Serdica J. Computing **5** (2011), 359–366

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

INTEGER PROGRAMMING APPROACH TO HP FOLDING*

N. Yanev, P. Milanov, I. Mirchev

Abstract. One of the most widely studied protein structure prediction models is the hydrophobic-hydrophilic (HP) model, which explains the hydrophobic interaction and tries to maximize the number of contacts among hydrophobic amino-acids. In order to find a lower bound for the number of contacts, a number of heuristics have been proposed, but finding the optimal solution is still a challenge. In this research, we focus on creating a new integer programming model which is capable to provide tractable input for mixed-integer programming solvers, is general enough and allows relaxation with provable good upper bounds. Computational experiments using benchmark problems show that our formulation achieves these goals.

1. Introduction. The challenge of inferring a protein's three-dimensional structure from its sequence is known as the Protein Structure Prediction (PSP) problem. The primary structure of the protein is the sequence of these amino-acid residues from amino terminus to carboxyl terminus. Thus the primary structure is a string over the twenty-letter alphabet of amino-acid types.

ACM Computing Classification System (1998): G.1.6.

Key words: Protein folding, HP model, integer programming.

*This work was supported by NFSR of Bulgaria, projects DOO2-162/16.12.2008, DOO2-135/31.07.2009 and DO 02-359

The goal of PSP is to predict the tertiary structure of proteins, which is the three-dimensional shape of the folded protein. PSP is a problem for which there is no unique formulation. There are several models of protein folding, and they generally fall into one of the two categories: off-lattice and on-lattice. Off-lattice models allow the protein's components to move, free-floating, in a continuous space. On-lattice models map the protein's components to points on a discrete lattice. The goal of on-lattice models is to confine the size of the energy landscape and reduce the protein-folding problem to its simplest form. The HP model generalizes amino-acids by partitioning them to two sets: hydrophilic (attracted to water) and hydrophobic (repelled by water). The HP model exploits the dominance of the hydrophobic-hydrophobic contact in the folding event in order to drastically reduce the problem's complexity. [5] is a comprehensive survey of combinatorial algorithms and theorems about lattice protein-folding models obtained 15 years after the publication in 1995 of the first protein-folding approximation algorithm with mathematically guaranteed error bounds [3]. The results presented here are mainly about the HP-protein-folding model introduced by Ken Dill in 1985 [4] and culminated later in [6].

The HP folding in a lattice differs from other folding approaches in two points. First, while most approaches rely on the full alphabet of amino-acids (20 letters), the HP folding uses a simplified two-symbol alphabet, in which each amino-acid is either Hydrophobic "H" or Polar "P" (as presented in table 1). Secondly, the 3D space in which the sequence is to fold is discretized (in lattice models) into a 2D or a 3D lattice (in our case, a 3D cubic lattice). For protein folding in HP-models, the optimization problem is defined as follows. Given an HP-model and a protein sequence over the binary alphabet of hydrophobic-hydrophilic amino-acids, find the protein fold in the model that has the maximum number of contacts. This optimization problem is indeed NP-complete; however, a collection of approximation algorithms exists for a variety of HP-models.

Definition 1. *In the* HP *folding in a lattice, a possible fold is called a* ***self-avoiding walk****, and consists of placing the amino-acids (*H/P *letters) from the sequence into the lattice, with the following constraints:*

– all amino-acids from the sequence must be placed into the lattice,

– a cell of the lattice can contain at most one amino-acid from the sequence,

– two amino-acids that are consecutive in the sequence must be placed in cells that are neighbors in the lattice.

Definition 2. *A hydrophobic contact (h-h contact) occurs when two hydrophobic amino-acids which are not neighbors in the sequence are placed in*

adjacent cells in the lattice.

Definition 3. *An **optimal fold** is a self-avoiding walk which possesses a maximum number of hydrophobic contacts.*

Table 1. Hydrophobic/Polar classification of the 20 α-amino-acids.

Name	Symbol	Classification	Name	Symbol	Classification
Alanine	A	Hydrophobic	Leucine	L	Hydrophobic
Arginine	R	Polar	Lysine	K	Polar
Asparagine	N	Polar	Methionine	M	Hydrophobic
Aspartic Acid	D	Polar	Phenylalanine	F	Hydrophobic
Cysteine	C	Polar	Proline	P	Hydrophobic
Glutamic Acid	E	Polar	Serine	S	Polar
Glutamine	Q	Polar	Threonine	T	Polar
Glycine	G	Polar	Tryptophan	W	Hydrophobic
Histidine	H	Polar	Tyrosine	Y	Polar
Isoleucine	I	Hydrophobic	Valine	V	Hydrophobic

Finding the optimal fold even in the simplest case of 2D lattice model is an NP-hard problem and still far from being efficiently solved. An attempt towards this end, focused on identifying the efficiently computable upper bound, is outlined in [2]. They present a new mathematical formulation of the HP model, which can provide an upper bound using a linear relaxation of the formulation. Computational experiments using benchmark problems show that the formulation provides a tight upper bound (see below).

Let E and O represent the sets of H amino-acids in an even position and an odd position of the sequence, respectively. Since a 2D square lattice is a bipartite graph, each H amino-acid in an even position can have contacts with H amino-acids in an odd position. Similarly, each H amino-acid in an odd position can have contacts with H amino-acids in an even position. If an H amino-acid is not in the first or last position of the sequence, it can have two HH pairs of contacts at most. Otherwise, it can have three HH pairs of contacts at most. Therefore, the following upper bound can be calculated:

$\mathrm{UB} = 2\min\{|\mathrm{E}|, |\mathrm{O}|\} + k$, where k is equal to 0, 1, 2, depending on the following characteristics: no H amino-acids are placed in the first and last positions, one H amino-acid is placed in the first or last position, two H amino-acids are placed in the first and last positions. The very sophisticated formulation in [2] provides this bound only experimentally and there are no indications as to the expense (the time needed to solve the respective linear programming relaxation). Our goal here is to derive an efficient integer programming model for the HP fold-

ing on lattices and even on graphs. The model is very compact (with respect to the number of integer variables) and such that the optimal value of the objective function of its LP relaxation is less than or equal to the above-mentioned bound.

2. Integer programming model. In order to derive the model, we will pose it as a contact map overlap problem (CMO) [1]. The input HP-sequence is considered as a graph $Seq = \{V, A\}$, with node and edge sets defined as: $V = \{v_1, v_2, \ldots, v_l\}$, $A = \{v_i, v_{i+1}\}$. The nodes are labeled $L(v) =$ "H/P" in accordance with the input sequence.

Let $Pair = \{i, i \mod 2 = 0,\ L(v_i) = \text{"H"}\}$, $Odd = \{i, i \mod 2 = 1,\ L(v_i) = \text{"H"}\}$. The sets B and B' will be frequently used below and are defined as: $B = Pair$, $B' = Odd$ if $|Pair| \leq |Odd|$ and vice versa.

Remark. The sets $Pair/Odd$ are eligible only if HP-folding is on a 2D square or a 3D cubic lattice. The model below is demonstrated on a 2D square but its extrapolating on a face-centered cubic lattice and/or graph models is self-evident.

The lattice is modeled as a graph $L = \{U, E\}$, $U = \{u_{ik}, i, k = 1, 2, \ldots, m\}$, $E = \{(u_{ik}, u_{jl})\}$, $|i - j| + |k - l| = 1$. (m is a reasonable estimate of the lattice size.) Any elementary chain (without cycles) in E is a self-avoiding walk and the set of those isomorphic to Seq define the set of feasible solutions. If the edge set A is extended to $A + A'$, $A' = \{e_{ij} = (v_i, v_j)\}$, $i \in B$, $j \in B'$, $|i - j| > 1$ then the HP folding becomes the problem of finding a feasible solution having a maximum of common edges with A'. Formally speaking, if $f : V \to U$ maps the input sequence to a feasible solution (self-avoiding walk) set, then the common edges are those with $e_{ij} \in A''$ and $(u_{f(i)}, u_{f(j)}) \in E$.

The only difference from CMO is that the set U is only partially ordered and this prevents the direct use of the algorithms for solving it, but at least the platform created for modeling maps such as f (matching in graphs) could be used for the purposes of this section.

Let us call an alignment graph the following $I \times l$ grid, with $|I|$, ($I = \{1, 2, \ldots, m^2\}$) rows and l columns. Let $g : U \to I$ be an arbitrary embedding of U in I, for instance $g(ik) = mk + i$. Then the edges in the grid say (ik, jl) (the two indices are for row-column coordinates of the grid nodes) if $k \in B/B'$, $l \in B'/B$ and $(g^{-1}(i), g^{-1}(j)) \in E$. Notice also that the grid vertices instantiate the f map. If x_{ik} is a binary variable equal to one when the vertex ik is to be chosen and vice-versa then a self-avoiding walk is a solution of the following

system:

$$\sum_{i=1}^{m^2} x_{ik} = 1, \quad k = 1, \ldots, l; \tag{1}$$

$$\sum_{k=1}^{l} x_{ik} \leq 1, \quad i = 1, \ldots, m^2; \tag{2}$$

$$x_{ik} \leq \sum_{j \in \delta(i)} x_{jk+1} \quad k = 1, \ldots, l-1, \quad i = 1, \ldots, m^2. \tag{3}$$

In the equations above: each v is aligned, each u is aligned with one v at most, neighbors in Seq are aligned to neighbors in L (in (3) neighbors of i are denoted as $\delta(i)$).

Let now t_{ik}, $k \in B$ denote the number of contacts in a self-avoiding walk, if the $k^{-\text{th}}$ H in a sequence is aligned with the $i^{-\text{th}}$ row of the alignment graph. Then

$$2x_{ik} \geq t_{ik}; \tag{4}$$

$$\sum_{j \in \delta(i)} \sum_{k \in B'} x_{jk} \geq t_{ik} \quad i = 1, m^2. \tag{5}$$

Finally, the objective function to be maximized is:

$$z_{ip} = \max \sum_{i \in B} \sum_{k \in B'} t_{ik}. \tag{6}$$

Remark. Equation (4) is a bit simplified to avoid tedious notations. More precisely, if $k = 1$ or $k = l$ and $L(v_k) = H$ the multiplier is 3 instead of 2. In a 3D lattice the multiplier is 4 and in a graph, the node degree minus 2.

Let z_{lp} be the optimal value of the linear programming relaxation of the mixed integer programming problem defined by (6) under constraints (1–4) and x binary. Then the following is true:

Proposition. $z_{lp} \leq \text{UB}$.

P r o o f. If x^*, t^* is the optimal LP solution then by summing on (4) we obtain the result.

3. Computational results. We present here computational results based on the benchmarks used in [2] to demonstrate the quality of LP bounds provided by their model (in fact several models aiming at approaching UB). In Table 2, we emphasized how easily these bounds are attained and not their values because due to the Proposition they are always equal to LB. The first column indicates the problems we used, where H_n (P_n) means n successive H amino-acids (P amino-acids).

Table 2. Computational results for the benchmark problems

Sequence	length	non-zeros	time
$HPHP_2H_2PHP_2HPH_2P_2HPH$	20	14144	0.1
$H_2P_2HP_2HP_2HP_2HP_2HP_2HP_2H_2$	24	20214	0.7
$P_2HP_2H_2P_4H_2P_4H_2P_4H_2$	25	23000	0.8
$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$	36	65856	2.15
$P_2HP_2H_2P_2H_2P_5H_{10}P_6H_2P_2H_2P_2HP_2H_5$	48	146315	4
$H_2PHPHPHPH_4PHP_3HP_3HP_4HP_3HP_3HPH_4PHPHPHPH_2$	50	170940	10
$P_2H_3PH_8P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$	60	513792	128
$H_{12}PHPHP_2H_2P_2H_2P_2HP_2H_2P_2H_2P_2HP_2H_2P_2H_2P_2HPHPH_{12}$	64	791120	157

The length is in the second column, the number of non-zero elements is given to demonstrate the size of the respective LP problem, and the time (in seconds) to solve it is in the last column. All experiments were run on a Server IBM, 2X Quad-Core, 4C, 2.26 GHz, 2×2 GB. We used CPLEX as optimization software. From table 2, we can find that the proposed model is a good candidate to use in a dedicated branch-and-bound (cut) algorithm at least for problem with lengths up to 100. Small problems (like the first one) could be solved by a direct call to "mipopt" in CPLEX. For this problem, the lower bound 9 is found in more than 10 hours in [2] with Xpress MP2007a software on AMD Athlon 64 X2 Dual Core (2.70 GHz) with 2 GB Ram. With our model, this bound was found in 8 sec. and proven to be optimal in 525 sec. An optimal solution is shown in fig. 1.

For the last demonstration of the comparative efficiency of the proposed model we run to optimality the 12H problem as one with a big duality gap. In [2] the bound UB = 13 was reduced to 6 in 9429 sec. by requiring integrality for only some of the integer variables. In our model this value is proven to be optimal in 17 sec.

Note. In all runs the parameter m was set to $2\sqrt{l}$.

4. Conclusions. In this research, we suggested a mathematical formulation of the HP model for the PSP problem, which can be used in lattices

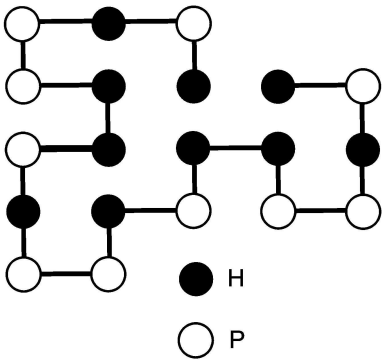

Fig. 1. Optimal solution for $HPHP_2H_2PHP_2HPH_2P_2HPH$

of various kinds, including arbitrary graphs. Our mathematical formulation provides the best known upper bound on the optimal value of the HP model when used as LP relaxation. The size of instances created in respect to the number of variables is $O(lm^2)$, which allows using general purpose mixed integer programming solvers not for toy problems only but also for a large class of real problems.

REFERENCES

[1] Andonov R., N. Malod-Dognin, N Yanev. Maximum Contact Map Overlap Revisited. *Journal of Computational Biology*, **18** (2011), No 1, 27–41.

[2] Ahn N., S. Park. Finding an Upper Bound for the Number of Contacts in Hydrophobic- Hydrophilic Protein Structure Prediction Model. *Journal of computational biology*, **4** (2010), No 17, 647–656.

[3] Hart W. E., S. Istrail. Fast protein folding in the Hydrophobic-Hydrophilic model within three- eighths of optimal. *Journal of Computational Biology*, **3** (1996), No 53, 157–168.

[4] Dill K. A. Theory for the folding and stability of globular proteins. *Biochemistry*, **24** (1985), No 6, 1501–1509.

[5] Istrail S., F. Lam. Combinatorial algorithms for protein folding in lattice models: A survey of mathematical results. *Commun. Inf. Syst.*, **9** (2009), 303–346.

[6] Dill K. A., S. B. Ozkan, M. S. Shell, T. R. Weikl. The protein folding problem. *Annu. Rev. Biophys*, **37** (2008), 289–316.

N. Yanev
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia, Bulgaria
e-mail: choby@math.bas.bg

Peter Milanov
Department of Informatics
Faculty of Mathematics and Natural Sciences
South-West University
66, Ivan Mihailov Str.
2700 Blagoevgrad, Bulgaria
and
Department of Operational Research
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia, Bulgaria

I. Mirchev
Faculty of Mathematics and Natural Sciences
South-West University
66, Ivan Mihailov Str.
2700 Blagoevgrad, Bulgaria
e-mail: mirchev@swu.bg

Received December 5, 2011
Final Accepted February 1, 2012