CMST 10(1) 7-19 (2004) DOI:10.12921/cmst.2004.10.01.07-19

COMPUTATIONAL METHODS IN SCIENCE AND TECHNOLOGY **10 (1)**, 7-19 (2004)

A TABU SEARCH STRATEGY FOR FINDING LOW ENERGY STRUCTURES OF PROTEINS IN HP-MODEL*

JACEK BŁAŻEWICZ[1, 2], KEN DILL[3], PIOTR ŁUKASIAK[1, 2], AND MACIEJ MIŁOSTAN[1]

[1] *Institute of Computing Science, Poznań University of Technology*
Piotrowo 3a, 60-965 Poznań, Poland

Maciej.Milostan@cs.put.poznan.pl

[2] *Institute of Bioorganic Chemistry, Polish Academy of Sciences*
Noskowskiego 12, 61-704Poznań, Poland

[3] *Department of Pharmaceutical Chemistry, University of California*
San Francisco, California, USA

(Rec. 18 November 2003)

Abstract: HP-model is one of the most successful and well-studied simplified lattice models of protein folding. It uses mathematical abstraction of proteins for hiding many aspects of the folding process and works as hypothesis generator. Due to the NP-hardness results of the protein folding problem many approximation algorithms, have been used to solve it. In the paper, the method for finding low energy conformations of proteins, based on the tabu search strategy, has been proposed. The algorithm has been extensively tested and the tests showed its very good performance.

1. INTRODUCTION

For the last few decades scientists have been trying to find a law that drives the protein folding process. Over forty years ago the experiments of Christian B. Anfinsen and co-workers [1,2] showed that proteins can fold reversibly, and native structures of some globular proteins are thermodynamically stable. They are the conformations at the global minima of their accessible free energies. Those observations quickly led to the puzzling aspects of theprotein folding problem, known as "Levinthal's Paradox" [15]. It can be paraphrased as follows "How can a folding protein choose so quickly among so many possible foldings the one with the minimum energy?" [5]. Throughout decades of studies and experiments many theories and models, concerning that question, have been widely investigated. Among these are some *simplified models,* sometimes called *simple exact models,* that use mathematical abstraction of proteins for hiding many aspects of the folding process and exaggerate the effects of the others.

One of the most successful and well-studied simplified models is the *HP-model* (H - hydrophobic amino acid, P - polar) proposed by Dill [6]. The aim of this model is to investigate the way hydrophobic interactions influence protein folds, whether delocalized "solvation" code is essential or not. HP-model was proposed on the basis of observations that

* This paper was written while the first author was visiting Laboratoire Leibnitz, IMAG, Grenoble, whose support is greatly appreciated.

most of hydrophobic amino acids are buried in the core of the protein, and moreover, there are very few conformations of the full chain that can bury nonpolar amino acids to the greatest possible degree [6-8].

In the basic HP-model each amino acid is represented as a bead (black bead - hydrophobic, white - polar) and the connecting bonds are represented as lines. The whole conformation is embedded in the two (or more) dimensional lattice. The background lattice simply divides the space into monomer-sized units. A lattice site may be either empty or filled with one bead. The bond angles have only a few discrete values, dictated by the structure of the lattice [7]. For each conformation one can compute the value of an energy function, which models free energies of protein folds. One of the simplest forms of the energy function counts each one of HH-contacts and multiplies it by a constant lower than zero. Two amino acids create HH-contact if they are topological neighbours and they are not connected by a bond (Fig. 1).

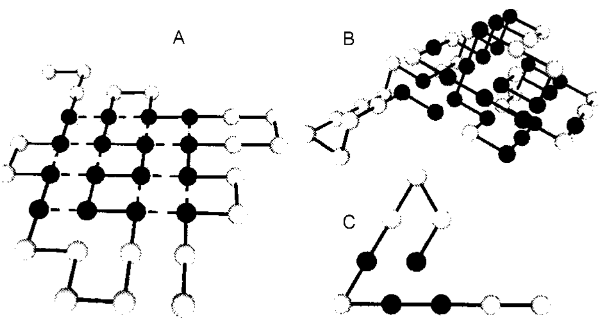

Fig. 1. Different types of conformations in HP-model (H - black, P - white): A) A native conformation on two dimensional lattice with included HH-contacts (dotted line) B) A suboptimal conformation on 3D lattice, C) A native conformation on 2D-triangular lattice

There have been several NP-hardness (or NP-completeness) results related to protein folding, in the simplified models. (Basically, it means that the problem considered cannot be solved optimally within reasonable time, i.e. polynomial time). In 1993 Unger and Moult showed that finding the native conformation in a number of simplified models is the NP-hard problem [22]. Four years later two independent teams proved that decision version of the folding problem in 2D and 3D HP-model is NP-complete [5, 4].

Due to the NP-hardness results many computer scientists have tried to find a fast approximation algorithms for searching an enormous number of the protein states. Most of the applied methods have been based on the search approaches such as Molecular Dynamics and Monte Carlo [3]. Ungert and Moult have shown an implementation of a Genetic Algorithm (GA), that is much faster than the traditional MC in the lattice tests [21]. O'Toole and Panagitopoulus have adapted some variants of chain growth method for folding co-polymer [19]. A core-directed growth method (CG), based on nearly the same idea have been proposed by Beutler and Dill [3]. Toma and Toma have shown the contact interactions method [20]. An approximation algorithm, that works in the linear time and guarantees that the energy of the found conformation is at least equal to 3/8 of the optimal value, have been shown by

Hart and Istrail [14]. Moreover at least one exact algorithm was proposed by Yue and Dill in [23], and is called CHCC. It searches the whole space of the conformations of a given sequence using Branch and Bound Method. An overview of the methods applied in HP-model can be found in [3]. In general, constraints and cutting that have been applied in the exact models restrict a complexity of the search space from 5^n (in case of 3D model, and sequence length n) to circa 1.125^n [24, 23], but the solution space remains still very large.

The aim of this paper is to implement the tabu search metaheuristic approach in the context of the considered problem [9, 10, 13]. Presented algorithm differs from the one proposed by Lesh and co-workers [16] at RECOMB 2003 mainly in the moves definition. It was proposed independently in the same time and presented during the poster session at the same conference. Extensive computational tests showed its very good performance for the HP-model in two-dimensional space.

The setup of the paper is as follows. Section 2 recalls basic HP-model. Section 3 presents the new method for the calculation of protein conformations based on the tabu search approach. Section 4 shows and discusses the results of the computational experiments. Section 5 contains conclusions.

2. PROBLEM FORMULATION

General idea of the HP-model is presented in the Introduction section, so in the following paragraphs more formal way of defining HP-model is given.

2.1. Two dimensional HP-model

A problem of finding a conformation with a minimal energy can be defined as a minimization of an energy function, defined as follows:

$$\min_{\mathbf{a}} E(\mathbf{s}, \mathbf{a}), \tag{1}$$

where:

$$E(\mathbf{s}, \mathbf{a}) = \xi \cdot HH_c(\mathbf{s}, \mathbf{a}), \tag{2}$$

where:

- **s** - a sequence of amino acids containing n elements; s_i — 1, if an amino acid on the *i*-th position in the sequence is hydrophobic; $s_i = 0$, if an amino acid on the *i*-th position is polar;
- **a** - a vector of (n - 2) angles defined by consecutive triples of amino acids in the sequence; a_i-∈ {0°, 90°, -90°};
- HH_c - a function that counts each pair of such hydrophobic amino acids, that are not neighbours in the sequence, but they are neighbours on the lattice (they are topological neighbours);
- ξ - a constant lower than zero, that defines the influence ratio of hydrophobic contacts on the value of conformational free energy; in most cases one can assume, that this ratio is equal to -1 ($\xi = -1$);

Additionally one has to introduce two more constraints: two neighbours in the sequence must be topological neighbours and each site of the lattice can be occupied only by one bead, so it is necessary to constrain values of Euclidean distance between a pair of amino acids - $dist_{i,j}$ **(a)**:

$$\forall_{(i \neq j)}\{dist_{i,j}(\mathbf{a}) > 0\}, \quad \text{where} \quad i, j \in \{1, 2 \ldots, n\} \tag{3}$$

$$\forall_{i}\{dist_{i,i+1}(\mathbf{a}) = 1\}, \quad \text{where} \quad i \in \{1, 2 \ldots, n-1\} \tag{4}$$

Each solution $x \in$

amino acid conformation. For each instance of the problem, the sequence **s** is explicitly given.

For the given sequence **s**, solution space Ω is a set of all possible pairs {**a**, **s**}. The size of the solution space $|\Omega|$ is the number of different values of vector **a**.

3. THE ALGORITHM

Following the model introduced in Section 2 we will present a new method for the calculation of the protein conformations based on the tabu search approach.

3.1. Tabu search for the basic HP-model

The proposed algorithm is based on Tabu Search (TS) strategy, which has been proposed by Fled Glover [9, 13]. One has to define some main elements of this strategy to adapt it to the problem. In the following paragraphs one can find definitions of such elements of the strategy in the questions like: neighbourhood, move, stop condition, tabu list, aspiration levels.

Move

Move r transforms each solution x from the set of all solutions Ω into another solution $x' \in \Omega$: $x \rightarrow^{r} x'$.

Three variants of the move are defined:

1. Move is defined as a change of a single angle in vector **a** for the sequence s as follows:
   (a) if the value of angle $a_i = 0°$, then the new value of angle a_i' can be taken as the one from the following set: {90°, -90°},
   (b) if the value of angle a_i - 90° or $a_i = -90°$, then the new value of angle a_i' can be equal to 0°.
2. Move is defined as a change of a single angle in vector **a** for the sequence s as follows:
   (a) value of angle a_i, can be substituted by the value $a_i' \in \{-90°, 0°, 90°\} \setminus \{a_i\}$.
3. Move is defined as a change of one, two or three consecutive angles from vector **a** for the given sequence s as follows:

(a) value of angle a_i, can be substituted by $a_i' \in \{-90°, 0°, 90°\}\backslash\{a_i\}$

(b) values of angles a_i and a_{i+1} where $i < n - 2$, can be substituted by a_i' and a_{i+1}' such that $a_i' \in \{-90°, 0°, 90°\}\backslash\{a_i\}$ and $a_{i+1} \in \{-90°, 0°, 90°\}\backslash\{a_{i+1}\}$

(c) values of angles a_i, a_{i+1} and a_{i+2}, where $i < n - 3$, can be substituted by a_i', $a_i'_{+1}$ and $a_i'_{+2}$, such that $a_i' \in \{-90°, 0°, 90°\}\backslash\{a_i\}$ and $a_i'_{+1} \in \{-90°, 0°, 90°\}\backslash\{a_{i+1}\}$ and $a_i{+}_2 \in \{-90°, 0°, 90°\}\backslash\{a,_{+2}\}$.

Proposed set of the moves guarantees fast search among different conformations of the proteins, because each move may have wide ranging effects. Changes of angles occur only on the consecutive positions, and in that sense moves are local.

Neighborhood

For the given sequence s solution x' is in the neighbourhood of solution x, if and only if a move such as $x \xrightarrow{r} x'$ exists.

Set of solutions $N(x)$ is called the neighbourhood if for each $x' \in N(x)$ move r such as $x \xrightarrow{r} x'$ exists.

Tabu list and tabu condition

The tabu list that contains forbidden moves, consists of single or triple angles changed in a single move, and the positions in which each change occurred. Formal definitions are as follows:

Definition 1

*The tabu list is a n-elementary cyclic list; each element of the list is an ordered pair {**b**, i}, where **b** is a m-elementary vector of modified (in single move) angle values; m is equal to a maximal number of modifications, that can be done in a single move; $i \in N$ and $i < n - 2$ is a position of the change (the position of the first angle from **b** in **a**).*

Into the tabu list inserted are those values of angles, that are being changed as the effect of doing a selected move. It means that reversed move becomes forbidden for next *n-iterations.*

Definition 2

*Set offeasible moves R is composed by only such moves r that transform solution x, described by angle vector **a**, into solution x', described by **a'**, such that conditions (3) and (4) are fulfilled.*

The tabu condition, that defines the set of forbidden moves $R_z \subseteq R$, can be defined on the basis of the definitions given above:

Definition 3

*Each move r belongs to the set offorbidden moves R_z, if and only if it transforms a solution given by **a** into a solution described by vector **a'**, such that a pair {**b**, i}, for which **b** = $[a_i'...,a_i'_{+m}]$, exists on the tabu list.*

Aspiration criteria

The searching procedure can make a transition from solution x into x' by making the move $r \in R_z$, if two conditions are simultaneously fulfilled:

- the value of energy $E(2)$ of solution x' is lower than energy E_{min} of the current best solution,
- there is no solution $x'' \in N(x')$, such that $r \notin R_z$, $x \longrightarrow^{r} x''$ and energy $E_{x''} \leq E$.

Stop criteria

The algorithm executes the searching procedure l_{runs} times (each execution is called a run), each time starting from a different solution (e.g. the best solution in previous run) and using the same or different variants of moves. Each run ends if one of the stop conditions is fulfilled:

- a given number l_{it} of iterations has been achieved,
- a given lower bound of energy E_{min} has been reached,
- a computed lower bound of energy E_{min} has been reached - computed by the sequence analysis,
- all possible moves are forbidden and aspiration criteria for any of them are not fulfilled,
- a given number of iterations $l_{it_{\max}}$ after reaching the upper bound, have been made.

Lower and upper bounds

A lower bound of energy is such in its value, that the optimal solution has the value equal or greater than the one mentioned. In fact, it is impossible to compute optimal energy only by analyzing amino acid sequence without conformations. In most cases the lower bound is underestimated (note, that the considered problem is a minimization one). It is lower than the energy of the optimal conformation. One can compute the lower bound on the basis of the observation, that in the optimal conformations, the shape of the core is usually the square like and the area of this "square" is equal to a number of hydrophobic amino acids. The simplest lower bound can be computed as follows:

$$E_{\min}(\mathbf{s}) = -n_H - a(n), \tag{5}$$

where:

- n_H-a number of hydrophobic amino acids in sequence **s**;
- $a(n) = \begin{cases} 1, & \text{if } \left\lceil\left(\sqrt{n}\right)\right\rceil^2 = \left\lceil\left(\sqrt{n}\right)\right\rceil \wedge \sqrt{n} \bmod 2 = 0 \\ 0, & \text{otherwise} \end{cases}$

An upper bound is such a value of the energy, that can be reached for sure and it can be given as follows:

$$E_{\max}(\mathbf{s}) = \max(-n_{H_e}, -n_{H_o}) + d(\mathbf{s}), \tag{6}$$

where:

$$-d(\mathbf{s}) = \begin{cases} 1, & \text{if in } \mathbf{s} \text{ not exists such } (s_i, s_j) \text{ that: } (j-i) > 1 \\ & \wedge (j-i) \text{ is odd } \wedge s_i = 1 \wedge s_j = 1 \wedge [s_{i+1,}\ s_{i+2} \dots, s_{j-1}] = 0 \\ 0, & \text{in other case} \end{cases}$$

n_{H_e} - a number of hydrophobic amino acids on even positions in sequence **s**,

n_{H_o} - a number of hydrophobic amino acids on odd positions in sequence **s**.

Generation of the starting solution

Some conformational motifs can be found quite often in optimal conformations, so the idea of creating the starting solutions is to use this knowledge to create some desirable conformational fragments at the beginning of the search. During the search this conformational fragments can be changed.

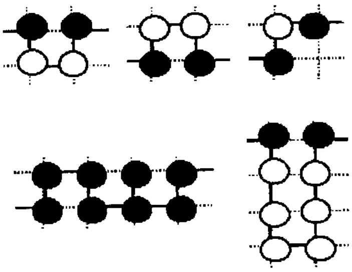

Fig. 2. Some short motifs from optimal conformations

Some samples of such motifs are presented in Fig. 2.

Strategy of choice

The neighbourhood $N(x)$ of given solution x can consist of many solutions with equally good energies. In such case one has to decide which one to choose for the next step. The strategy of choice is the way of choosing two elements:

- turning point(s) - an element of angle vector **a** that should be changed,
- angle(s) value(s) - the value of its change.

Turning points can be chosen randomly, or the next element from the vector of the angles can be preferred. Additionally the choice of a particular angle value can be related to the frequency of its occurrence on the given position. One can also search for the particular motifs in the sequence, where changes have any sense, e.g. one can assume, that the most preferred turning points are triplets *hph* in the sequence. In fact, the choice of a particular method or a motif is greatly dependent on the problem instance.

Diversification

The aim of the diversification procedure is to change the focus of the searching procedure from a current fragment of the conformational space to another, probably not yet explored.

One of the diversification methods can be rewriting **s** in the reverse order, without making any changes to the vector **a** of the angles.

Constraints on the solution space

Solution space Ω can be easily constrained on the basis of the observation that for each solution x with nonzero angle vector **a**, solution x' exists, such that angle vector $\mathbf{a}' = (-1)\mathbf{a}$. A conformation defined by a is the mirror image of the conformation defined by $\mathbf{a}'$, so both conformations have the same energy.

Taking into account the above observations one can easily propose the searching strategy, that constrains solution space Ω to $\Omega \subset \Omega$, such that:

$$|\Omega'| = \frac{|\Omega|}{3}\left(\frac{3}{2} - \frac{1}{2\cdot 3^{n-4}}\right) = \frac{3^{n-2}}{3}\left(\frac{3}{2} - \frac{1}{2\cdot 3^{n-4}}\right) = \frac{3^{n-2}}{2} - \frac{3^{n-3}}{2\cdot 3^{n-4}} = \frac{3^{n-2}-3}{2} = \frac{|\Omega|-3}{2} \quad (7)$$

Additionally one can apply some simple conditions to constrain the solution space more efficiently - for example one can add definitions of some forbidden combinations of the angles values (e.g. values of three consecutive angles cannot be equal to 90°) in the aim not to compute values of energy for unrealistic conformations.

4. RESULTS OF COMPUTATIONAL EXPERIMENTS AND THEIR DISCUSSION

The algorithm was implemented in *C* language and tested for a wide range of parameters on several benchmark sequences commonly found in the literature [3, 16, 20, 21].

In this section the results of the tests are presented. The relations among the tabu list sizes, the energy values, and the time of computations are shown on the charts.

Instances used for testing purposes can be found in Table 1.

Most of all tests have been done on PC with AMD Duron 700 MHz processor, under Linux OS.

The computational experiment contained the following phases:

1. A test of the algorithm for the parameters (see below) chosen during the pre-tests in an implementation phase. Tests where made on the whole space of sequences from 5 to 12 amino acids length. The optimal solutions were found before using an exhaustive enumeration.
2. A test of the algorithm on instances from [3] with loosely-tuned parameters (cf. Table 1).
3. Final tuning of the algorithm parameters for the above instances (cf. Table 1).

In the first phase of the test, the tabu algorithm with the following parameters has been used:

- the second variant of moves;
- the stop condition: as defined above with a number of runs equal to two; a number of iterations for the first phase was equal to 1000 (per run);

- the aspiration criteria were used;
- for choosing the turning point a pseudorandom generator was used;
- a procedure that generates the starting solution on the basis of the conformational motifs, was used.

Table 1. Test sequences for two dimensional HP-model from [3] and [16].H - hydrophobic, P - polar

No.	Len.	Sequence	E_{opt}
1	20	HPHPPHHPHP PHPHHPPHPH	–9
2	24	HHPPHPPHPP HPPHPPHPPH PPHH	–9
3	25	PPHPPHHPPP PHHPPPPHHP PPPHH	–8
4	36	PPPHHPPHHP PPPPHHHHHH HPPHHPPPPH HPPHPP	–14
5	48	PPHPPHHPPH HPPPPPHHHH HHHHHHPPPP PPHHPPHHPP HPPHHHHH	–23
6	50	HHPHPHPHPH HHHPHPPPHP PPHPPPPHPP PHPPPHPHHH HPHPHPHPHH	–21
7	60	PPHHHPHHHH HHHHPPPHHH HHHHHHHPHP PPHHHHHHHH HHHHPPPPHH HHHHPHHPHP	–35
8	64	HHHHHHHHHH HHPHPHPPHH PPHHPPHPPH HPPHHPPHPP HHPPHHPPHP HPHHHHHHHH HHHHH	–42
9	85	HHHHPPPPHH HHHHHHHHHH PPPPPPHHHH HHHHHHHHPP PHHHHHHHHH HHHPPPHHHH HHHHHHHHPP PHPPHHPPHH PPHPH	–53
10	100	PPPPPPHPHH PPPPPHHHPH HHHHPHHPPP PHHPPHHPHH HHHPHHHHHH HHHHPHHPHH HHHHHPPPPP PPPPPPHHHH HHHPPHPHHH PPPPPPHPHH	–48
11	100	PPPHHPPHHH HPPHHHPHHP HHPHHHHPPP PPPPPHHHHH HPPHHHHHHP PPPPPPPPHP HHPHHHHHHH HHHHPPHHHP HHPHPPHPHH HPPPPPPHHH	–50

The optimal solutions for all sequences in range from 5 to 12 were found in the first phase. A size of the tabu list is strongly dependent on the instance of the problem, but some sizes of the tabu list gave the best results. In figures (Fig. 3 and 4) one can find average distances from the optimum and the percentage of the optimal solutions found in relation to the tabu list sizes, for different lengths of the sequences. Most of the optimal solutions were found (at least for one the tabu list size) after the first run (before diversification).

In the second phase of the tests, firstly computations for loosely-tuned (the same for all sequences) parameters have been made, and later on better parameters for each sequence have been found. The tabu algorithm with the following parameters has been used:

- the second variant of the moves was used in the first phase, and the third in the second phase;
- the stop condition: as defined above with a number of runs equal to two, a number of iterations for the second phase was equal to 1000 iterations;
- the aspiration criteria were used;
- for choosing turning points frequencies of the angles' values were used;

- the procedure generating the starting solution was used;
- a more preferred change of an angle in the chosen position was a change from value 0° to 90°;
- a tabu list size was equal to 20.

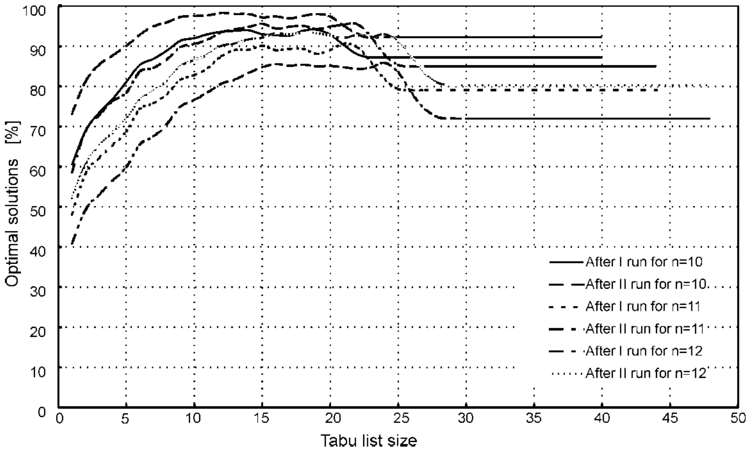

Fig. 3. A relation between optimal solutions found and the tabu list sizes for sequences from 10 to 12 amino acids length

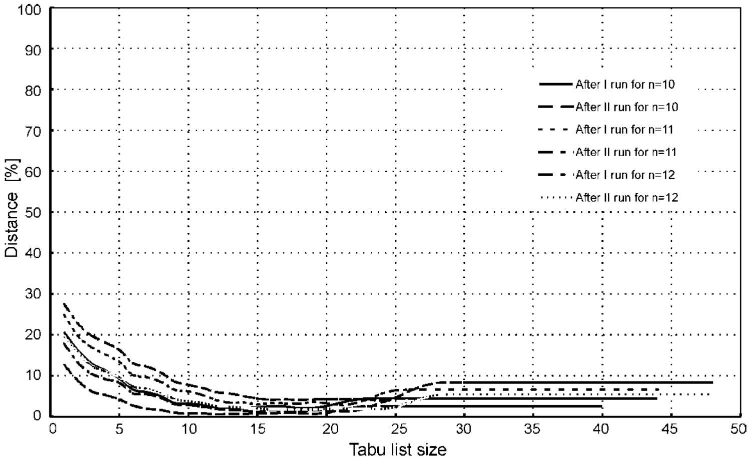

Fig. 4. A relation between average distances from the optimal solutions (in sense of differences in values of energy) and the tabu list sizes for sequences from 10 to 12 amino acids length (the whole spectrum of sequences has been considered)

The computations for each sequence have lasted no longer then 30 seconds, and the optimal solution for sequences 1, 2, 3, 4, 6 has been found, for the sequence no. 5 conformation with the energy equal to -22 has been found, for the sequence no. 7 the energy was equal -33 and for the sequence no. 8-39 (cf. Table 1).

In the third phase the parameters of the tabu algorithm were fine-tuned for each sequence independently, and then for the sequences no. 7 and 8 optimal solutions have been found. That is, the tabu list sizes were changed in the way shown in Table 2, while variant of the moves in the first run was changed from the second one into the first (for sequences 1, 3, 6). What's more for sequence no. 7 a number of iterations in each run was increased up to 10000, aspiration conditions was turn off in the first run, the tabu list size was lowered to 10. For the sequence no. 8 a change of the tabu list size only was necessary - the optimal solution were found just after 378 iterations and 43597 energy computations (cf. [16] sequence S64). For sequence no. 5 the tabu list size was lowered to 5, for choosing turning points a pseudo-random generator was used, a number of iterations was increased up to 2000 per run, the procedure generating starting solution was not used and the search space was constrained by fixing the value of the first angle in the sequence to 90° (after a diversification of the last angle in the sequence).

Table 2. A comparison of the results obtained by methods GA, CI, CG, and TS, based on [3] for 2D HP-model. Times for GA (Genetic Algorithms) [21] and CI (Contact Interactions) [20] methods were estimated. For CG [3] method an average time to find the optimum were presented. Empty field in the table means, that no optimal solution was found for the method in question

No.	Length	t_{GA} [s]$^{1)}$	t_{CI} [s]$^{1)}$	t_{CG} [s]$^{1)}$	t_{TS} [s]$^{2)}$	E_{opt}	$E_{TS}$$^{3)}$	$TL_{size}$$^{4)}$	$N_{hydr}$$^{5)}$
1	20	5.60	0.11	2.16	0.03	–9	–9	10	10
2	24	6.00	0.24	6.60	0.02	–9	–9	20	10
3	25	3.66	0.18	4.68	<0.01	–8	–8	6	9
4	36	54.60	6.60	132.00	0.34	–14	–14	20	16
5	48		34.80	378.00	3.70	–23	–23	5	25
6	50	3180.00	0.24	18600.00	0.57	–21	–21	10	24
7	60		60.00	5820.00	170.00	–35	–35	10	43
8$^{6)}$	64			546.00	0.54	–42	–42	23	42
9$^{6)}$	85				80.00	–53	–51	10	59
10$^{6)}$	100				2199.24	–48	–45	40	55
11$^{6)}$	100				6068.00	–50	–48	20	56

1) The algorithm performed on SUN SPARC1.
2) The algorithm performed on AMD Duron 700 MHz.
3) The best energy found by TS.
4) The tabu list size.
5) Number of hydrophobic amino acids in a sequence.
6) The optimal solution was also found by GTabu method used by Lesh [16] in time < 1800 [s]. For sequences 1-7 results for that method were not presented.

For longer sequences from [16] (no. 9-11) our algorithm did not find optimal solutions, but for roughly chosen parameters quite good suboptimal solutions were found in a reasonable time (cf. Table 2).

Method comparison

In comparison to the well known strategies, mentioned in the introduction, the tabu search algorithm (TS) gives very good results for the HP-model with a rectangular lattice. In Table 2 one can see a comparison of the tabu search (TS) results with those given in the Beutler and Dill [3] for Genetic Algorithms (GA), Contact Interactions (CI) and Coredirected Growth Method (CG).

In fact, one can say that ranges of the computation times for TS place it as one of the leaders.

5. CONCLUSIONS

In this paper, a new method for finding low energy conformations of proteins in the HP-model, has been proposed. It is a variant of the tabu search strategy adapted for the considered problem. The method was implemented independently from that presented in [16] and was presented during the same conference - RECOMB 2003. It uses the problem domain knowledge, such as conformational motifs. The proposed algorithm has a very good performance and finds the optimal conformations for all short sequences from 5 to 12 amino acids length. For eight benchmark sequences for the 2D HP-model with the rectangular lattice optimal solutions were found in a very short time in comparison with other heuristic approaches.

Acknowledgements

This work was partially supported by the KBN grant.

References

[1] C. B. Anfinsen, E. Haber, M. Sela, F. H. White, Jr., *The kinetics offormation of native ribonuclease during oxidation of the reduced polypeptide chain,* Proc. Natl. Acad. Sci. USA **47**, 1309-1314 (1961).

[2] C. B. Anfinsen, *Principles that govern thefolding ofprotein chains,* Science **181**, 223-230 (1973).

[3] T. C. Beutler, K. A. Dill, *A fast conformational search strategy for finding low energy structures ofmodel proteins,* Protein Sci. **5**, 2037-2043 (1996).

[4] B. Berger, T. Leighton, *Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete,* J. Comp. Biol. **5(1),** 27-40 (1998).

[5] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, M. Yannakakis, *On the complexity of proteinfolding,* Proc. 1998 STOC, andJ. Comp. Biol.. **5(2)** (1998)

[6] K. A. Dill, *Theory for the folding and stability of globular proteins,* Biochemistry **24**, 1501-1509 (1985).

[7] K. A. Dill, S. Bomberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, H. S. Chan, *Principles of proteinfolding: Aperspectivefrom simple exact models,* Protein Sci. **4**, 561-602 (1995).

[8] K. A. Dill, *Polymer principles andproteinfolding,* Protein Sci. **8**, 1166-1180 (1999).

[9] F. Glover, *Tabu Search - Parti, ORSA* J. Comp. **1**, 190-206 (1989).

[10] F. Glover, *Tabu Search Fundamentals and Uses,* Graduated School of Business, University of Colorado, Condensed version published in Mathematical Programming: State ofthe Art, Birge and Murty (eds.) 64-92 (1994).

[11] F. Glover, *Tabu Search and Adaptive Memory Programing - Advances, Applications and Challenges. Interfaces in Computer Science an Operations Research* (1996), Kluwer Academic Publishers, 1-75.

[12] F. Glover, M. Laguna, *Tabu Search. Modern Heuristic Techniques for Combinatorial Problems,* Blackwell Scientific Publishing, Oxford, 70-141.

[13] F. Glover, M. Laguna, *Tabu Search,* Kluwer Academic Publishers (1997), Boston, USA, 1-357.

[14] W. E. Hart, S. Istrail, *Fast protein folding in the hydrophobic-hydrophilic model. within three-eights of optimal,* J. Comp. Biol. **3(1)**, 53-96 (1996).

[15] C. Levinthal, *Are there pathways for proteinfolding?* Chem. Phys. **65**, 44-45 (1968).

[16] N. Lesh, M. Mitzenmacher, S. Whitesides, *A complete and effective move setfor simplifiedprotein folding,* RECOMB Proc. 188-195 (2003).

[17] K. F. Lau, K. A. Dill, *A lattice statistical mechanics model of the conformational and sequence space of proteins,* Macromolecules **22**, 3986-3997 (1989).

[18] J. T. Ngo, J. Marks, M. Karplus, *Computational complexity, protein structure prediction, and the Levinthal paradox.* in *The protein folding problem and tertiary structure prediction,* edited by K. M. Merz and S. M. Le Grand, Birkhauser, Boston (1994).

[19] E. M. O'Toole, A. Z. Panagiotopoulos, *Monte Carlo simulation of folding transitions of simple model proteins using a chain growth algorithm,* J. Chem. Phys. **97**, 8644-8652 (1992).

[20] L. Toma, S. Toma, *Contact Interactions Method: A new algorithm for proteinfolding simulations,* Protein Sci. **5,** 147-153 (1996).

[21] R. Unger, J. Moult, *Genetic algorithms for protein folding simulations,* J. Mol. Biol. **231**, 75-81 (1993).

[22] R. Unger, J. Moult, *Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications,* Bull. Math. Biol. **55(6)** 1183-1198 (1993).

[23] K. Yue, K. A. Dill, *Sequence-structure relationships in proteins and copolymers,* Phys. Rev. E **48(3)** 2267-2278 (1993).

[24] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. L. Shakhnovich, K. A. Dill, *A test of lattice proteinfolding algorithms,* Proc. Natl. Acad. Sci. USA **92**, 325-329 (1995).