Optimally Compact Finite Sphere Packings — Hydrophobic Cores in the FCC

Rolf Backofen and Sebastian Will*

Institut für Informatik, LMU München
Oettingenstraße 67, D-80538 München
{backofen,wills}@informatik.uni-muenchen.de

Abstract. Lattice protein models are used for hierarchical approaches to protein structure prediction, as well as for investigating principles of protein folding. The problem is that there is so far no known lattice that can model real protein conformations with good quality, *and* for which there is an efficient method to prove whether a conformation found by some heuristic algorithm is optimal. We present such a method for the FCC-HP-Model [3]. For the FCC-HP-Model, we need to find conformations with a maximally compact hydrophobic core. Our method allows us to enumerate maximally compact hydrophobic cores for sufficiently great number of hydrophobic amino-acids. We have used our method to prove the optimality of heuristically predicted structures for HP-sequences in the FCC-HP-model.

1 Introduction

The protein structure prediction is one of the most important unsolved problems of computational biology. It can be specified as follows: Given a protein by its sequence of amino acids, what is its native structure? NP-completeness of the problem has been proven for many different models (including lattice and off-lattice models) [8, 9]. These results strongly suggest that the protein folding problem is NP-hard in general. Therefore, it is unlikely that a general, efficient algorithm for solving this problem can be given. Actually, the situation is even worse, since the general principles why natural proteins fold into a native structure are unknown. This is cumbersome since rational design is commonly viewed to be of paramount importance e.g. for drug design, where one faces the difficulty to design proteins that have a unique and stable native structure.

To tackle structure prediction and related problems simplified models have been introduced. They are used in hierarchical approaches for protein folding (e.g., [21], see also the meeting review of CASP3 [15], where some groups have used lattice models). Furthermore, they have became a major tool for investigating general properties of protein folding.

Most important are the so-called lattice models. The simplifications commonly used in this class of models are: 1) monomers (or residues) are represented

* Supported by the PhD programme "Graduiertenkolleg Logik in der Informatik" (GKLI) of the "Deutsche Forschungsgemeinschaft" (DFG).

using a unified size 2) bond length is unified 3) the positions of the monomers are restricted to lattice positions and 4) a simplified energy function.

In the literature, many different lattice models (i.e., lattices and energy functions) have been used. Examples of how such models can be used for predicting the native structure or for investigating principles of protein folding were given [20, 1, 11, 19, 12, 2, 17, 21]. Of course, the question arises which lattice and energy functions has to be preferred. There are two (somewhat conflicting) aspects that have to be evaluated when choosing a model: 1) the accuracy of the lattice in approximating real protein conformations, and the ability of the energy function to discriminate native from non-native conformations, and 2) the availability and quality of search algorithm for finding minimal (or nearly minimal) energy conformations.
While the first aspect is well-investigated in the literature (e.g., [18, 10]), the second aspect is underrepresented. By and large, there are mainly two different heuristic search approaches used in the literature: 1) Ad hoc restriction of the search space to compact or quasi-compact conformations (a good example is [20], where the search space is restricted to conformations forming an $n \times n \times n$-cube). The main drawback here is that the restriction to compact conformation is not biologically motivated for a complete amino acid sequence (as done in these approaches), but only for the hydrophobic amino acids. In consequence, the restriction either has to be relaxed and then leads to an inefficient algorithm or is chosen to strong and thus may exclude optimal conformations. 2.) Stochastic sampling like Monte Carlo methods with simulated annealing, genetic algorithms etc. Here, the degree of optimality for the best conformations and the quality of the sampling cannot be determined by state of the art methods.[1]

On the other hand, there are only three exact algorithms known [23, 4, 6] which are able to enumerate minimal (or nearly minimal) energy conformations, all for the cubic lattice. However, the ability of this lattice to approximate real protein conformations is poor. For example, [3] pointed out especially the parity problem in the cubic lattice. This drawback of the cubic lattice is that every two monomers with chain positions of the same parity cannot form a contact.

In this paper, we follow the proposal by [3] to use a lattice model with a simple energy function, namely the HP (hydrophobic-polar) model, but on a better suited lattice (namely the face-centered cubic). There are two reasons for this approach:
1) The FCC can model real protein conformations with good quality (see [18], where it was shown that FCC can model protein conformations with coordinate root mean square deviation below 2 Å)
2) The HP-model models the important aspect of hydrophobicity. Essentially it is a polymer chain representation (on a lattice) with one stabilizing interaction each time two hydrophobic residues have unit distance. This enforces compactification while polar residues and solvent is not explicitly regarded. It follows the

[1] Despite there are mathematical treatments of Monte Carlo methods with simulated annealing, the partition function of the ensemble (which is needed for a precise statement) is in general unknown.

assumption that the hydrophobic effect determines the overall configuration of a protein (for a definition of the HP-model, see [16, 10]).

Once a search algorithm for minimal energy conformations is established for this *FCC-HP-model*, one can employ it as a filter step in an hierarchical approach. This way, one can improve the energy function to achieve better biological relevance and go on to resemble amino acid positions more accurately.

Contribution of the paper In this paper, we present the first algorithm for enumerating maximal compact hydrophobic cores in the face-centered cubic lattice. For a given conformation of the FCC-HP-model, the hydrophobic core is the set of all positions occupied by hydrophobic (H) monomers. A hydrophobic core is maximally compact if the number of contacts between neighbored positions is maximized. Thus, a conformation which has a maximally compact hydrophobic core has minimal energy in the HP-model.

There are mainly two applications of the algorithm for finding hydrophobic cores. The first is that it provides a method to check minimality of conformations found by an heuristic algorithm. We have used an heuristic algorithm described earlier [7]. For the first time, we were able to find minimal energy conformations (and to prove their optimality) for HP-sequences in the FCC-HP-model. So far, the only known results for the FCC-HP-models were approximation results with an guaranteed ratio of 60% ([3], [13] provides a general approximation scheme for HP-models on arbitrary lattices; [14] gives an approximation scheme for the HP-model on the cubic lattice).

The second application is that the hydrophobic cores are a promising intermediate step for an algorithm to enumerate *all* minimal energy conformations. This technique has already been used successfully in [23].

2 Preliminaries

For a vector $\boldsymbol{p}$, we denote with $\boldsymbol{p}_x$ (resp. $\boldsymbol{p}_y$ or $\boldsymbol{p}_z$) its x-coordinate (resp. y- or z-coordinate). We use a transformed representation of the FCC-lattice (for a detailed description, see [5]. We define the FCC-isomorphic lattice D_3' to be the lattice that consists of the following sets of points: $D_3' = \{\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mid \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{Z}^3 \text{ and } x \text{ even}\} \uplus \{\begin{pmatrix} x \\ y+0.5 \\ z+0.5 \end{pmatrix} \mid \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{Z}^3 \text{ and } x \text{ odd}\}$. The first set consist of the points in even x-layers, the second of the points in odd x-layers. The set $N_{D_3'}$ of *minimal vectors* connecting neighbors in D_3' is given by $N_{D_3'} = \left\{ \begin{pmatrix} 0 \\ \pm 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \pm 1 \end{pmatrix} \right\} \uplus \left\{ \begin{pmatrix} \pm 1 \\ \pm 0.5 \\ \pm 0.5 \end{pmatrix} \right\}$. The vectors in the second set are the vectors connecting neighbors in two successive x-layers. Two points $\boldsymbol{p}$ and $\boldsymbol{p'}$ in D_3' are *neighbors* if $\boldsymbol{p} - \boldsymbol{p'} \in N_{D_3'}$.

A *coloring* is a function $f : D_3' \to \{0, 1\}$, where $f^{-1}(1) \neq \emptyset$. We will identify a coloring f with the set of all *points colored by* f, i.e. $\{\boldsymbol{p} \mid f(\boldsymbol{p}) = 1\}$. Hence, for colorings f_1, f_2 we will use standard set notation for size $|f_1|$, union $f_1 \cup f_2$, disjoint union $f_1 \uplus f_2$, and intersection $f_1 \cap f_2$. Given a coloring f, we define the *number of contacts of* f by $\mathrm{con}(f) := \frac{1}{2} \left| \{(\boldsymbol{p}, \boldsymbol{p'}) \mid f(\boldsymbol{p}) \wedge f(\boldsymbol{p'}) \wedge (\boldsymbol{p} - \boldsymbol{p'}) \in N_{D_3'}\} \right|$.

A coloring f is called a *coloring of the plane* $x = c$ if $f(x, y, z) = 1$ implies $x = c$. We say that f is a *plane coloring* if there is a c such that f is a coloring of plane $x = c$. We define $\mathrm{Surf}_{pl}(f)$ to be the surface of f in the plane $x = c$, i.e. $\mathrm{Surf}_{pl}(f) = |\{(\boldsymbol{p}, \boldsymbol{p'}) \mid (\boldsymbol{p} - \boldsymbol{p'}) \in N_{D_3'} \wedge f(\boldsymbol{p}) \wedge \neg f(\boldsymbol{p'}) \wedge \boldsymbol{p'}_x = c\}$. With $\min_x(f)$ we denote the integer $\min\{\boldsymbol{p}_x \mid \boldsymbol{p} \in f\}$. $\max_x(f)$, $\min_y(f)$, $\max_y(f)$, $\min_z(f)$ and $\max_z(f)$ are defined analogously.

3 Description of the Method

Our aim is to determine maximally compact hydrophobic cores. A *hydrophobic core* is just a coloring f. A *maximally compact* hydrophobic core for n points is a coloring f of n points that maximizes $\mathrm{con}(f)$. Without loss of generality, we can assume that $\min_x(f) = 1$. Let $k = \max_x(f)$. Then, we partition f into plane colorings $f_1, \ldots, f_k$ of the layers $x = 1, \ldots, x = k$. For searching a maximal coloring f, we do a branch-and-bound search on k and $f_1 \ldots f_k$.

Of course, the problem is to give good bounds that allow us to cut off many k and $f_1 \ldots f_k$ that will not maximize $\mathrm{con}(f_1 \uplus \ldots \uplus f_k)$. For this purpose, we distinguish between contacts in a single layer ($= \mathrm{con}(f_i)$ for $1 \leq i \leq k$), and interlayer contacts $\mathrm{IC}_{f_i}^{f_{i+1}}$ for $1 \leq i < k$ between two successive layers (i.e., pairs $(\boldsymbol{p}, \boldsymbol{p'})$ such that $\boldsymbol{p}$ and $\boldsymbol{p'}$ are neighbors, $\boldsymbol{p} \in f_i$ and $\boldsymbol{p'} \in f_{i+1}$). We then give two different bounds on the layer and interlayer contacts, provided some parameters restricting the f_i's.

For every plane coloring f_i, these parameters are the size n_i of f_i, the number a_i of rows that contain a point of f_i, and the number b_i of columns that contain a point of f_i. Given these parameters, it is known [23] that the layer contacts of f_i are given by $2n_i - a_i - b_i$. In this paper, we present for *any set of parameters* n_i, a_i, b_i and n_{i+1} an upper bound on the number of interlayer contacts $B_{n_i,a_i,b_i}^{n_{i+1}} \geq \max \left\{ \mathrm{IC}_{f_i}^{f_{i+1}} \middle| \begin{array}{l} f_i \text{ satisfies } n_i, a_i, b_i \\ \text{and } |f_{i+1}| = n_{i+1}. \end{array} \right\}$

So far, the only related bound was given in our own work [5]. Although there a bound $B_{n_i,a_i,b_i}^{n_{i+1}}$ was given, this bound does not hold for arbitrary sets of parameters n_i, a_i, b_i and n_{i+1}. Instead, the bound is valid for sufficiently filled plane colorings (called normal), which was sufficient for the purpose of [5].

The bound $B_{n_i,a_i,b_i}^{n_{i+1}}$ is used in searching for a maximally compact core for n H-monomers as follows. Instead of directly enumerating k and all possible colorings $f_1 \uplus \ldots \uplus f_k$, we search through all possible sequences of parameters $((n_1, a_1, b_1) \ldots (n_k, a_k, b_k))$ with the property that $n = \sum_i n_i$. By using the $B_{n_i,a_i,b_i}^{n_{i+1}}$, only a few layer sequences have to be considered further. For these optimal layer sequences, we then search for all admissible colorings $f_1 \uplus \ldots \uplus f_k$.

For calculating the bound $B_{n_i,a_i,b_i}^{n_{i+1}}$, we need to introduce additional parameters, namely the number of non-overlapping and unconnected rows in layer $x = i$. These additional parameters allow us to determine the maximal number of interlayer contacts between layer $x = i$ and $x = i + 1$. Further note that only few combinations of (n_i, a_i, b_i) and these additional parameters are admis-

sible. Thus, for every (n_i, a_i, b_i), we search through all admissible numbers of non-overlapping rows in layer $x = i$ to determine $B^{n_{i+1}}_{n_i,a_i,b_i}$.

In Section 4, we define the parameters of a plane coloring and determine which combinations of parameters are admissible. In Section 5, the number of interlayer contacts is given provided the parameters and the number of points with three interlayer contacts, called 3-points, is fixed. In the following section, we determine the number of 3-points that maximizes the interlayer contacts.

4 Properties of Overlapping and Non-overlapping Colorings

Let f be a coloring of plane $x = c$. A *horizontal caveat in* f is a k-tuple of points $(\boldsymbol{p_1}, \ldots, \boldsymbol{p_k})$ such that $\forall 1 \leq j < k : ((\boldsymbol{p_{j+1}} - \boldsymbol{p_j})_y = 1)$, $\{\boldsymbol{p}_1, \boldsymbol{p}_k\} \in f$ and $\forall 1 < j < k : \boldsymbol{p}_j \notin f$. A *vertical caveat in* f is defined analogously satisfying $\forall 1 \leq j < k : ((\boldsymbol{p_{j+1}} - \boldsymbol{p_j})_z = 1)$ instead. We say that f contains a *caveat* if there is at least one horizontal or vertical caveat in f. f is called *caveat-free* if it does not contain a caveat. We will handle only caveat-free colorings. The methods can be extended to treat caveats as well, but we suppress them for simplicity.

We now introduce the parameters of a plane coloring f that will allows us to determine layer *and* to bound interlayer contacts. The first set of parameters are the rows and columns occupied by f. For an arbitrary plane coloring f of $x = c$ define $\mathrm{occz}(f, z) := \exists y : f(c, y, z)$ and $\mathrm{occy}(f, y) := \exists z : f(c, y, z)$. Furthermore, we define $\mathrm{oylines}(f) := \left|\left\{\, y \middle| \mathrm{occy}(f, y) \,\right\}\right|$ and $\mathrm{ozlines}(f) := \left|\left\{\, z \middle| \mathrm{occz}(f, z) \,\right\}\right|$. For notational convenience define $\mathrm{olines}(f) := (\mathrm{oylines}(f), \mathrm{ozlines}(f))$. For a coloring f, we call rows z, where $\mathrm{occz}(f, z)$ holds, and columns y, where $\mathrm{occy}(f, y)$, *occupied*, and *unoccupied* otherwise.

For a plane coloring f, we define the *layer contacts* LC_f to be $\mathrm{con}(f)$. We define

$$\mathrm{LC}_{n,a,b} := \max \left\{ \mathrm{LC}_f \middle| \begin{array}{l} f \text{ is a coloring of plane } x = c \\ \wedge f \text{ has lines } (a, b) \wedge |f| = n \end{array} \right\}.$$

Proposition 1. *For every caveat-free coloring f with* $\mathrm{olines}(f) = (a, b)$*, we get* $\mathrm{LC}_{n,a,b} = 2n - \frac{1}{2}\mathrm{Surf}_{pl}(f)$ *and* $\mathrm{Surf}_{pl}(f) = 2(a + b)$.

Proof (sketch). Each of the n points colored by f has 4 neighbors, which are either occupied by another point, or by a surface point. Hence, $4n = 2\mathrm{LC}_{n,a,b} + \mathrm{Surf}_{pl}(f)$. For the second claim, note that by definition, every occupied row and column must generate 2 surface contacts, and, by caveat-free, there can be no more than 2.

The second set of parameters are the number of unconnected and non-overlapping rows. Let f be a coloring of plane $x = c$. We define a row z to be *non-overlapping in* f if z is occupied, there is an occupied row $z' > z$, and there is no y such that $f(c, y, z) \wedge f(c, y, z + 1)$. A row z is called *unconnected* if it is non-overlapping and not $\exists y, y' : f(c, y, z) \wedge f(c, y', z + 1) \wedge |y - y'| \leq 1$.

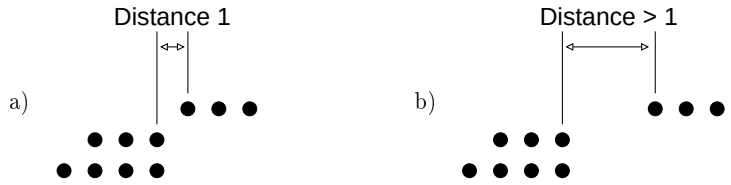


Fig. 1. a) Non-overlapping vs. b) unconnected

The number of non-overlapping rows is denoted by #non-overlaps(f) and the number of unconnected rows by #non-connects(f).

To illustrate the terms, Figure 1a) shows a coloring with #non-overlaps(f) = 1 and #non-connects(f) = 0, whereas the coloring in Figure 1b) satisfies that #non-overlaps(f) = 1 and #non-connects(f) = 1.

We will call a coloring f with #non-overlaps(f) = 0 *overlapping* (otherwise *non-overlapping*). A coloring with #non-connects(f) = 0 is called *connected* (otherwise *unconnected*).

In the rest of this section, we give precise bounds on the number of colored points, given the parameters of the plane coloring. We will first state some properties of colorings with respect to olines(f), #non-overlaps(f) and #non-connects(f).

Proposition 2. *For every caveat-free coloring f, we have* $|f| \geq \max(\text{olines}(f))$.

Since by definition the maximal occupied row z can not be non-overlapping we immediately get that #non-overlaps(f) is less than oylines(f). The next lemma states in addition that #non-overlaps(f) is less than ozlines(f). Intuitively, this is a consequence of the (non-trivial) fact that every non-overlapping row produces exactly one non-overlapping column.

Lemma 1. *For a caveat-free coloring f, we get*

$$\#\text{non-overlaps}(f) < \min(\text{olines}(f)).$$

A caveat-free coloring can be split at non-overlapping rows into sub-colorings with the nice property that the parameters of the coloring can be calculated from the sub-colorings in a simple way. This fact will be employed for inductive arguments. Given a plane coloring f and a row $\min_z(f) \leq z_s < \max_z(f)$, we define $f_{\theta z_s} = \{(c, y, z) \in f \mid z \theta z_s\}$ for $\theta \in \{\leq, >\}$. Note that the restriction on z_s is required, since splitting at row $z_s = \max_z(f)$ would produce an empty sub-coloring $f_{>z_s}$. Further note that this restriction is trivially satisfied by any non-overlapping row.

Lemma 2 (Split). *Let f be a caveat-free coloring of the plane $x = c$ with* #non-overlaps(f) ≥ 1, *and let z_s be a non-overlapping row. Then,*

1. $f = f_{\leq z_s} \uplus f_{>z_s}$ *and the sub-colorings $f_{\leq z_s}$ and $f_{>z_s}$ are caveat-free*
2. $\text{olines}(f) = (\text{oylines}(f_{\leq z_s}) + \text{oylines}(f_{>z_s}), \text{ozlines}(f_{\leq z_s}) + \text{ozlines}(f_{>z_s}))$
3. $\#\text{non-overlaps}(f) = \#\text{non-overlaps}(f_{\leq z_s}) + \#\text{non-overlaps}(f_{>z_s}) + 1$.

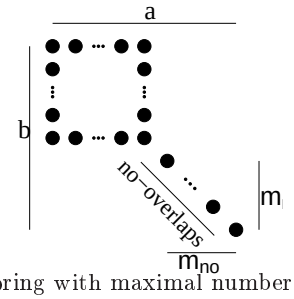


Fig. 2. Coloring with maximal number of elements.

There is a dependency of the admissible numbers #non-overlaps(f) and the number of elements in a coloring f, given the number of occupied lines in y and z direction. Think of (a, b) (resp. $\mathrm{m_{no}}$) as representing olines(f) (resp. #non-overlaps(f)). We define $n_{\max}(a, b, \mathrm{m_{no}}) := \mathrm{m_{no}} + (a - \mathrm{m_{no}})(b - \mathrm{m_{no}})$ and $n_{\min}(a, b, \mathrm{m_{no}}) := a + b - 1 - \mathrm{m_{no}}$. The idea of the definition of $n_{\max}(a, b, \mathrm{m_{no}})$ is that the number of elements is maximized if we have one big overlapping region and waste as little space as possible for the non-overlapping region. Hence, in this maximal coloring, all of the non-overlapping rows contain exactly one point. Such a coloring is shown in Figure 2.

Lemma 3. *All caveat-free colorings f satisfy $|f| \leq n_{\max}(a, b, \mathrm{m_{no}})$, where $\mathrm{m_{no}} =$ #non-overlaps(f) and $(a, b) =$ olines(f).*

Lemma 4. *For all caveat-free colorings f holds $n_{\min}(a, b, \mathrm{m_{no}}) \leq |f|$, where $(a, b) =$ olines(f) and $\mathrm{m_{no}} =$ #non-overlaps(f).*

Proof (sketch). For the case $\mathrm{m_{no}} = 0$, a coloring f of plane $x = c$ with minimal number of points and olines(f) $= (a, b)$ is given by the coloring that has b points $(c, 1, 1) \ldots (c, 1, b)$ in the column $y = 1$, and a points $(c, 1, b) \ldots (c, a, b)$ in the row $z = b$. Clearly, f has $a + b - 1$ points since $(c, 1, b)$ is in the first column *and* last row. For $\mathrm{m_{no}} > 0$, the claim follows by induction using the split lemma 2, Claim 3.

For convenience, we define the following bounds on the number of non-overlapping rows:

$$\mathrm{no_{min}}(n, a, b) := \min\{\mathrm{m_{no}} \mid 0 \leq \mathrm{m_{no}} \leq \min(a, b) - 1 \wedge n \geq n_{\min}(a, b, \mathrm{m_{no}})\}$$
$$\mathrm{no_{max}}(n, a, b) := \max\{\mathrm{m_{no}} \mid 0 \leq \mathrm{m_{no}} \leq \min(a, b) - 1 \wedge n \leq n_{\max}(a, b, \mathrm{m_{no}})\}$$

Proposition 3. *For any caveat-free coloring f with* olines(f) $= (a, b)$ *and* $|f| = n$ *holds* $\mathrm{no_{min}}(n, a, b) \leq$ #non-overlaps(f) $\leq \mathrm{no_{max}}(n, a, b)$.

5 Number of i-points for caveat-free colorings

In the next two sections, we will provide a bound on interlayer contacts. For this purpose, we calculate for a coloring f of plane c the numbers of points having

4,3,2, and 1 contacts to f (in the following called i-points). Theorem 1 will state that we can achieve the maximal number of interlayer contacts between $x = c$ and $x = c + 1$ if we fill the 4-points first, then (if points are left) the 3-points and so on. Before, we need some definitions and auxiliary lemmata.

In the following, let f be a plane coloring of plane $x = c$ and f' a plane coloring of plane $x = c'$, where $c \neq c'$. We define the *number of interlayer contacts of f and f'* by $\mathrm{IC}_f^{f'} = \mathrm{con}(f \uplus f') - \mathrm{LC}_f - \mathrm{LC}_{f'}$. We define $\mathrm{contacts}_{\max}(f, n)$ as

$$\max \left\{ \mathrm{IC}_f^{f'} \,\middle|\, f' \text{ is a plane coloring of } x = c + 1 \text{ with } |f'| = n \right\}.$$

A point $\boldsymbol{p}$ is called a 4-*point for* f if $\boldsymbol{p}$ is in plane $x = c+1$ or $x = c-1$ and $\boldsymbol{p}$ has 4 neighbors $\boldsymbol{p_1}, \ldots, \boldsymbol{p_4} \in f$. Analogously, we define 3-points, 2-points and 1-points. Furthermore, we define $\#4_{c-1}(f) = |\{\boldsymbol{p} \mid \boldsymbol{p} \text{ 4-point for } f \text{ in } x = c - 1\}|$. Analogously, we define $\#4_{c+1}(f)$ and $\#i_{c\pm1}(f)$ for $i = 1, 2, 3$. We will show that the number of i-points for every $i \in \{1, 2, 3, 4\}$ depend only on the number of non-overlaps, the number of non-connects, and the number of x-steps. An *x-step for a plane coloring* f is a triple $(\boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3})$ such that $f(\boldsymbol{p_1}) = 0$, $f(\boldsymbol{p_2}) = 1 = f(\boldsymbol{p_3})$, $\boldsymbol{p_1} - \boldsymbol{p_2} = \pm \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ and $\boldsymbol{p_1} - \boldsymbol{p_3} = \pm \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$. With $\mathrm{xsteps}(f)$ we denote the number of x-steps of f. Now we can define the number of i-points, depending on $n = |f|$, $s = \mathrm{Surf}_{pl}(f)$, $m_x = \mathrm{xsteps}(f)$, $\mathrm{m_{no}} = \#\text{non-overlaps}(f)$ and $\mathrm{m_{nc}} = \#\text{non-connects}(f)$:

$$\begin{aligned}
\#4\binom{n,s}{\mathrm{m_{no}},\mathrm{m_{nc}},m_x} &= n - \frac{1}{2}s + 1 + \mathrm{m_{no}} \\
\#2\binom{n,s}{\mathrm{m_{no}},\mathrm{m_{nc}},m_x} &= s - 4 - 2\#3\binom{n,s}{\mathrm{m_{no}},\mathrm{m_{nc}},m_x} - 3\,\mathrm{m_{no}} - \mathrm{m_{nc}} \\
\#3\binom{n,s}{\mathrm{m_{no}},\mathrm{m_{nc}},m_x} &= m_x - 2(\mathrm{m_{no}} - \mathrm{m_{nc}}) \\
\#1\binom{n,s}{\mathrm{m_{no}},\mathrm{m_{nc}},m_x} &= \#3\binom{n,s}{\mathrm{m_{no}},\mathrm{m_{nc}},m_x} + 2\,\mathrm{m_{no}} + 2\,\mathrm{m_{nc}} + 4
\end{aligned}$$

For preparation, we state two lemmas that investigate how to calculate the i-points of f from the two sub-colorings generated by splitting f at a non-overlapping or unconnected row.

Lemma 5 (Split 3-Points). *Let f be a caveat-free coloring of plane $x = c$ with* $\#\text{non-overlaps}(f) \geq 1$*, and let z_s be a non-overlapping row. Then,* $\#3(f) = \#3(f_{\leq z_s}) + \#3(f_{>z_s})$.

Proof (sketch). We can show that neither $f_{\leq z_s}$, nor $f_{>z_s}$, nor f has a 3-point that lies between rows z_s and $z_s + 1$. This implies that every 3-point for f is either below z_s and is therefore also a 3-point for $f_{\leq z_s}$, or above $z_s + 1$ and is therefore also a 3-point for $f_{>z_s}$.

Lemma 6 (Split at minimal unconnected row). *Let f be a caveat-free coloring of plane $x = c$ with* $\#\text{non-connects}(f) \geq 1$*, and let z_s be the minimal unconnected row. Then,* $\#\text{non-connects}(f_{\leq z_s}) = 0$, $\#\text{non-connects}(f_{>z_s}) =$

$\#\text{non-connects}(f) - 1$ *and*

$$\text{xsteps}(f_{\leq z_s}) + \text{xsteps}(f_{>z_s}) = \text{xsteps}(f), \tag{1}$$

$$\forall i \in \{1,2,3,4\}: \ \#i(f_{\leq z_s}) + \#i(f_{>z_s}) = \#i(f). \tag{2}$$

Proof (sketch). The first two claims are trivial. For claims (1) and (2), one shows that if z_s is unconnected, the y-distance between points in $f_{\leq z_s}$ and $f_{>z_s}$ is always greater than 1. This implies that the sets of i-points and x-steps of $f_{\leq z_s}$ and $f_{>z_s}$ are disjoint.

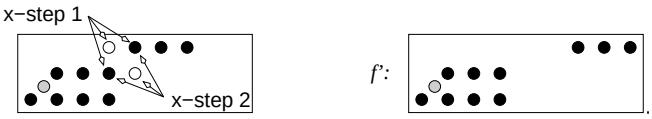
Lemma 7. *Let f be a caveat-free coloring. Then*

$$\forall i \in \{1,2,3,4\}: \#i(f) = \#i\binom{|f|, \text{Surf}_{pl}(f)}{\#\text{non-overlaps}(f), \#\text{non-connects}(f), \text{xsteps}(f)}.$$

Proof (sketch). The case $\#\text{non-connects}(f) = 0$ is equivalent to the formula already proven in [5]. For the case $\#\text{non-connects}(f) = \text{m}_{\text{nc}} > 0$, we do induction on m_{nc}. The claim for $\#4(f)$, $\#3(f)$ and $\#1(f)$ follow from the Split-Lemmata 2, 5 and 6 by simple calculation (recall that by definition, every unconnected line is also non-overlapping). For $\#2(f)$, the claim follows by simple calculation from the equation $4\#4(f) + 3\#3(f) + 2\#2(f) + 1\#1(f) = 4|f|$. This equation holds since the sum of all interlayer contacts between f and the next plane is $4|f|$.

6 Maximal number of 3-points.

Due to the last lemma, if we consider colorings with given $n, a, b, \text{m}_{\text{no}}$, and m_{nc}, then m_x does not affect the number of 4-points, but increases the number of 3-points and 1-points, while decreasing the number of 2-points. The increase of 3- and 1-points is 1 per x-step, the decrease of 2-points is 2 per 3-point. This pattern grants that we maximize the possible number of interlayer contacts to a second plane with a given number of elements, if we maximize the number of 3-points in the first plane. For this purpose, we first show that we need not to distinguish between unconnected and non-overlapping rows for the number of 3-points. The reason is that number of 3-points does not change if one transforms a non-overlapping row into into a unconnected row. Consider as an example the two colorings



Then both f and f' have one 3-point (indicated in grey). By transforming the non-overlapping row in f into a unconnected row, f' looses two x-steps. Thus, the effects of increasing $\#\text{non-connects}(\cdot)$ by 1 are diminished by decreasing $\text{xsteps}(\cdot)$ by 2.

Note that such a bound for the interlayer contacts using a bound for 3-points that does not distinguish between non-overlapping and unconnected rows slightly overestimates, since we assume the best case for the number of 2- and 1-points (note that in contrast to the number of 3-points, the number of 2- and 1-points depend on the exact number of unconnected rows).

We start with the extension of the bound for 3-points, as given in [5] in the case of "sufficiently filled" and overlapping colorings, to arbitrary overlapping colorings. We need to recall some definitions from [5]. For an overlapping coloring f with olines$(f) = (a, b)$, a and b are the side lengths of the minimal rectangle around the points in f (called frame(f) in the following). The detailed frame of a coloring f is the tuple $(a, b, i_{lb}, i_{lu}, i_{rb}, i_{ru})$, where (a, b) is the frame of f and i_{lb} is the number of diagonals that can be drawn from the left-bottom corner. i_{lu}, i_{rb}, i_{ru} are defined analogously. For a coloring f with detailed frame $(a, b, i_{lb}, i_{lu}, i_{rb}, i_{ru})$, we call $\boldsymbol{i} = (i_{lb}, i_{lu}, i_{rb}, i_{ru})$ the *indent vector of* f. As shown in [5], the indent vector gives a precise bound on the $\#3(f)$, since in this case, xsteps$(f) = \#3(f)$ and xsteps$(f) = i_{lb} + i_{lu} + i_{rb} + i_{ru} -$ diagcav(f). Here, diagcav(f) counts the number of *diagonal caveats*, which are defined analogous to vertical and horizontal caveats. For example, consider the plane coloring f_{ex} as given in Figure 6. Then the detailed frame of f_{ex} is $(6, 9, 3, 2, 1, 2)$. The number of 3-points (indicated by $\times$) for f_{ex} is $8 = 3 + 2 + 1 + 2$, since f_{ex} does not contain diagonal caveats.

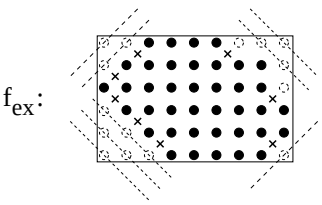


Fig. 3. Detailled Frame

In the overlapping case, we search for a given number of points n and a frame (a, b) the maximal number of x-steps. For this purpose, we define for some indent vector $\boldsymbol{i} = (i_1, i_2, i_3, i_4)$, vol$(a, b, \boldsymbol{i}) := ab - \sum_{1 \leq j \leq 4} \frac{i_j (i_j + 1)}{2}$. vol$(a, b, \boldsymbol{i})$ is the maximal number of points that can be colored by any f that has indent vector $\boldsymbol{i}$ and frame (a, b). $\boldsymbol{i} = (i_1, i_2, i_3, i_4)$ is called *maximal* for (a, b) iff $\sum_{1 \leq j \leq 4} i_j = 2(\min(a, b) - 1)$. For example, if $b \leq a$, then the indent vector $\boldsymbol{i}$ is maximal for (a, b) if every coloring with frame (a, b) and indent vector $\boldsymbol{i}$ has exactly one colored point in the first and last column.

vol$(a, b, \boldsymbol{i})$ can now be used to calculate the maximal number of x-steps that can be achieved given n colored points and frame (a, b). The maximal number of x-steps is achieved if we make the indents as uniform as possible. For this purpose, define edge$(n, a, b) = \max\{k \in \mathbb{N} \mid \text{vol}(a, b, (k, k, k, k))\}$. $k =$ edge(n, a, b) defines the maximal possible uniform indent. Then $r = \text{ext}(n, a, b) = \lfloor \frac{ab - 4\frac{k(k+1)}{2} - n}{k+1} \rfloor$ defines the number of times r we can extend the uniform indent by 1. n is called *normal for* (a, b) if either $4k + r < 2(a - 1)$, or $4k + r = 2(a - 1)$ and $ab - 4\frac{k(k+1)}{2} - r(k + 1) = n$.

Now there are two upper bounds that can be given for the number of x-steps, given n colored points and frame (a, b). The first is given by the indent vector. The second by the fact, that in caveat-free and overlapping colorings, there may be at most between every two successive lines 2 x-steps, which gives at most

$2\min(a,b)-1$. Thus, the bound given in [5] is as follows:

$$\mathrm{xsteps}_{\mathrm{bnd}}(n,a,b) = \min(4\,\mathrm{edge}(n,a,b) + \mathrm{ext}(n,a,b), 2(\min(a,b)-1)).$$

We improve the bound in the case of quadratic frames (a,a) and n is not normal for (a,a). Here, we show that we have an upper bound of $2a-3$ instead of $2a-2$ if there is no maximal indent $\boldsymbol{i}$ with $n = \mathrm{vol}(a,a,\boldsymbol{i})$. We show in this case, that there must be a diagonal caveat.

Lemma 8. *For every overlapping caveat-free coloring f we get*

$$\#3(f) \le \#3_{\mathrm{bound}}(|f|,a,b),$$

where $(a,b) = \mathrm{frame}(f)$ *and*

$$\#3_{\mathrm{bound}}(n,a,b) := \begin{cases} \mathrm{xsteps}_{\mathrm{bnd}}(n,a,b) & n \textit{ is normal for frame } (a,b) \\ 2\min(a,b)-2 & \textit{else if } a \neq b \\ 2a-2 & \textit{else if } \exists \boldsymbol{i} : \boldsymbol{i} \textit{ are maximal indents} \\ & \qquad \textit{for } (a,a) \wedge n = \mathrm{vol}(a,a,\boldsymbol{i}) \\ 2a-3 & \text{otherwise} \end{cases}$$

For the general case of possibly non-overlapping colorings, Lemmata 3, 4, and 1 imply that any coloring f with $\mathrm{olines}(f) = (a,b)$ and $\#\text{non-overlaps}(f) = \mathrm{m_{no}}$ satisfies $\mathrm{valid}(n,a,b,\mathrm{m_{no}}) := \big(\mathrm{m_{no}} < \min(a,b) \wedge n_{\min}(a,b,\mathrm{m_{no}}) \le n \le n_{\max}(a,b,\mathrm{m_{no}})\big)$. Hence, we define $\#3_{\mathrm{bound}}(n,a,b,\mathrm{m_{no}})$ to be $-\infty$ in the case that $\mathrm{valid}(n,a,b,\mathrm{m_{no}})$ does *not* hold. Otherwise, we define $\#3_{\mathrm{bound}}(n,a,b,\mathrm{m_{no}})$ by $\#3_{\mathrm{bound}}(n,a,b)$ if $\mathrm{m_{no}} = 0$ and

$$\max\left\{ \begin{array}{c} \#3_{\mathrm{bound}}(n',a',b',0) \\ +\,\#3_{\mathrm{bound}}(n-n',a-a', \\ b-b',\mathrm{m_{no}}-1) \end{array} \middle| \begin{array}{l} 1 \le n' \le n-1, \\ 1 \le a' \le a-1, \\ 1 \le b' \le b-1, \end{array} \right\},$$

otherwise.

Lemma 9. *For every caveat-free coloring f, holds $\#3(f) \le \#3_{\mathrm{bound}}(n,a,b,\mathrm{m_{no}})$, where $n = |f|$, $(a,b) = \mathrm{olines}(f)$, and $\mathrm{m_{no}} = \#\text{non-overlaps}(f)$.*

Proof (sketch). The case $\mathrm{m_{no}} = 0$ is treated in Lemma 8. For n,a,b and $\mathrm{m_{no}} > 0$ with $\mathrm{valid}(n,a,b,\mathrm{m_{no}})$, we can split a coloring f at the minimal non-overlapping line z_s and into $f_{\le z_s}$ and $f_{>z_s}$ and get $\#3(f) = \#3(f_{\le z_s}) + \#3(f_{>z_s})$ by Lemma 5. Considering all possible rows for splitting will give the second case of $\#3_{\mathrm{bound}}(n,a,b,\mathrm{m_{no}})$.

The bound on the number of 3-points can now be used to derive a bound on the number of interlayer contacts for arbitrary colorings. Summarizing, we get the following bound:

$$\begin{aligned} \mathrm{BNMIC}^{n_2}_{n_1,a_1,b_1}(\mathrm{m_{no1}}) :=\ & 4\min(n_2,\#4) + 3\min(\#3,\max(n_2-\#4,0)) \\ & + 2\min(\#2,\max(n_2-\#4-\#3,0)) + \min(\#1,\max(n_2-\#4-\#3-\#2,0)) \end{aligned}$$

where $\#4 = n - a_1 - b_1 + 1 + \mathrm{m_{no1}}$ $\qquad \#3 = \#3_{\mathrm{bound}}(n_1,a_1,b_1,\mathrm{m_{no1}})$
$\#2 = 2(a_1+b_1) - 4 - 2\#3 - 3\,\mathrm{m_{no1}}$ $\qquad \#1 = \#3 + 2\,\mathrm{m_{no1}} + 4.$

Theorem 1. *Let f_1 and f_2 be coloring of planes $x = c$ and $x = c + 1$, respectively. Let $n_1 = |f_1|$,olines$(f_1) = (a_1, b_1)$, $|f_2| = n_2$ and* olines$(f_2) = (a_2, b_2)$. *Then* $\mathrm{IC}_{f_1}^{f_2} \leq \min(\mathrm{BNMIC}_{n_1,a_1,b_1}^{n_2}, \mathrm{BNMIC}_{n_2,a_2,b_2}^{n_1})$.

7 Constructing the Compact Cores

We will now show how to compute the optimally compact cores for a given number of elements, thereby employing the given bound on interlayer contacts, for a branch-and-bound approach. Due to space restrictions, we have to omit many details of the approach.

W.l.o.g, let a coloring f be decomposed into plane colorings $f_1 \uplus \cdots \uplus f_k$. A dynamic programming algorithm allows one to efficiently compute bounds $\mathrm{BMC}(n, n_1, a_1, b_1)$ such that for every coloring $f = f_1 \uplus \cdots \uplus f_k$, it holds that $\mathrm{BMC}(n, n_1, a_1, b_1) \geq \mathrm{con}(f)$, where $|f| = n$, $|f_1| = n_1$, and olines$(f_1) = (a_1, b_1)$. From this algorithm we get immediately a maximal number of contacts in any coloring with n elements. Further, let a *layer sequence* be a sequence of triples (n_i, a_i, b_i). A coloring f is called *s-compatible*, if every plane restriction f_i of f is compatible to $s_i = (n_i, a_i, b_i)$, i.e. $|f_i| = n_i$ and olines$(f_i) = (a_i, b_i)$.

By traceback from the above dynamic programming algorithm one efficiently obtains the set of all layer sequences s, where there may exist (by our bound) an s-compatible coloring f with b contacts. That is, we define this set of sequences by $S(n, b) := \left\{ s \text{ layer sequence} \middle| \text{bound for } s \text{ greater or equal } b \right\}$.

To find optimally compact colorings it remains to search by constraint based search through the colorings of candidate layer sequences.

Now, we assume that the sets $S(n, b)$ are already precomputed by the dynamic programming algorithm. To find one optimally compact coloring with n elements do the following. Let b_n be the contacts bound for colorings with n elements. For ascending $i \geq 0$, iteratively search for a coloring f with $b_n - i$ contacts in all layer sequences $s \in S(n, b_n - i)$. Clearly, the first coloring f_b found by this procedure has maximal contacts. To find all colorings with a given number k of contacts (e.g. all best colorings) we perform an analogous search in all layer sequences $s \in S(n, b)$.

8 Results

We have computed all sets of layer sequences $S(n, b)$ for $n \leq 100$ in about 10 days on a standard PC. For a given layer sequence one optimally compact core is usually found within a few seconds by our constraint based search program. Some results are shown in Table 1.

We present some of the optimal cores for $n = 60$ and $n = 100$ elements in Figures 4 and 5. The cores are shown as plane sequence representation. This representation shows a coloring by the sequence of its occupied x-layers in the lattice D_3'. For each x-layer $x = x_0$ the lower left corner of the grid has coordinates $(x_0, 0, 0)$. The grid-lines have distance 1. The core points in each x-layer are

Table 1. Search for one optimally compact core with n elements, given a layer sequence. We give the number of contacts, as well as nodes and time of the constraint search.

n	# contacts	# search-nodes	time in s
23	76	15	0.1
60	243	150	0.7
89	382	255	2.1
100	436	82	1.2

shown as filled circles. There is a noteworthy difference between layers $x = x_0$, where x_0 is even and those where it is odd. In the latter ones the points have non-integer y and z coordinates.

Further, we folded some proteins of the FCC-HP-model using a program from [7] to their now proven optimum. The results are shown in Table 2.

References

1. V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich. Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *Journal of Molecular Biology*, 252:460–471, 1995.
2. V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich. Computer simulations of prebiotic evolution. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, and Teri E. Klein, editors, *PSB'97*, pages 27–38, 1997.
3. Richa Agarwala, Serafim Batzoglou, Vlado Dancik, Scott E. Decatur, Martin Farach, Sridhar Hannenhalli, S. Muthukrishnan, and Steven Skiena. Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP-model. *Journal of Computational Biology*, 4(2):275–296, 1997.
4. Rolf Backofen. The protein structure prediction problem: A constraint optimisation approach using a new lower bound. *J. Constraints*, 2000. accepted for publication, special issue on 'Constraints in Bioinformatics/Biocomputing'.
5. Rolf Backofen. An upper bound for number of contacts in the HP-model on the face-centered-cubic lattice (FCC). In Raffaele Giancarlo and David Sankoff, editors, *Proc. of the 11th Annual Symposium on Combinatorial Pattern Matching*

Table 2. Sequences L_1-L_5 (taken from [22]) with absolute walks of optimal conformations in FCC-HP-model. The steps of the walk are given by points of the compass. The $+$ and $-$ indices indicate an additional 45° walk out of the plane.

L_1 HPPPPHHHHPPHPHPHHHPHPPHHPPH :
$N_e^- E N_w^+ S_e^+ S N_w^- S_w^+ S_w^- S_w^- S_e^+ N_e^- N_w^- N_e^+ S_e^- N_e^+ S_w^+ E S N_w^- S W N_e^+ N_w^+ S_e^+ E N_w^-$
L_2 HPPPHHHHPHPHHPPPHPHHPHPPPHP :
$S_e^- S_w^- N S_w^+ N_w^+ N_e^- N_w^- N S_e^+ N_e^+ S W N_e^+ N_w^- N_w^- S S_w^+ S_e^- S_e^- W S_e^+ S_w^+ N_e^+ E N N_e^+$
L_3 HPHHPPHHPPHHHHPPPHPPPHHHPPH :
$S_w^+ E E N_e^- N_w^- S_w^+ N E N_w^+ W S_w^- S_e^+ W S_w^+ S_e^+ E N_e^- S_e^+ N_e^- N S_w^- N_w^+ W W S_e^+ S_e^-$
L_4 HHPHHPHHPHHHHHHPPHHHHPPHHHHHH :
$S_e^+ N_w^+ N_e^- N_w^- N_w^- E E N_e^- S S S_w^+ S_w^+ N_w^- N_e^- N_e^- N S_w^+ S_w^+ S_e^+ N S_e^+ N_e^+ N_w^- S_e^- S_e^- W S_e^+ S_w^+ S_e^- N_e^-$
L_5 PHPPHPPHPPHPPHPPHPPHPPHPPHPPHPPHPPHPPHP :
$S_e^+ S_w^- S_w^+ E S N_w^- E S_e^- E N_w^+ N_e^- W N_w^+ N_w^- E S_e^+ N_e^- S_e^+ S_w^+ E N W N_e^+ W S_w^- N_w^+ S_e^+ S_w^- S_e^+ E N_w^- S S_w^- N S_w^+$

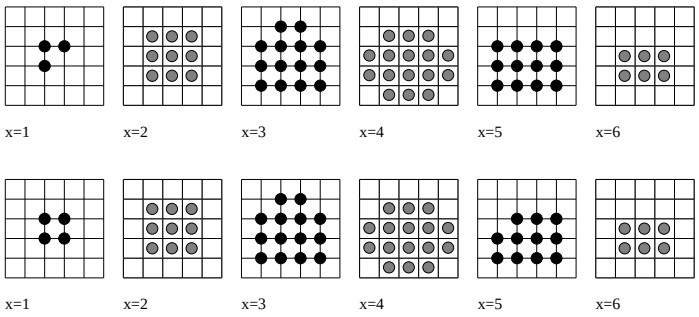


Fig. 4. Plane sequence representation of two optimally compact coloring with $n = 60$ elements.

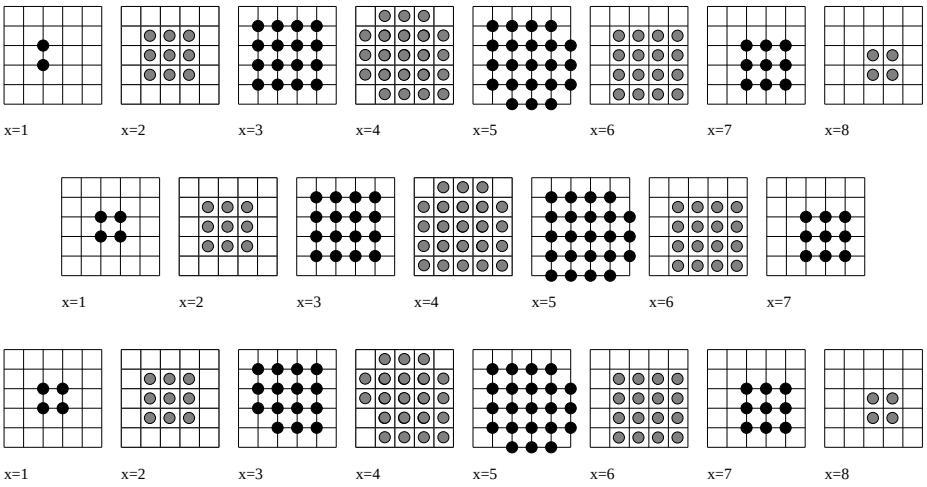


Fig. 5. Plane sequence representations of three optimally compact coloring with $n = 100$ elements.

(CPM2000), volume 1848 of *Lecture Notes in Computer Science*, pages 277–292, Berlin, 2000. Springer–Verlag.
6. Rolf Backofen, Sebastian Will, and Erich Bornberg-Bauer. Application of constraint programming techniques for structure prediction of lattice proteins with extended alphabets. *J. Bioinformatics*, 15(3):234–242, 1999.
7. Rolf Backofen, Sebastian Will, and Peter Clote. Algorithmic approach to quantifying the hydrophobic force contribution in protein folding. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, and Teri E. Klein, editors, *Pacific Symposium on Biocomputing (PSB 2000)*, volume 5, pages 92–103, 2000.
8. B. Berger and T. Leighton. Protein folding in the hydrophobic-hydrophilic (HP) modell is NP-complete. In *Proc. of the Second Annual International Conferences on Compututational Molecular Biology (RECOMB98)*, pages 30–39, New York, 1998.
9. P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. In *Proc. of STOC*, pages 597–603, 1998. Short version in *Proc. of RECOMB'98*, pages 61–62.
10. K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. Principles of protein folding – a perspective of simple exact models. *Protein Science*, 4:561–602, 1995.
11. Aaron R. Dinner, Andreaj Šali, and Martin Karplus. The folding mechanism of larger model proteins: Role of native structure. *Proc. Natl. Acad. Sci. USA*, 93:8356–8361, 1996.
12. S. Govindarajan and R. A. Goldstein. The foldability landscape of model proteins. *Biopolymers*, 42(4):427–438, 1997.
13. William E. Hart and Sorin Istrail. Invariant patterns in crystal lattices: Implications for protein folding algorithms. *J. of Universal Computer Science*, 6(6):560–579, 2000.
14. William E. Hart and Sorin C. Istrail. Fast protein folding in the hydrophobid-hydrophilic model within three-eighths of optimal. *Journal of Computational Biology*, 3(1):53 – 96, 1996.
15. Patrice Koehl and Michael Levitt. A brighter future for protein structure prediction. *Nature Structural Biology*, 6:108–111, 1999.
16. Kit Fun Lau and Ken A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22:3986 – 3997, 1989.
17. Hao Li, Robert Helling, Chao Tnag, and Ned Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273:666–669, 1996.
18. Britt H. Park and Michael Levitt. The complexity and accuracy of discrete state models of protein structure. *Journal of Molecular Biology*, 249:493–507, 1995.
19. Ron Unger and John Moult. Local interactions dominate folding in a simple protein model. *Journal of Molecular Biology*, 259:988–994, 1996.
20. A. Šali, E. Shakhnovich, and M. Karplus. Kinetics of protein folding. *Journal of Molecular Biology*, 235:1614–1636, 1994.
21. Yu Xia, Enoch S. Huang, Michael Levitt, and Ram Samudrala. Ab initio construction of protein tertiary structures using a hierarchical approach. *Journal of Molecular Biology*, 300:171 – 185, 2000.
22. Kaizhi Yue and Ken A. Dill. Sequence-structure relationships in proteins and copolymers. *Physical Review E*, 48(3):2267–2278, September 1993.
23. Kaizhi Yue and Ken A. Dill. Forces of tertiary structural organization in globular proteins. *Proc. Natl. Acad. Sci. USA*, 92:146 – 150, 1995.