different types of algorithms have been developed to predict the tertiary or secondary structure of proteins. All these algorithms use a model that is an abstraction of real proteins and describes important characteristics. An important class of models are the lattice models. A lattice model consists of a lattice that describes possible positions for the amino acids and an energy function that is to be minimized and depends on the positions of the amino acids on the lattice. The most simplest lattice model is the HP model which is based on the observation that hydrophobic forces are very important factors that drive the protein folding process. Advantages of the HP model are simplicity, that it shows several aspects of real proteins, and remains the hardness features of the biological problem.

In this paper we propose a hybrid population based Ant Colony Optimization algorithm called PFold-P-ACO for solving the protein folding in the HP model. PFold-P-ACO is the first population based Ant Colony Optimization Algorithm (P-ACO) algorithm for the problem domain of bioinformatics.

The paper is organized as follows. Section 2 describes the HP model and mentions some heuristics from the literature for the protein folding problem in the HP model. An introduction to ACO and ACO approaches for the protein folding problem is given in Section 3. Population based ACO and our algorithm PFold-P-ACO are described in Section 4. The experiments and the results are presented in Section 5. Conclusions are given in Section 6.

2 The HP model

The HP model is introduced by Dill [8, 16]. It is a lattice model that is based on the fact that for folding the most important difference between amino acids is their hydrophobicity, i.e., how much they are repelled from a mass of water. The reason is that hydrophobicity is the main driving force to fold a molecule into the native conformation (at least for of small globular proteins). In the HP model all 20 different amino acids are classified into two types: hydrophobic or non-polar (H) and hydrophilic or polar (P).

*Torsten Thalheim is with the Helmholtz-Centre for Environmental Research, Permoserstrae 15, 04318 Leipzig, Germany (e-mail: torsten.thalheim@ufz.de). Daniel Merkle is with the Dept. of Mathematics & Computer Science, University of Southern Denmark Campusvej 55, DK-5230 Odense M, Denmark (e-mail: daniel@imada.sdu.dk). Martin Middendorf is with the Department of Computer Science, University of Leipzig, Postfach 100920, D-04009 Leipzig, Germany (phone: 49-341-9732275; fax: 49-341-9732252, e-mail: middendorf@informatik.uni-leipzig.de) send correspondence to M. Middendorf

A primary structure with n amino acids is viewed as a sequence $S = s_1, \ldots, s_n$ with $s_i \in \{H, P\}$ for $i = 1, \ldots, n$. A conformation is a mapping C of the amino acids s_i to the points of a cartesian lattice. Two and three dimensional cartesian lattices are used here. In the following we describe the 2-dimensional model. The definitions for the 3-dimensional model are analogous. We use the following notation: if C is a conformation then (x_i, y_i) denotes the position in the lattice to which s_i is mapped by C. All valid conformations are self-avoiding paths on the cartesian lattice. A mapping is a path when amino acids s_i, s_j that are consecutive in the molecule, i.e., $|i - j| = 1$, are mapped to neighbored positions (x_i, y_i), (x_j, y_j) on the lattice, i.e., $|x_i - x_j| + |y_i - y_j| = 1$. A path is self-avoiding when all two different amino acids s_i, s_j, $i \neq j$ are mapped to different positions, i.e., $(x_i, y_i) \neq (x_j, y_j)$.

The energy function in the HP model reflects the fact that the hydrophobic amino acids have a propensity to form a hydrophobic core. Therefore, the energy function adds a value -1 for every pair of hydrophobic amino acids (H) that are adjacent on the lattice but not consecutive in the sequence. Formally, the energy $E(C)$ of a conformation C is $\sum_{1 \leq i \leq j-2 \leq n} I(i, j)$ where $I(i, j) = -1$ if $|x_i - x_j| + |y_i - y_j| = 1$ and $I(i, j) = 0$ otherwise.

The protein folding problem in the HP model — called HP-Protein Folding problem — is to find for a given protein $S = s_1 \ldots s_n$, $s_i \in \{H, P\}$ a valid conformation C on the cartesian lattice such that the energy $E(C)$ is minimum. In [5] it was shown that the HP-Protein Folding problem is NP hard, i.e., it is very unlikely that there exists a polynomial time algorithm for solving the problem. Therefore, it is interesting to find heuristics for solving the HP-Protein Folding problem.

The variety of heuristics that have been developed for HP-Protein Folding problem that include (Metropolis) Monte Carlo algorithms, chain growth algorithms, evolutionary algorithms, memetic algorithms, immune algorithms and ACO algorithms. An exact branch-and-bound algorithm has been presented in [31]. In the following we shortly present some of the heuristics (for recent more complete overviews see [26, 33]) but the ACO algorithms are described in more detail the next section.

Unger and Moult presented a genetic algorithm [30] that incorporates a Monte Carlo method as mutation operator to change the individual conformations in the population. Each individual is encoded as a sequence of sequence of moves (left, right, ...) in the lattice. As crossover operator a one-point crossover is applied. The probability to be selected for crossover depends on the energy the conformations so that low energy structure have a higher chance. Patton et al. [21] described a standard GA that uses a penalty method to enforce the self-avoiding constraints. Liang and Wong [18] proposed a hybrid between Monte Carlo optimization and GA.

A memetic algorithm [4] that uses a self-adaptive strategy for Local Search (LS). Depending on the degree of convergence LS can act toward either exploitation or diversification. A temperature parameter that is chosen according to a Boltzmann distribution is used to determine the chances that a LS move that causes a decrease in fitness is accepted. Krasnogor et al. [15] extended the memetic algorithm by the introduction of a contact map memory of current solutions in the mating strategy.

A filter-and-fan algorithm has been proposed by Rego and Glover [23]. This algorithm switches between a local search to identify a local optimum and a filter and fan search which is as a restrictive form of tabu search to explore larger neighborhoods in order to overcome local optimality.

A chain growth method (CG) by Beutler and Dill [6] biases the construction towards finding a good hydrophobic core. The so called 'pruned-enriched Rosenbluth method' (PERM) which is also a biased chain growth algorithm has been applied by Hsu et al. [13] for the Protein Folding problem. The algorithm evaluates partial conformations and employs pruning and enrichment operators to explore partial solutions. This algorithm and its variant by Huang and Lü [14] are among the best algorithms for the protein Folding problem is.

3 Ant Colony Optimization

Ant Colony Optimization (ACO) is a metaheuristic that is inspired by the foraging behaviour of real ants ([9]). ACO has been applied successfully to solve various combinatorial optimization problems (see [10, 20]). It is an iterative method where artificial ants search for good solutions. Every ant of an iteration builds up a solution stepwise thereby going through several decisions. The ants that found a good solution mark their paths through the decision space by putting some amount of (artificial) pheromone on the edges of the path. The following ants of the next iteration are attracted by the pheromone and search in the solution space near good solutions.

Shmygelska and Hoos proposed several variants of an ACO algorithm for the HP-Protein Folding problem ([24, 25, 26]). The latest algorithm ACO-HPPFP-3 iterates over the following three phases: construction phase, local search phase, pheromone update phase. In the construction phase each ant constructs a candidate solution by sequentially growing a conformation of the given HP sequence, starting from a folding point that is chosen uniformly at random among all sequence positions. Since conformations are rotationally invariant, the position of the first two amino acids can be fixed without loss of generality. A candidate conformation for a HP sequence of length n corresponds to a decision sequence of length $n - 2$. Each decision extends the subsequence of amino acids that have already placed by placing either the pre-

ceding or the following amino acid in the chain (i.e., either the direct predecessor of the first amino acid in the subsequence or the direct successor of the last amino acid on the subsequence). The decision indicates the position of the newly placed amino acid on the 2D or 3D lattice relative to its two direct predecessors in the given sequence. Possible decisions are whether the chain folds straight (S), left (L), right (R) in 2D, (and also up (U), down (D) in 3D). Whether the partial conformation is extended to the front or to the back, is done such that the ratio of the lengths of the unfolded residues at each end of the protein remains (roughly) unchanged. The relative direction $d \in \{S, L, R\}$ in which the conformation is extended in construction step i is determined probabilistically based on a heuristic function $\eta_{i,d}$ and pheromone values $\tau_{i,d}$ according to the formula:

$$p_{i,d} = \frac{\eta_{i,d}^{\alpha} \cdot \tau_{i,d}^{\beta}}{\sum_{e \in \{S,L,R\}} \eta_{i,e}^{\alpha} \cdot \tau_{i,e}^{\beta}} \quad (1)$$

where $\alpha > 0$ and $\beta > 0$ are parameters that determine the relative influence of pheromone and heuristic information. The pheromone values $\tau_{i,d}$ indicate the desirability of using direction d at sequence position i. Initially, all $\tau_{i,d}$ values are equal. Throughout the search process, the pheromone values are updated to bias the folding towards the use of local directions that occur in low-energy structures.

In the pheromone update phase each pheromone value $\tau_{i,d}$ is evaporated according to $\tau_{i,d} := \rho \cdot \tau_{i,d}$ where $\rho < 1$ is the pheromone persistence parameter. Subsequently, selected ants with low-energy conformations update the pheromone values according to $\tau_{i,d} := \tau_{i,d} + \Delta_{i,d,c}$ where $\Delta_{i,d,C}$ is the relative solution quality of the given ant's candidate conformation C if that conformation contains local direction d at sequence position i, and zero otherwise. For further details and information on the following parts of ACO-HPPFP-3 see [26]: i) the heuristic method, ii) the backtracking method that is used when a direction is not possible because the chain would run into itself, i.e., it would invalidate the self-avoiding property, iii) the local search method that is used to improve conformations that have been found be the ants.

4 P-ACO and PFold-P-ACO

One of the main characteristics of an ACO algorithm is the pheromone information which stores information on good solutions that have been found by ants of former iterations. The pheromone information is what is transferred from one iteration of the algorithm to the next. An alternative to this scheme has been proposed by Guntsch and Middendorf [11] and is called Population based ACO (P-ACO). Instead of pheromone information as in ACO, in P-ACO a population of solutions is transferred from one iteration of the algorithm to the next. The ants in the new iteration use this population to construct pheromone information from it and then proceed as in standard ACO for solution construction by using also Formula (1). Instead of pheromone update P-ACO uses a population update and several strategies have been proposed for a solution to enter or leave the population (see [11]). Two potential advantages of P-ACO compared to standard ACO are: i) the population update and pheromone construction needs for typical applications (e.g., for permutation problems like the Traveling Salesperson problem) time $O(n)$ where n is the problem size instead of time $O(n^2)$ that is necessary for standard ACO pheromone update, ii) the population can be used to apply operations on the solutions (e.g. crossover). A potential disadvantage of P-ACO compared to ACO is that the number of different pheromone values is small (typically, the population size) whereas for standard ACO it is potentially infinite. Thus, a P-ACO algorithm might be faster than a standard ACO algorithm but it is not clear whether it can achieve the same solution quality.

It has been shown experimentally that P-ACO works equally good as ACO even when the P-ACO algorithm uses only a small population ([11]). It was also shown that P-ACO can be used for multi-objective problems [1, 12, 7]. An extended P-ACO algorithm with niching has been proposed in [2]. The P-ACO principle has also been used to develop an ACO algorithm for continuous optimization problems [27]. Since P-ACO has been tested on classical problem like TSP ([1, 3, 11, 12, 7]) or single machine scheduling problems [12] only it is interesting to apply it to other problem domains.

In the following subsections we describe our hybrid P-ACO algorithm called PFold-P-ACO. It consists of two parts: a P-ACO part and a branch-and-bound part. When the P-ACO part has not found an improvement over a certain number of iterations, the branch-and-bound part starts. The branch-and-bound part uses the pheromone information frome the P-ACO part and is a heuristic that does not do a complete enumeration.

4.1 ACO Part

The population that is used by PFold-P-ACO contains always the best 10 conformations that have been found. But to keep enough diversity a new found conformation that has the same HH-contacts as a conformation that is already in the population is not allowed to enter the population. If a new found conformation has the same energy as a conformation in the population it replaces the one in the population with probability 0.5. The Construction Phase and the Local Search Phase are described in the following subsections.

Construction Phase. Similar as in ACO-HPPFP-3 from [26] each ant constructs a solution by sequentially growing a conformation of the given sequence, starting

from an element that is chosen uniformly at random. Different from algorithm ACO-HPPFP-3 in PFold-P-ACO the probability that the next element to be placed is chosen in direction of the beginning or end of the sequence equals the relative length of the remaining prefix respectively suffix of the sequence. But if it is possible that a long subsequence of P amino acids can be extended or certain prescribed HH-contact can be realized this is always done (details are described later.

All pheromone values $\tau_{i,d}$ $i = 1, \ldots, n-1$, $d = S, L, R$ are initialized with value 1. Each ant chooses randomly a conformation from the population and sets all corresponding pheromone values to 4. As heuristic values the ants use the change of the energy ΔE of the partial conformation that occurs when a decision is made, i.e., $\lambda_{i,d} = e^{\Delta E}$. An ant that has to decide how to fold at s_{i+1} makes with probability $\sigma \geq 0$ a decision randomly with equal probability for S, L, or R (if all are possible). Otherwise, the ant considers the pheromone values and decides for direction d with probability $\tau_{i,d}^{\alpha} \cdot \lambda_{i,d}^{\beta} / (\sum_{h \in \{S,L,R\}} \tau_{i,h}^{\alpha} \cdot \lambda_{i,h}^{\beta})$ if all directions are possible where parameters α and β define the relative influence of pheromone and heuristic. If not all directions are possible only the pheromone values corresponding to the allowed directions are taken into account. If for example directions S and R are allowed but direction L is forbidden by a constraint as described in the following then the probability to fold in direction d is $\tau_{i,d} \cdot \lambda_{i,d} / (\sum_{h \in \{S,R\}} \tau_{i,h} \cdot \lambda_{i,h})$.

Several constraints are used that might forbid some decisions for an ant. If the ant can not make an alternative decision it makes a backtrack step and revises the former decision. This is done until the ant finds a decision that is not forbidden or until a Time-to-Live (TTL) counter that counts the number of backtrack steps stops the ant.

Constraint 1. Inspired by an idea from [19] a set of prescribed amino acid contacts is used to guide the search for good conformation. Different from [19] the set $\mathcal{H}$ that is used by the ants in the PFold-P-ACO contains only HH-contacts. Further, set $\mathcal{H}$ contains only local restrictions, i.e., for every HH-contact in $\mathcal{H}$ the distance of the two H elements in the sequence is at most 9. Moreover, two HH-contacts are not both included into the set $\mathcal{H}$ if they are in conflict with each other, i.e., when it is not possible that both contacts can be realized by a conformation. The following easy criterion is used to detect such a case. Let x_1, x_2 are the sequence positions of the H elements of one HH-contact and y_1, y_2 are the sequence positions of the H elements of another HH-contact. Then both HH-contacts are in conflict if $y_1 - x_2 + p > y_2 - y_1$ where $p = 2$, if $(y_1 - x_2 > 1) \wedge (y_1 - x_2 > x_2 - x_1)$ and otherwise $p = 0$. The conflict criterion is illustrated in Figure 3.

If during the construction phase an ant places the first element of a HH-contact in $\mathcal{H}$ it initializes a vector of counters —- one counter for every moving direction (up, down, left, right). The counters are used to check for every decision in the construction process whether the current element is placed near enough to the location of first element of the required HH-contact so that the HH-contact can still be realized with respect to distance (it is not checked if it is really possible to realize the HH-contact).

Constraint 2. It is required that the elements of P-rich subsequences are placed near to each other. A P-rich subsequence is defined such that before and after it comes an H element and it contains at least 75% percent P elements and does not contain a singleton P element that has no P neighbor. Similar as for the required HH-contacts a vector of counters is used for every P-rich subsequence. It is checked when an element of a P-rich subsequences is to be placed that its location would not be too far from the location of the first element of this subsequence. If, the location is too far the ant has to make an alternative decision.

The reason to introduce Constraint 2 was that it can be seen that the ants tend to create conformations several small hydrophobic cores if PFold-P-ACO runs without Constraint 2 (an example for this is given in figure 1). This is a problem because most proteins the minimum solution contain only one hydrophobic core (as mentioned, e.g., in [6]). For the ants it is difficult to find a conformation with one core because a conformation often reflects a solution with local optimization that often has an energy value that is close to the known minimum. The reason why the ants tend to create several small cores are the P-rich subsequences. Due to the fact that the P elements did not directly influence the energy calculation the pheromone intensity for all 3 directions when placing a P element are often not very different and therefore the probability for all three directions are similar. As a result when placing a P rich subsequence the ants often fold the molecule away from a hydrophobic core. Therefore, Constraint 2 forces the ants to fold P-rich subsequences within a small area.

Constraint 3. HP-contacts are not allowed.

Local Search Phase. A local search is used that is similar to the filter-and-fan approach from [23]. The move operator is called pull-move introduced in [17] and described in detail in [23]. The pull-move is initiated by moving one node of the current conformation to one of its empty diagonal adjacent positions in a square of the lattice where the positions of one side of the square are occupied be the node itself and one of its direct neighbors. Depending on the structure of the conformation the displacement of the initiating node may require other nodes to change their positions. In a pull-move, displaced nodes are only allowed to occupy vacant adjacent positions in the lattice.

Table 1: Test HP sequences D100-x of length 100 and their best known energy values E_{min}. The corresponding best folded conformations are shown in Figure 2.

ID	E_{min}	Sequence
D100-1	-24	$P_2HP_5HPHP_{12}HP_4H_3P(P_2H_2)_2(PH)_3P_4HP_{17}HP_2HP_3H_3PHP_2HP_2(PH)_2P_2H_2P_6H$
D100-2	-42	$H_2P_4H_2(PH_5)_2(PH)_3P_{11}HPH_3P(HP_2)_2H_2P_4H_2PHP_2H_2PHPH_6P_2H4P_3(H_2P)_4P_2HP_3H_4$
D100-3	-52	$P_2H_2P_3HPH_5P_2HPH_{10}PH_2P_2H(P_2H_2)_2P_3HPH_3PH_2(P_2H_3)_2H(PH_3)_2H_2P_3HP_2H(PH_3)_2HP_2H_2P_4H_2$

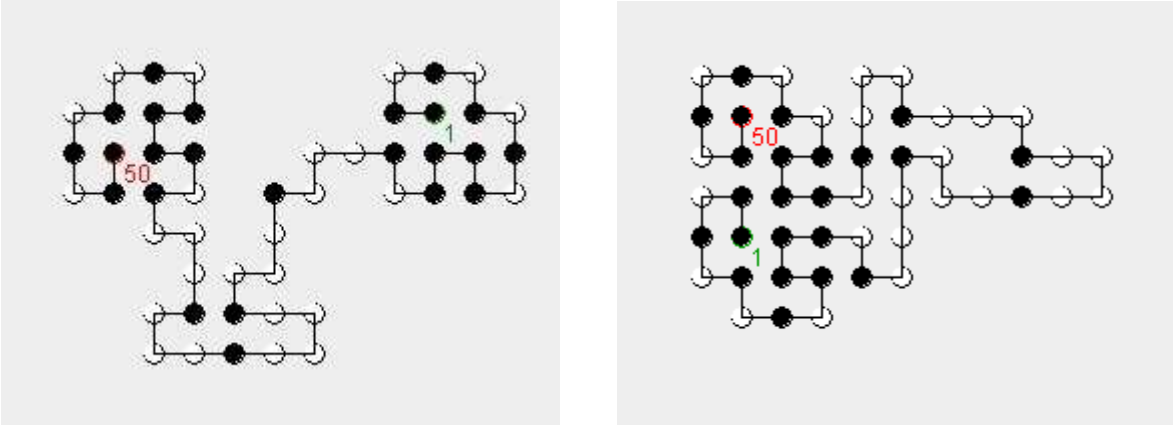

Figure 1: Example sequence S1-6: conformation (left) with energy value -20 that was found by a variant of PFold-P-ACO that did not use Constraint 2 and an optimal conformation (right) with energy value -21.

The η best conformations from the construction phase are selected as start conformations for local search. To each selected conformation the 4 best pull moves are applied. From all conformations that have been obtained the 4 best ones are selected for the next iteration of local search. After each iteration of local search it is checked whether a better conformation has been found. If so an iteration counter is set to zero, otherwise the iteration counter is increased by one. If the value of the iteration counter equals 10 the local search procedure is stopped. Some restrictions are applied during the local search: i) a conformation is selected only if it has at least 0.7 as much HH-contacts as the so far best found conformation, ii) for each conformation a tabu list that contains the last 5 pull moves is used in order to hinder that pull moves are reversed, iii) a conformation is only accepted if its diameter is at most $(4/3)\sqrt{n}$ or if it has more HH-contacts than the so far best found solution.

It should be mentioned that even when our tests have shown that these constraints make it more easier for the algorithm to find good conformations for the test sequences it can in principal happpen that the constraints make it impossible for the algorithm to find the optimum for some HP sequences.

4.2 Branch-and-Bound Part

The branch-and-bound process starts with two traces. One trace starts folding at sequence position s_1 and the other at sequence position s_n. Both traces work independently but exchange information on the energy of new best conformations. This information is used to estimate whether a partial conformation can potentially reach a new best energy value or should be cut. A mix between breadth first search and depth first search is done. More exactly, the algorithm searches on a level l of the tree (breadth first search) until it contains more than 300 nodes. The 150 best of these partial conformations are extended to level $l+1$ and so on. Only when the algorithm does not find a conformation (because all branches on corresponding subtrees are cut as described afterwards) the other 150 partial conformations on level l are used. The following five criteria are used to heuristically decide whether the search tree is extended or cut at a leaf.

Criterion 1. This criterion uses the pheromone information as constructed by the ACO part. Consider a leaf of the search tree and assume that the corresponding partial conformation C consists of $s_1, \ldots, s_{i-1}$ and the element to be placed is s_i. Let $I_{max} = \sum_{j=1}^{i} max\{S_j, L_j, R_j\}$ be the sum of the maximum pheromone values for the first i decisions. Then an extension of C with decision $d \in \{S, L, R\}$ for placing s_i is not considered if the sum of pheromone values corresponding to the extended partial conformation is smaller than $I_{max} \cdot \Phi_l$ where $0 < \Phi_l < 1$ is a parameter.

Criterion 2. For each H element i in the sequence a minimum energy value is computed that has to be reached by a partial conformation that consist of elements $s_1, \ldots, s_i$. This minimum value is based on the average energy value

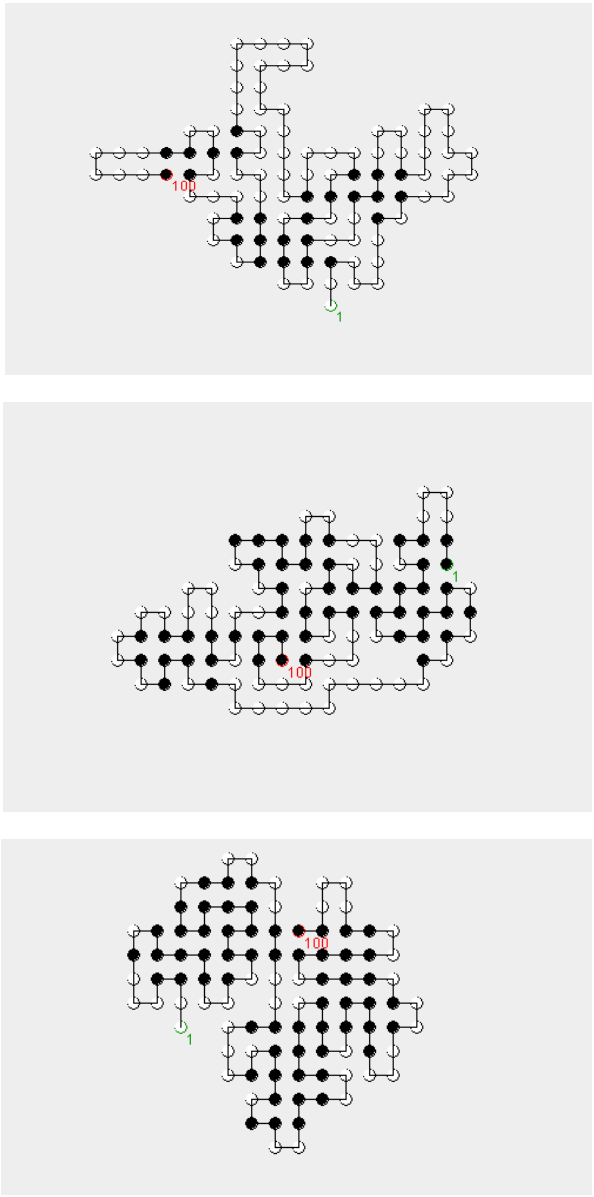

Figure 2: Best found conformations for D100-1 (upper), D100-2 (middle) and D100-3 (bottom).

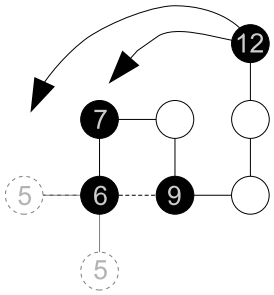

Figure 3: HH-contacts (6,9) and (7,12) are in conflict: if a connection between 6 and 9 is established it is not possible to connect 7 and 12.

$E_{avg}(i)$ of the prefixes of length i of the conformation in the population delivered by the ACO part. The longer the conformation becomes the higher the required energy value. For $i \in [1, n/2]$ the minimum value $\lfloor E_{avg}(i) \rfloor + 1$, for i from $n/2+1$ up to the position before the last few H elements the minimum value $\lfloor E_{avg}(i) \rfloor$ is used. For the rest of the sequence it is required that the best so far found energy value can still be obtained.

Criterion 3. It is checked whether it is possible to extend the current partial conformation so that it can become a new best found conformation. This is done by computing the potential that the partial conformation and the rest of the sequence have. This potential can be calculated according to the following formula:

$$
\begin{aligned}
E_{pot} = \\
& E_{cur} - min\{2 * \#even + even_0, \sum_{i=1}^{\sharp e} free_{even,i}\} \\
& \quad - min\{2 * \#odd + odd_0, \sum_{i=1}^{\sharp o} free_{odd,i}\} \\
& \quad - min\{ \\
& \qquad max\{0, 2 * \#even + even_0 - \sum_{i=1}^{\sharp e} free_{even,i}\}, \\
& \qquad max\{0, 2 * \#odd + odd_0 - \sum_{i=1}^{\sharp o} free_{odd,i}\} \\
& \quad \}
\end{aligned}
$$

The computation considers: i) the energy value of the current partial conformation E_{cur}, ii) the number $free_{even,i}$ ($free_{odd,i}$) of free and reachable locations next to an H element in the partial conformation where i is an index to number the H elements at even (respectively odd) positions within the sequence and $\sharp e$ (respectively $\sharp o$) is the total number of H elements with an even (respectively odd) index the partial conformation, iii) the number of H elements with even (odd) indices $\#even$ (respectively $\#odd$) that have not been placed. It should be noted that a location is considered "reachable" in the definition of $free_{even,i}$ ($free_{odd,i}$) if at least two neighbored locations are not occupied by the partial conformation or at least one neighbored position is free and the sequence ends with an H element on an even (odd) position and the empty location can potentially be filled by an even (odd) element. Note that every inner H element in the rest of the sequence could (potentially) have two possible H neighbors when it is placed. If the last element of the sequence is an H element it could (potentially) have three H neighbors when it is placed. Accordingly, the value of $even_0$ (odd_0) is 1, if the sequence ends with an H element and the sequence length is even (odd) and otherwise 0.

Criterion 4. For a long subsequence that consists only of P elements (pure P subsequence) there exists many possibilities how to fold it but the energy value of the partial conformation will not change during. Therefore, for all

Table 2: Energy of best conformation found with PERM [13, 26], filter-and-fan [23] (F&F), ACO-HPPFP-3 [26] (HPPFP-3), and PFold-P-ACO (PFold-P); ^{a} (^{b},^{c},^{d}) energy value obtained only for 2/5 (2/5,1/2, 9/10) of the runs; ^{e} energy value obtained only for 3/5 of the runs, the other results obtained -47

Protein	PERM	F&F	HPPFP-3	PFold-P
S1-1	-9	-9	-9	-9
S1-2	-9	-9	-9	-9
S1-3	-8	-8	-8	-8
S1-4	-14	-14	-14	-14
S1-5	-23	-23	-23	-23
S1-6	-21	-21	-21	-21
S1-7	-36	-36	-36	-36
S1-8	-42	-42	-42	-42
S1-9	-53	-53	-53^b	-53
S1-10	-50	-50	-49	-49^d
S1-11	-48	-48	-47	-48^e
B30-6	-13	-	-13	-13
B30-9	-18	-	-18	-18
B50-5	-22	-	-22	-22^c
B50-7	-17	-	-17	-17^c
D1	-19	-	-19	-19
D2	-17	-	-17	-17
D100-1	-	-	-24^a	-24
D100-2	-	-	-42^a	-42
D100-3	-	-	-52	-52^c

Table 3: Average computation times for PERM [13, 26], filter-and-fan [23] (F&F), ACO-HPPFP-3 [26] (HPPFP-3), and PFold-P-ACO (PFold-P); the run times for s1-11 for PFold-P-ACO are averages over the runs that produced a conformation with energy -48.

Protein	PERM	F&F	HPPFP-3	PFold-P
S1-1	<1s	0s	<1s	0.06s
S1-2	<1s	2s	<1s	0.4s
S1-3	2s	0.5s	<1s	0.2s
S1-4	<1s	4s	4s	1.1s
S1-5	2s	10s	1m	13.3s
S1-6	3s	22s	15s	15.4s
S1-7	4s	56s	20m	4m
S1-8	78h	24s	1.5h	35m
S1-9	60s	1.3m	24h	4.5h
S1-10	-	8.6h	12h	15h
S1-11	8m	9h	10h	1.5h(-47)
				8.5h(-48)
B30-6	1.6s	-	70.9s	8.5h
B30-9	0.06s	-	0.06s	0.9s
B50-5	9.4s	-	13m	9.3m
B50-7	4.5m	-	2m	2.5m
D1	2s	-	4m	2.2m
D2	3.5h	-	16m	25m
D100-1	-	-	42.4h	3.5h
D100-2	-	-	38.5h	1.5h
D100-3	-	-	25.8h	14.5h

partial conformations for $s_1, \ldots, s_j$ where the last two elements of a pure P subsequence $s_i, \ldots, s_j$, $j \geq i+2$ are placed on the same location and which have an equal prefix of length i all those are cut which satisfy the following criterion: the weight of the prefix of length i of the conformation is less than the average weight of the prefixes of length i of this conformations. Basically the weight is high when the energy of the partial conformation is small, the corresponding pheromone value are high, and its potential is high (see for details [29]).

Criterion 5. If the weight of a partial sequence of length k is not at least 5% higher than the weight of its prefix of length $k-5$ the node is cut.

5 Experiments and Results

The parameter values used for PFold-P-ACO are: $\alpha = 1.2$, $\beta = 1.6$ and $\sigma = 0.05$. Each TTL counter has initial value $2.5 \cdot n$ when n is the length of the sequence. The ACO part of PFold-P-ACO stops after a maximum number of 100000 0iterations. The population size is 10 and the number of ants per iteration is 20. Test runs have been executed on a 2.8 GHz Intel Xeon double processor PC with 4GB RAM. The HP test sequences are 11 standard benchmark sequences from [28] (S-1,...S-11), 4 sequences that have been used in [26] from the PDB [22] (B-30-6,B-30-9, B-50-5,B-50-7), and 2 sequences from [26] (D-1, D-2). Moreover 3 sequences that are shown in Table 1 have been created by us using a method provided in [26] (D100-x, $x \in \{1,2,3\}$). For the sequences of length ≥ 85 10 runs have been made per test sequences and for the shorter sequences 100 runs.

In addition to the best existing ACO algorithm ACO-HPPFP-3 from [26] we compare PFold-P-ACO with another state of the art algorithm PERM [13] and with the very good algorithm filter-and-fan algorithm of Rego et al. [23] (F&F). A variant of PERM is used which folds from both sides and is called PERMt_{exp} in [26] where also the run times results for PERM and F&F can be found.

A comparison between PERM, F&F, ACO-HPPFP-3 and PFold-P-ACO can be found in tables 2 and 3. All algorithms produce very good results on the S1-x sequences but only PERM and F&F found the best results for all them. PFold-P-ACO found the optimal results for all these sequences with the exception of sequence S1-10. The other ant algorithm ACO-HPPFP-3 found the optimal values for all but the two sequences S1-10 and S1-11. With respect to run time PERM is often relatively fast, but has serious problems with some sequences, e.g., symmetric sequences (see also [13]). This can be seen for sequence S1-8 where PERM needs 78h, but ACO-HPPFP-3 needs only 1.5h and the other two algorithms need significantly less than $1h$. F&F has a similar run-

time on most S1-x Sequences as the ACO algorithms and is significantly faster on sequences S1-8 and S1-9. Unfortunately, so far we could not get results of F&F for the other sequences from the authors of [23]. On sequences B30-x, B50-x, and Dx algorithms PERM, ACO-HPPFP-3 and PFold-P-ACO obtained conformations with the same minimum free energy. On the long sequences D100-x both ant algorithms found conformations of the same minimum free energy values. The only difference is that ACO-HPPFP-3 could for two of the sequences (D100-1, D100-2) not find the minimum value in all runs whereas PFold-P-ACO could not find for one sequence (D100-3) the minimum value in all runs.

Comparing the ACO algorithms it should be mentioned that the results on S1-x, B30-x, B50-x, D-1 and D-2 from [26] were obtained on a one 2.4GHz processor PC, whereas for PFold-P-ACO we used a two processor 2.8GHz PC. On the other hand ACO-HPPFP-3 is written in C whereas PFold-P-ACO is written in Java. The results for sequences D100-x have been obtained by us for PFold-P-ACO and ACO-HPPFP-3 on the same two processor PC. Taking all this into account, it seems that PFold-P-ACO is slightly faster on the S-4,...,S-11 sequences (an exception is S-10) and seems slightly slower on the B30-x, B50-x, D-1 and D-2 sequence. On the long sequences D100-x PFold-P-ACO is clearly faster. Altogether, PFold-P-ACO seems faster on long sequences whereas both algorithms are similar on small and medium length sequences.

Figure 4 shows the influence of relative size of $\mathcal{H}$ compared to number of estimated HH-contacts with distance ≤ 9 in the final conformation (for details see [29]). It can be seen that the size of $\mathcal{H}$ has a strong influence on the run time. The results indicate that for long sequences a medium size number of prescribed HH-contacts is advantageous whereas for small sequences a larger number of prescribes HH-contacts is better.

The influence of parameters α and β (see Formula (1)) on the run time until an optimal solution is found is shown in Figure 5. Values $\alpha > 1$ and $\beta > 1$ are clearly advantageous. The best values for the tested proteins where for $1.1 \leq \alpha \leq 1.2$ and $1.6 \leq \beta \leq 1.9$.

6 Conclusions

A hybrid population based Ant Colony Optimization (ACO) algorithm for the HP-Protein folding problem has been proposed. The algorithm is called PFold-P-ACO and consists of an ACO part and a heuristic branch-and-bound part. The branch-and -bound part uses the pheromone information that is delivered by the ACO part. PFold-P-ACO is the first population based ACO (P-ACO) algorithm that has been designed for solving a bioinformatics problem. It was shown experimentally that algorithm PFold-P-ACO achieves on nearly all test

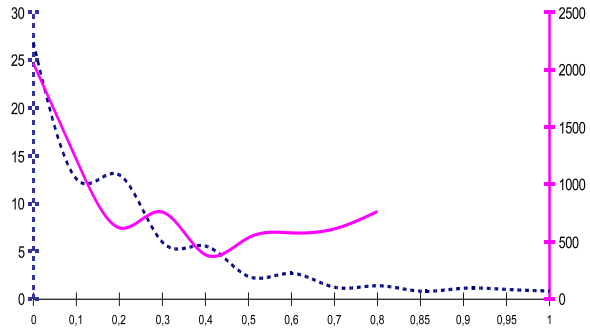

Figure 4: Run times in seconds until the optimum is found for S1-4 (dotted line, left scale) and S1-7 (right scale) for different values of set $\mathcal{H}$ compared to the number of estimated HH-contacts with distance at most 9 in the final conformation.

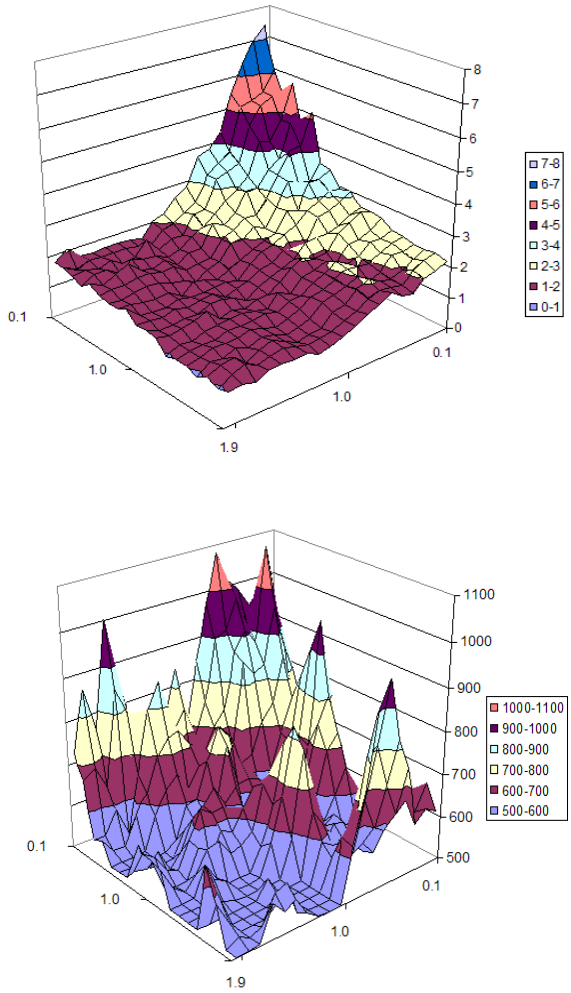

Figure 5: Run times in seconds until the optimum is found for S1-4 (upper) and S1-7 (lower) for different values of α and β. In both charts the left bottom scale represents α and the right bottom scale is β. The vertical scale shows the average run times.

sequences conformations with the same free energy values as other state of the art algorithms. Moreover, PFold-P-ACO is in general slightly better and also clearly faster on long sequences than the existing best ACO algorithm.

References

[1] Angus, D., "Population-Based Ant Colony Optimisation for Multi-objective Function Optimisation," *Proceedings ACAL 2007*: pp. 232-244, 2006.

[2] Angus, D., "Niching for Population-Based Ant Colony Optimization," *e-Science*, p. 115,2006.

[3] Angus, D., "Crowding Population-based Ant Colony Optimisation for the Multi-objective Travelling Salesman Problem," *Proc. IEEE Symp. Comp. Intell. Multicrit. Decision Making*, pp. 333-340, 2007.

[4] Bazzoli, A., Tettamanzi, G.B., "A Memetic algorithm for protein structure prediction in a 3D-lattice HP model," *Proceedings of EvoWorkshop*, LNCS3005, pp. 110, 2004.

[5] Berger, B., Leight, T., "Protein folding in the hydrophobichydrophilic (HP) model is NP-complete," *J. Comp. Biol.*, V5, N1, pp. 2740, 1998.

[6] Beutler, T. C., Dill, K. A., "A fast conformational search strategy for finding low energy structures of model proteins," *Protein Science*, V5, N10, pp. 2037-2043, 1996.

[7] García-Martínez, C., et al., "A Taxonomy and an Empirical Analysis of Multiple Objective ACO for Bi-criteria TSP," *EJOR*, V180, pp. 116-148, 2007.

[8] Dill, K. A., "Theory for the folding and stability of globular proteinst," *Biochemistry*, V24, pp. 15011509, 1985.

[9] Dorigo, M., Maniezzo, V., Colorni, A., "Positive feedback as a search strategy," *Tech Rep.*, 91-016, Dip Elettronica, Politecnico di Milano, Italy, 1991.

[10] Dorigo, M., Stützle, T., *Ant Colony Optimization*, The MIT Press, 2004.

[11] Guntsch, M., Middendorf, M., "A Population Based Approach for ACO," *Proc. EvoCOP-2002*, LNCS 2279, pp. 72-81, 2002.

[12] Guntsch, M., Middendorf, M., "Solving Multi-Objective Permutation Problems with Population Based ACO," *Proc. EMO 2003*, LNCS 2636, pp. 464-478, 2003.

[13] Hsu, H.-P., Mehra, V., Nadler, W., Grassberger, P., "Growth Alg. for Lattice Heteropolymers at Low Temp.," *J. Chem. Phys.* V118, pp. 444-452, 2003.

[14] Huang, W., Lü, Z., "Personification algorithm for protein folding problem: Improvements in PERM," *Chinese Sci. Bull.*, V49, N19, pp. 20922096, 2004.

[15] Krasnogor, N., Blackburne, P. B., Burke, E. K., Hirst, J. K., "Multimeme algorithms for protein structure prediction," *Proc. PPSN VII*, LNCS 2439, 2002, pp. 769778; .

[16] Lau, K. F., Dill, K. A., "A lattice statistical mechanics model of the conformation and sequence spaces of proteins," *Macromolecules*, V22, pp. 39863997, 1989.

[17] Lesh, N., M. Mitzenmacher, Whitesides, S., "A complete and effective move set for simple protein folding," *Proc. 7th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, ACM Press, pp. 188-195, 2003.

[18] Liang, F., Wong, W. H., "Evolutionary Monte Carlo for protein folding simulations," *J. Chem. Phys.*, V115, N7, pp. 33743380, 2001.

[19] Liebich, D., *Packungsprobleme bei Proteinen*, Shaker, 2005.

[20] Merkle, D., Middendorf, M., "Swarm Intelligence," *Search Methodologies - Introductory Tutorials* , Springer, pp. 401-435, 2006.

[21] Patton A. L., Punch, W., Goodman, E., "A standard GA approach to native protein conformation prediction," *Proceedings of the Sixth International Conference on Genetic Algorithms*, pp. 574581, 1995.

[22] *RCSB Protein Data Bank*, www.pdb.org

[23] Rego, C., Li, H., Glover, F., "A Filter-and-Fan Approach to the 2D Lattice Model of the Protein Folding Problem," *Manusscript under revision*, 2006.

[24] Shmygelska, A, Hernandez, R. A., Hoos, H. H., "An ant colony optimization algorithm for the 2D HP protein folding problem", *Proceedings ANTS 2002*, LNCS 2463, pp. 4052, 2002.

[25] Shmygelska, A., Hoos, H. H., "An improved ant colony optimization algorithm for the 2D HP protein folding problem," *Proc. Canadian Conf. on Artificial Intelligence*, LNCS 2671, pp. 400417, 2003.

[26] Shmygelska, A., Hoons, H. H., "An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem," *BMC Bioinformat.*, V6, p. 30, 2005.

[27] Socha, K., Dorigo, M., "Ant Colony Optimization for Continuous Domains." *European Journal of Operational Research*, V185, N3, pp. 1155-1173, 2008.

[28] *Standard Tortilla Benchmarks*, http://www.dmi.unict.it/∼mpavone/psp.html

[29] Thalheim, T., "Hybrid Ant Colony Optimization Algorithms for Solving the Protein Folding Problem," Diploma thesis, University of Leipzig, 2007.

[30] Unger, R., Moult, J, "Genetic algorithm for protein folding simulations," *J. Mol. Biol.*, V231, N1, pp. 7581, 1993.

[31] Yue, K., Dill, K. A., "Forces of Tertiary Structural Organization in Globular Proteins," PNAS, V92, N1, pp. 146-150, 1995.

[32] Yue, K., et al., "A Test of Lattice Protein Folding Algorithms," *PNAS*, V92, pp. 325329, 1995.

[33] Zhao, X., "Advances on protein folding simulations based on the lattice HP models with natural computing," *Applied Soft Computing*, in Press, 2007.