In: CPM2000

An Upper Bound for Number of Contacts in the HP-Model on the Face-Centered-Cubic Lattice (FCC)

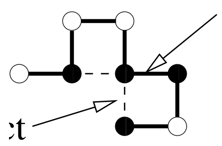
Rolf Backofen

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

Abstract. Lattice protein models are a major tool for investigating principles of protein folding. For this purpose, one needs an algorithm that is guaranteed to find the minimal energy conformation in some lattice model (at least for some sequences). So far, there are only algorithm that can find optimal conformations in the cubic lattice. In the more interesting case of the face-centered-cubic lattice (FCC), which is more protein-like, there are no results. One of the reasons is that for finding optimal conformations, one usually applies a branch-and-bound technique, and there are no reasonable bounds known for the FCC. We will give such a bound for Dill's HP-model on the FCC.

1 Introduction

Simplified protein models such as lattice models are used to investigate the protein folding problem, the major unsolved problem in computational biology. An important representative of lattice models is the HP-model, which has been introduced by [7]. In this model, the 20 letter alphabet of amino acids (called monomers) is reduced to a two letter alphabet, namely H and P. H represents *hydrophobic* monomers, whereas P represent *polar* or hydrophilic monomers. A *conformation* is a self-avoiding walk on the cubic lattice. The energy function for the HP-model simply states that the energy contribution of a contact between two monomers is -1 if both are H-monomers, and 0 otherwise. Two monomers form a *contact* in some specific conformation if they are not connected via a *bond*, and the euclidian distance of the positions is 1. One searches for a conformation which maximizes the number of contacts, which is a conformation whose hydrophobic core has minimal surface. Just recently, the structure prediction problem has been shown to be NP-hard even for the HP-model [4, 5] on the cubic lattice. A sample conformation for the sequence PHPPHHPH in the two-dimensional lattice with energy -2 is the following:



The white beads represent P, the black ones H monomers. The two contacts are indicated via dashed lines.

For investigating general properties of protein-folding, one needs an algorithm which is guaranteed to find a conformation with maximal number of contacts (at least for some sequences, since the problem is NP-hard in general). Although there are approximation algorithms for the HP-model in the cubic lattice [6] and FCC [1], the need of an optimal conformation in this case implies that one cannot use approximate or heuristic algorithms for this purpose. To our knowledge, there are two algorithms known in the literature that find conformations with maximal number of contacts (optimal conformations) for the HP-model, namely [8, 2]. Both use some variant of Branch-and-Bound.

The HP-model is original defined for the cubic lattice, but it is easy to define it for any other lattice. Of special interest is the face-centered-cubic lattice (FCC), which models protein backbone conformations more appropriately than the cubic lattice. When considering the problem of finding an optimal conformation, the problem occurs that no good bound on the number of contacts for the face-centered cubic lattice is known, in contrast to the HP-model. Both known algorithm for finding the optimal conformation search through the space of conformations using the following strategy:

- fix one coordinate (say x) of all H-monomers first
- calculate an upper bound on the number of contacts, given fixed values for the H-monomers.

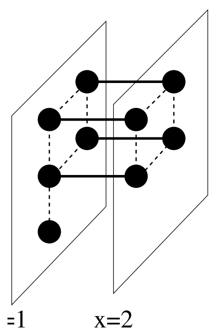


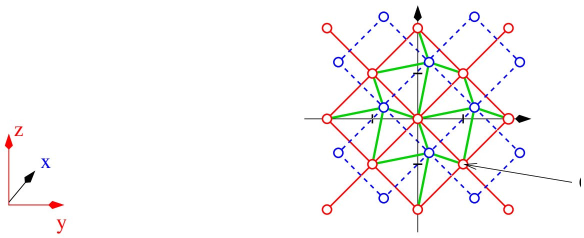
Fig. 1. H-Positions

An upper bound can easily be given in the case of the HP-model, if only the number of H-monomers are known in every plane defined by an equation $x = c$ (called x-layer in the following). For this purpose, one counts the number of HH-contacts and HH-bonds (since the number of HH-bonds is constant, and we do not care in which layer the HH-bonds actually are). Let us call this generalized contacts in the following. Then one distinguishes between generalized contacts within an x-layer, and generalized contacts between x-layers. Suppose that the positions occupied by H-monomers are given by black dots in Figure 1. Then we have 5 H-monomers in layer $x = 1$, and 4 H-monomers in $x = 2$. Furthermore, we have 4 generalized contacts between the layer $x = 1$ and $x = 2$ (straight lines), 5 contacts within $x = 1$ and 4 contacts within $x = 2$ (dashed lines). This coincide with the upper bound given 5 H-monomer in $x = 1$, and 4 H-monomers in $x = 2$, which is calculated as follows. For the number of interlayer contacts, we know that every interlayer contact consumes 1 H-monomer in every layer. Hence, the maximal number of interlayer contacts is the minimum of the number of H-monomer in each layer, in this case $\min(5, 4) = 4$. The upper bound for the layer contacts is a bit more complicated, since it uses the concept of a frame. Consider some layer with n

H-monomers. Let $a = \lceil \sqrt{n} \rceil$ and $b = \lceil \frac{n}{a} \rceil$. (a, b) is the minimal rectangle (frame) around n H-monomers. Then the maximal number of contacts within this layer is upper bound by $2n - a - b$. In our example, we get for the first layer $n = 5$, $a = 3$ and $b = 2$, and the maximal number of layer contacts is then $10 - 3 - 2 = 5$, as it is the case in our example. For $n = 4$, we get $a = 2$, $b = 2$ and the maximal number is then $8 - 2 - 2 = 4$, as in our case. For details, see [8, 2].

The problem with the face-centered-cubic lattice is, that there is no similar bound known (given the number of H-monomers in every x-layer). Part of the problem is that the interlayer contacts are much more complex to determine. The reason is that the face-centered cubic lattice has 12 neighbors (position with minimal distance), whereas the cubic lattice has only 6. Thus, in any representation of FCC, we have more than one neighbor in the next x-layer for any point $\boldsymbol{p}$, which makes the problem complicated. Such an upper bound will be given in this paper.

2 Preliminaries



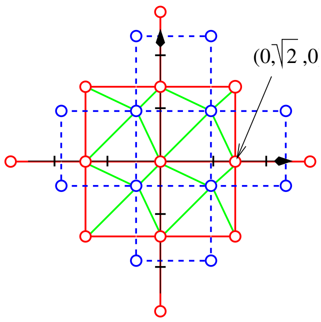


Fig. 2. In the first figure, we have shown two x-layers (where the x-axis is shown as the third dimension). The red circles are the lattice points in the first x-layer, where the red lines are the nearest neighbor connections. The blue circles are the points in the second x-layers. The green lines indicate the nearest neighbor connections between the first and the second x-layer. The second figures shows FCC after rotation by 45°

Given vectors $\boldsymbol{v_1}, \ldots, \boldsymbol{v_n}$, the lattice generated by $\boldsymbol{v_1}, \ldots, \boldsymbol{v_n}$ is the minimal set of points L such that $\forall \boldsymbol{u}, \boldsymbol{v} \in L$, both $\boldsymbol{u} + \boldsymbol{v} \in L$ and $\boldsymbol{u} - \boldsymbol{v} \in L$. An *x-layer* in a lattice L is a plane orthogonal to the x-axis (i.e., is defined by the equation $x = c$) such that the intersection of the points in the plane and the points of L is non-empty. The *face-centered cubic lattice* (FCC) is defined as the lattice $D_3 = \{ \left(\begin{smallmatrix} x \\ y \\ z \end{smallmatrix}\right) \mid \left(\begin{smallmatrix} x \\ y \\ z \end{smallmatrix}\right) \in \mathbb{Z}^3 \text{and } x + y + z \text{ is even}\}$. For simplicity, we use a representation of D_3 that is rotated by $\phi = 45°$ along the x-axis. Since we want to have distance 1 between successive x-layers, and distance 1 between neighbors in one x-layer, we additionally scale the y- and z-axis, but leave the x-axis as it is. A partial view of the lattice and its connections, as well as the rotated lattice is given in Figure 2. Thus, we can define the lattice D_3' to be the lattice that consists of the following sets of points in real coordinates: $D_3' = \{ \left(\begin{smallmatrix} x \\ y \\ z \end{smallmatrix}\right) \mid \left(\begin{smallmatrix} x \\ y \\ z \end{smallmatrix}\right) \in \mathbb{Z}^3\} \uplus \{ \left(\begin{smallmatrix} x+0.5 \\ y+0.5 \\ z+0.5 \end{smallmatrix}\right) \mid \left(\begin{smallmatrix} x \\ y \\ z \end{smallmatrix}\right) \in \mathbb{Z}^3\}$. The first is the set of points in even x-layers, the second the set of point in odd x-layers. A generator matrix for D_3' is given in [3].

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 different successive x-layers. Two points $\boldsymbol{p}$ and $\boldsymbol{p'}$ in D_3' are *neighbors* if $\boldsymbol{p} - \boldsymbol{p'} \in N_{D_3'}$.

Colorings The positions occupied by H-monomers in some conformation of the HP-model can be defined by colorings. A *coloring* is a function $f : D_3' \to \{0, 1\}$. We denote with points(f) the set of all points colored by f, i.e., points$(f) = \{\boldsymbol{p} \mid f(\boldsymbol{p}) = 1\}$. With num$(f)$ we denote $|$points$(f)|$. Let f_1 and f_2 be colorings. With $f_1 \cup f_2$ we denote the coloring f with points$(f) =$ points$(f_1) \cup$ points(f_2). Two colorings f_1, f_2 are *disjoint* if their set of points are disjoint. $f_1 \uplus f_2$ denotes the disjoint union of colorings. Given a coloring f, we define the number of contacts con(f) of f by $\mathrm{con}(f) = \frac{1}{2}|\{(\boldsymbol{p}, \boldsymbol{p'}) \mid f(\boldsymbol{p}) = 1 = f(\boldsymbol{p'}) \wedge (\boldsymbol{p} - \boldsymbol{p'}) \in N_{D_3'}\}$.

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., $\mathrm{Surf}_{pl}(f)$ is the number of pairs $(\boldsymbol{p}, \boldsymbol{p'})$ such that $\boldsymbol{p} - \boldsymbol{p'} \in N_{D_3'}$ (i.e., they are neighbors), $f(\boldsymbol{p}) = 1$, $f(\boldsymbol{p'}) = 0$, and $\boldsymbol{p'}$ is also in plane $x = c$. With $\min_y(f)$ we denote the integer $\min\{y \mid \exists z : f(c, y, z) = 1\}$. $\max_y(f)$, $\min_z(f)$ and $\max_z(f)$ are defined analogously.

3 Description of the upper bound

Our purpose is to give a upper bound on the number of contacts, given that n_c H-monomers are in the x-layer defined by $x = c$. Thus, we need to find a function $b(n_1, \ldots, n_k)$ with

$$b(n_1, \ldots, n_k) \geq \max \left\{ \mathrm{con}(f) \middle| \begin{array}{l} f = f_1 \uplus \ldots \uplus f_k,\ \mathrm{num}(f_c) = n_c \\ \text{and } f_c \text{ is a coloring of plane } x = c \end{array} \right\}.$$

To develop $b(n_1, \ldots, n_k)$, we distinguish between contacts $(\boldsymbol{p}, \boldsymbol{p'})$ where both $\boldsymbol{p}$ and $\boldsymbol{p'}$ are in one x-layer, and contacts $(\boldsymbol{p}, \boldsymbol{p'})$ where $\boldsymbol{p}$ is in an layer $x = c$, and $\boldsymbol{p'}$ is in the layer $x = c + 1$. The contacts within the same x-layer are easy to bound by bounding the surface $\mathrm{Surf}_{pl}(f_c)$. Since every point in layer $x = c$ has four neighbors, which are either occupied by an colored point, or an uncolored point, we get $4 \cdot \mathrm{num}(f) = \mathrm{Surf}_{pl}(f_c) + 2 \cdot LC$, where LC is the number of layer contacts. The hard part is to bound the number of contacts between two successive layers.

For defining the bound on the number of contacts between two successive layers, we introduce the notion of a i-point, where $i = 1, 2, 3, 4$. Given any point in $x = c + 1$, then this point can have at most 4 neighbors in the plane $x = c$. Let f be a coloring of the plane $x = c$. Then a point $\boldsymbol{p}$ in plane $x = c + 1$ is a *i-point for f* if it has i neighbors in plane $x = c$ that are colored by f (where $i \leq 4$). Of course, if one colors a i-point in plane $x = c + 1$, then this point generates i contacts between layer $x = c$ and $x = c + 1$. In the following, we will restrict ourself to the case where $c = 1$ for simplicity. Of course, the calculation is independent of the choice of c.

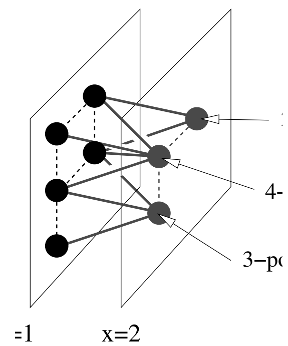


Fig. 3. H-Positions in FCC

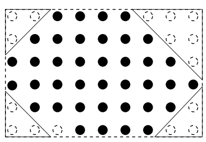
Consider as an example the two colorings f_1 of plane $x = 1$ and f_2 of plane $x = 2$ as shown in Figure 3. f_1 consists of 5 colored points, and f_2 of 3 colored points. Since f_2 colors one 4-point, one 3-point and one 2-point of f_1, there are 9 contacts between these two layers. It is easy to see that we generated the most contacts between layers $x = 1$ and $x = 2$ by first coloring the 4-points, then the 3 points and so on until we reach the number of points to be colored in layer $x = 2$.[1]

For this reason, we are interested to calculate the maximal number of i-points (for $i = 1, 2, 3, 4$), given only the number of colored points n in layer $x = 1$. But this would overestimate the number of possible contacts, since we would maximize the number of 4-, 3-, 2- and 1- point independently from each other. We have found a dependency between these numbers, which requires to fix the side length (a, b) of the minimal rectangle around all colored points in layer $x = 1$ (called the *frame*). In our example, the frame is $(3, 2)$. Of course, one has to search through all "reasonable frames" to find the maximal number of contacts between the two layers. This will be treated in a later section.

Denote with $max_i(a, b, n)$ the maximal number of i-points in layer $x = 2$ for any coloring of layer $x = 1$ with n-colored points and frame (a, b). Then we have found that

$$\begin{aligned} max_4(a,b,n) &= n + 1 - a - b \qquad & max_2(a,b,n) &= 2a + 2b - 2\ell - 4 \\ max_3(a,b,n) &= \ell & max_1(a,b,n) &= \ell + 4. \end{aligned}$$

The remaining part is to find $\ell = max_3(a, b, n)$, which is a bit more complicated. Define $k = \mathrm{edge}(a, b, n) = \max\{k \in \mathbb{N} \mid ab - 4\frac{k(k+1)}{2} \geq n\}$, and $r = \mathrm{ext}(a, b, n) = \lfloor \frac{ab - 4\frac{k(k+1)}{2} - n}{k+1} \rfloor$. Then $max_3(a, b, n) = \begin{cases} 4k + r & \text{if } 4k + r < 2(a-1) \\ 2(a-1) & \text{else.} \end{cases}$
$4k + r$ has a geometric interpretation. It is the maximal sum of the side length (one short side) of *uncolored* triangles at the corners of the frame for colorings with n points and frame (a, b). Consider the coloring



[1] Note that this might not necessarily be the coloring with the maximal number of contacts, since we might loose contacts within the layer $x = 2$; although this could be included in the calculation of the upper bound, we have excluded this effect in for simplicity

with $n = 38$, $a = 6$ and $b = 9$. The sum of the triangle side lengths (one short side) is $2+2+2+3 = 9$. This is exactly the same as $4\mathrm{edge}(a,b,n)+\mathrm{ext}(a,b,n)$.

Plan of the Paper In Section 4, we will determine the number of points having n possible contacts, given some parameter of the coloring f of plane $x = c$. The parameters are the surface $\mathrm{Surf}_{pl}(f)$, and the number of points with 3 possible contacts.

In Section 5, we will then show how we can determine the number of points having 3 possible contacts, given $\mathrm{Surf}_{pl}(f)$. $\mathrm{Surf}_{pl}(f)$ is determined by the minimal rectangle (called frame) around all points colored by f. Thus, we get an upper bound for both the contacts in the plane $x = c$, and the contacts between $x = c$ and $x = c+1$ by enumerating all possible frames for f. Of course, we cannot enumerate *all* frames. Thus, we introduce in Section 6 a concept of "sufficiently filled frames", i.e. frames that are not too big for the number of points to be colored within the frame. These frames will be called normal. We then prove that it is sufficient to enumerate only the normal frames to get an upper bound. In fact, this is the most tedious part of the construction.

4 Number of Points with 1, 2, 3, 4-Contacts

In the following, we want to handle caveat-free, connected colorings. A coloring is connected if for every two points colored by f, there is a path $\boldsymbol{p} = \boldsymbol{p_1} \dots \boldsymbol{p_n} = \boldsymbol{p'}$ such that all points $\boldsymbol{p_1} \dots \boldsymbol{p_n}$ are colored by f, and $\boldsymbol{p_i} - \boldsymbol{p_{i+1}} \in N_{D'_3}$. Let f be a coloring of plane $x = c$. A *horizontal caveat in* f is a k-tuple of points $(\boldsymbol{p_1}, \dots, \boldsymbol{p_k})$ such that $f(\boldsymbol{p}_1) = 1 = f(\boldsymbol{p}_k)$, $\forall 1 < j < k : f(\boldsymbol{p}_j) = 0$ and $\forall 1 \leq j < k : \left(\boldsymbol{p_{j+1}} = \boldsymbol{p_j} + \left(\begin{smallmatrix}0\\1\\0\end{smallmatrix}\right)\right)$. A *vertical caveat* is defined analogously satisfying $\forall j < k : \left(\boldsymbol{p_{j+1}} = \boldsymbol{p_j} + \left(\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right)\right)$ instead. We say that f contains a *caveat* if there is at least one horizontal or vertical caveat in f.

Definition 1. *Let f be a coloring of plane $x = c$. We say that a point $\boldsymbol{p}$ is a 4-point for f if $\boldsymbol{p}$ is in plane $x = c+1$ or $x = c-1$ and $\boldsymbol{p}$ has 4 neighbors $\boldsymbol{p_1}, \dots, \boldsymbol{p_4}$ in plane $x = c$ with $f(\boldsymbol{p_1}) = \cdots = f(\boldsymbol{p_4}) = 1$. Analogously, we define 3-points, 2-points and 1-points. Furthermore, we define $\#4_{c-1}(f) = |\{\boldsymbol{p} \mid \boldsymbol{p}$ is a 4-point for f in $x = c-1\}|$. Analogously, we define $\#4_{c+1}(f)$ and $\#i_{c\pm1}(f)$ for $i = 1, 2, 3$.*

Trivially, we get for any coloring f of plane $x = c$ that $\forall i \in [1..4] : \#i_{c-1}(f) = \#i_{c+1}(f)$. Hence, we define for a coloring f of plane $x = c$ that $\#i(f) = \#i_{c-1}(f) \quad (= \#i_{c+1}(f))$ for every $i = 1 \dots 4$. For calculating the number of i-points for a coloring f of plane $x = c$, we need the additional notion of x-steps for f. An x-step f consists of 3 points in $x = c$ that are sufficient to characterize one 3-point.

Definition 2 (X-step). *Let f be a coloring of plane $x = c$. An x-*step for f *is a triple $(\boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3})$ such that $f(\boldsymbol{p_1}) = 0$, $f(\boldsymbol{p_2}) = 1 = f(\boldsymbol{p_3})$, $\boldsymbol{p_1} - \boldsymbol{p_2} = \pm \left(\begin{smallmatrix}0\\1\\0\end{smallmatrix}\right)$ and $\boldsymbol{p_1} - \boldsymbol{p_3} = \pm \left(\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right)$. With* $\mathrm{xsteps}(f)$ *we denote the number of x-steps of f.*

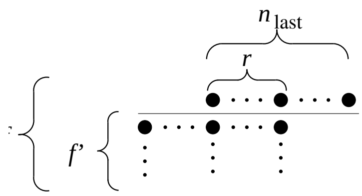
Lemma 1. *Let f be a connected, horizontal caveat-free coloring of the plane $x = c$. Then the following equations are valid:*

$(a)\quad \#4(f) = n + 1 - \frac{1}{2}\mathrm{Surf}_{pl}(f)$ $\qquad\qquad$ $(b)\quad \#3(f) = \mathrm{xsteps}(f)$

$(c)\quad \#2(f) = 2n - 2\#4(f) - 2\#3(f) - 2$ $\qquad\qquad$ $(d)\quad \#1(f) = \mathrm{xsteps}(f) + 4$

Proof (sketch). The proof is by induction on the number of rows. We will handle only the simple case of plane colorings where the different rows do overlap (since we can show that non-overlapping colorings are not optimal). The base case is trivial. For the induction step, let f be a plane coloring of height $h + 1$. Let the coloring f' be f with the row $z = \max_z(f)$ deleted.

Equation (a)**:** Let n_{last} be the number of points introduced in the last row of f, and let r be the number of points where the last row of f overlaps with the last row of f' as shown below:

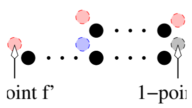


By our assumption of overlapping colorings we get $r > 0$. Then

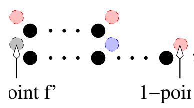
$$\begin{aligned} \#4(f) &= \#4(f') + (r - 1) \\ \text{and} \qquad \mathrm{Surf}_{pl}(f) &= \mathrm{Surf}_{pl}(f') + 2(n_{\mathrm{last}} - r) + 2. \end{aligned}$$

Using these values and the induction hypotheses, a simple calculation proves the claim.

Eqns. (b) **and** (d)**:** We will show only one case for the arrangement of the last two rows. Assume that the last two rows are of the forms



or



,

where the points colored by f are given in black. By this arrangement of the last two rows of f, f contains one more x-steps than f'. Thus, one can easily check that

$$\begin{aligned} \#3(f) &= \#3(f) + 1 \\ \text{and} \qquad \#1(f) &= \#1(f) + 1, \end{aligned}$$

from which one can verify Eqns. (b) and (d) by the induction hypotheses.

Equation (d) follows from the other equations and the equation

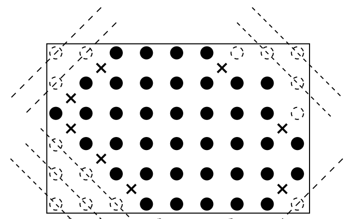
$$4n = 4 \cdot \#4(f) + 3 \cdot \#3(f) + 2 \cdot \#1(f) + 1 \cdot \#1(f).$$

With this lemma we need only to bound $\mathrm{Surf}_{pl}(f)$ and $\mathrm{xsteps}(f)$ (which is a bound on the number of 3-points) to calculate bounds on the number of 4-, 3-, 2- and 1-points.

5 Bound on the Number of 3-Points.

Given a plane coloring f, then we denote with frame(f) the pair (a, b), where $a = \max_z(f) - \min_z(f) + 1$ and $b = \max_y(f) - \min_y(f) + 1$. a is called the *height of* f, and b is called the *width of* f. The frame gives us the possibility to given a lower bound on the surface of a plane coloring, which is then an upper bound on the layer contacts. We need more characteristics of a coloring than the frame to generate a bound for xsteps(f), which will be captured by the notion of a detailed frame. The formal definition can be found in [3]. In principle, the detailed frame just counts for every corner, how many diagonals we can draw (starting from the corner) without touching a point that is colored by f. E.g., consider the following plane coloring f_{ex} given by the black dots:

ex:



Note that there are 8 positions in the next layer that are 3-points for this coloring. We have indicated these points with a $\times$. We can draw 3 diagonals from the left-lower corner, 2 from the left upper, 1 form the right lower, and 2 from the right upper corner. Note that the number of 3-points near every corner is exactly the same. We will prove this relationship later.

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 f_{ex}, the detailed frame is $(6, 9, 3, 2, 1, 2)$. The interesting part is that the the number of diagonals to be drawn gives an upper bound for the number of points to be colored *and* for the number of x-steps.

Proposition 1. *Let* $(a, b, i_1, i_2, i_3, i_4)$ *be the detailed frame of a plane coloring* f. *Then* $\mathrm{num}(f) \leq ab - \sum_{j=1}^{4} \frac{i_j(i_j+1)}{2}$.

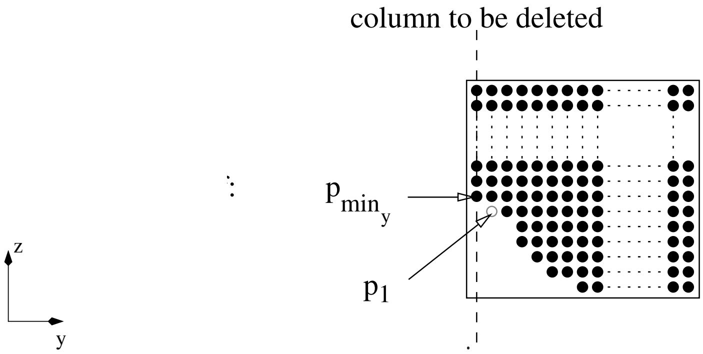
Definition 3 (Diagonal caveat). *A* diagonal caveat *in* f *is a k-tuple of points* $(\boldsymbol{p_1}, \ldots, \boldsymbol{p_k})$ *of* D_3' *with* $k \geq 3$ *such that* $f(\boldsymbol{p}_1) = 1 = f(\boldsymbol{p}_k)$, $\forall 1 < j < k : f(\boldsymbol{p}_j) = 0$ *and* $\forall 1 \leq j < k : \left(\boldsymbol{p_{j+1}} = \boldsymbol{p_j} + \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}\right) \;\vee\; \forall 1 \leq j < k : \left(\boldsymbol{p_{j+1}} = \boldsymbol{p_j} + \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}\right)$.

The number of diagonal caveats in f is denoted by diagcav(f). The next lemma gives us a good bound on the number of 3-points of a plane coloring f, given its edge characteristics. Recall the above example coloring f_{ex} with characteristic $(6, 9, 3, 2, 1, 2)$. Since the coloring does not have any diagonal caveats, the next lemma will show that xsteps(f) is given by $3 + 2 + 1 + 2 = 8$, as we have indicated.

Lemma 2. *Let f be a connected, caveat-free coloring of the plane $x = c$ which has a detailed frame $(a, b, i_1, i_2, i_3, i_4)$. Then*

$$\text{xsteps}(f) = \sum_{j \in [1..4]} i_j - \text{diagcav}(f).$$

Proof (sketch). By induction on the number of columns. The base case is trivial. For the induction step, we will illustrate only the case where the left lower corner of f is not completely filled. I.e., the coloring f is of the form



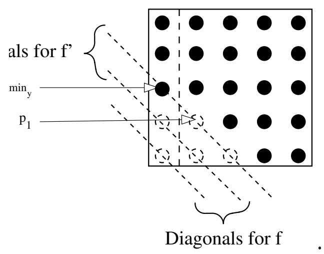
Note that this implies that f has the characteristics $(a, b, i_{lb}, 0, 0, 0)$, and that we have to show that $\text{xsteps}(f) = i_{lb} - \text{diagcav}(f)$. Let f' be generated from f by deleting the first column (as indicated above). Let $\boldsymbol{p}_{min_y}$ be the lowest colored point in the first column of f. We distinguish the following cases for

$$\boldsymbol{p}_1 = \boldsymbol{p}_{min_y} - \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}.$$

1. $f(\boldsymbol{p}_1) = 0$ and $\boldsymbol{p}_{min_y}$ does not start a diagonal caveat in f. This implies that we cannot delete a diagonal caveat by deleting the first column. Additionally, we cannot delete or add an x-step. Hence,

$$\begin{aligned} \text{diagcav}(f') &= \text{diagcav}(f) \\ \text{xsteps}(f') &= \text{xsteps}(f). \end{aligned}$$

For determining the characteristics of f', consider the following example coloring f for this case:



.

Then by deleting the first column, we loose the diagonal through the left lower corner, but we can draw an additional one through $\boldsymbol{p}_1$. Thus, f' has the characteristics $(a, b-1, i_{lb}, 0, 0, 0)$. From this the claim follows from the induction hypotheses.

2. $f(\boldsymbol{p}_1) = 0$ and $\boldsymbol{p}_{min_y}$ starts a diagonal caveat. Then we get

$$\begin{aligned} \text{diagcav}(f') &= \text{diagcav}(f) - 1 \\ \text{xsteps}(f') &= \text{xsteps}(f) \end{aligned}$$

   Since $\boldsymbol{p}_{min_y}$ starts a diagonal caveat, we know that $\boldsymbol{p}_1$ is part of this diagonal caveat. Hence, we cannot draw an additional diagonal in f' through $\boldsymbol{p}_1$ as in the previous case. Thus, f' has the characteristics $(a, b-1, i_{lb}-1, 0, 0, 0)$, and we get the claim again by the induction hypotheses.
3. $f(\boldsymbol{p}_1) = 1$. By deleting the first column, we deleted one x-step, but do not introduced or deleted a diagonal caveat. Hence,

$$\begin{aligned} \text{diagcav}(f') &= \text{diagcav}(f) \\ \text{xsteps}(f') &= \text{xsteps}(f) - 1 \end{aligned}$$

   Furthermore, f' has the characteristics $(a, b-1, i_{lb}-1, 0, 0, 0)$. From this the claim follows from the induction hypotheses.

Proposition 2. *Let f be a caveat-free and connected coloring of plane $x = c$ with frame (a, b). Then* $\text{xsteps}(f) \leq 2(\min(a, b) - 1)$.

Proof (sketch). This proposition follows from the fact that in every connected and caveat-free coloring, there can be at most 2 x-steps in very row (resp. in very column).

6 Number of Contacts between two Planes

As already mentioned in Section 2, for every coloring f we need to distinguish between contacts, where both points are in the same layer, and contacts, where the two corresponding points are in successive layers. The first one are called *layer contacts of f* (denoted by LC_f^c), whereas the later one are called *interlayer contacts*. Since we can split every coloring into a set of plane colorings, we define this notions for plane colorings.

Layer and Interlayer Contacts Let f be a coloring of plane $x = c$. Since all colored points of f are in plane $x = c$, we can define the *layer contacts* LC_f^c of f in the plane $x = c$ by $\text{LC}_f^c = \text{con}(f)$. We define $\text{LC}_{n,a,b}$ to be the maximum of all LC_f^c with $\text{num}(f) = n$, f has frame (a, b) and f is a coloring of some plane $x = c$.

Proposition 3. *Under assumption of caveat-free colorings,* $\text{LC}_{n,a,b} = 2n - a - b$.

Proof. Let f be a coloring of an arbitrary plane $x = c$. If f is caveat-free, then the surface of f in the plane $x = c$ is $2a + 2b$. Now we know that each of the n points has 4 neighbors, which are either occupied by another point, or by a surface point. Hence, we get $4n = 2\mathrm{LC}_{n,a,b} + 2a + 2b$.

Let f be a coloring of plane $x = c$, and f' be a coloring of plane $x = c'$. If $c' = c + 1$ (resp. $c - 1$), then we define the *interlayer contacts* $\mathrm{IC}_f^{f'}$ to be the number of contacts between plane $x = c$ and $x = c + 1$ (resp. $x = c - 1$) in the coloring $f \uplus f'$, i.e.: $\mathrm{IC}_f^{f'} = \left| \left\{ (\boldsymbol{p}, \boldsymbol{p'}) \middle| f(\boldsymbol{p}) = 1 \wedge f'(\boldsymbol{p'}) = 1 \wedge \boldsymbol{p'} - \boldsymbol{p} = \left(\begin{smallmatrix} \pm 1 \\ \pm 0.5 \\ \pm 0.5 \end{smallmatrix}\right) \right\} \right|$. Otherwise, $\mathrm{IC}_f^{f'} = 0$.

Let f be a coloring of plane $x = c$. With $\mathrm{contacts}_{\max}(f, n)$ we denote the maximal number of contacts between plane $x = c$ and $x = c+1$ by placing n points in $x = c+1$. I.e. $\mathrm{contacts}_{\max}(f, n) = \max \left\{ \mathrm{IC}_f^{f'} \middle| \begin{array}{l} f' \text{ is a plane coloring of } x = c + 1 \\ \text{with } \mathrm{num}(f') = n \end{array} \right\}$

Lemma 3. *Let f be a plane coloring of $x = c$. With $\delta_0(k)$ we denote* $\max(k, 0)$. *Then*

$$\begin{aligned} \mathrm{contacts}_{\max}(f, n) = {} & 4 \min(n, \#4(f)) + 3 \min(\delta_0(n - \#4(f)), \#3(f)) \\ & + 2 \min(\delta_0(n - \sum_{i=3}^{4} \#i(f)), \#2(f)) + 1 \min(\delta_0(n - \sum_{i=2}^{4} \#i(f)), \#1(f)) \end{aligned}$$

Next, we want not to consider a special coloring, but only the frame the coloring has. With $\mathrm{MIC}_{n_1,a_1,b_1}^{n_2,a_2,b_2}$ we denote

$$\max \left| \left\{ \mathrm{IC}_{f_1}^{f_2} \middle| \begin{array}{l} \mathrm{num}(f_1) = n_1 \wedge \mathrm{frame}(f_1) = (a_1, b_1) \wedge \\ \mathrm{num}(f_2) = n_2 \wedge \mathrm{frame}(f_2) = (a_2, b_2) \end{array} \right\} \right|$$

We define $\mathrm{MIC}_{n_1,a_1,b_1}^{n_2} = \max\limits_{a_2,b_2} \mathrm{MIC}_{n_1,a_1,b_1}^{n_2,a_2,b_2}$.

Proposition 4. $\mathrm{MIC}_{n_1,a_1,b_1}^{n_2} = \max \left\{ \mathrm{contacts}_{\max}(f, n_2) \middle| \begin{array}{l} f \textit{ is a plane coloring} \\ \textit{with frame } (a_1, b_1) \\ \textit{and } \mathrm{num}(f) = n_1. \end{array} \right\}$

Normal Colorings Now we proceed as follows. We will first consider the case that the frame is sufficiently filled (where we define what this means in a moment). After that, we will show that we do not have to consider the frames which are not sufficiently filled (the pathological cases). We start with defining what "sufficiently filled" means.

Let a, b, n be positive numbers such that $ab \geq n$. We define $\mathrm{edge}(a, b, n)$ by $\mathrm{edge}(a, b, n) = \max\{k \in \mathbb{N} \mid ab - 4\frac{k(k+1)}{2} \geq n\}$. Let $k = \mathrm{edge}(a, b, n)$. Then we define $\mathrm{ext}(a, b, n) = \lfloor \frac{ab - 4\frac{k(k+1)}{2} - n}{k+1} \rfloor$. Intuitively, $\mathrm{edge}(a, b, n)$ is the lower bound for the indent from the corners of a coloring of n points with frame (a, b), if we try to make the indents as uniform as possible (since uniform indents generate the maximal number of x-steps). $\mathrm{ext}(a, b, n)$ is the number of times we can add 1 to $\mathrm{edge}(a, b, n)$. Using this definitions, we can say what sufficiently filled means.

Proposition 5. $0 \leq \mathrm{ext}(a, b, n) \leq 3$

Definition 4 (Normal). *Let n be an integer, (a, b) be a frame with $a \leq b$. Furthermore, let $k = \mathrm{edge}(a, b, n)$ and $r = \mathrm{ext}(a, b, n)$. We say that n is* normal for *(a, b) if either $4k+r < 2(a-1)$, or $4k+r = 2(a-1)$ and $ab-4\frac{k(k+1)}{2}-r(k+1) = n$.*

The reason for using this notion is that if n is normal for (a, b), $\mathrm{edge}(a, b, n)$ and $\mathrm{ext}(a, b, n)$ yield a good bound on the number of x-steps of a plane coloring f. This will be shown in the next two lemmas.

Lemma 4. *If n is normal for (a, b), then there exists a caveat-free, connected plane coloring f such that $\mathrm{xsteps}(f) = 4k + r$, where $k = \mathrm{edge}(a, b, n)$ and $r = \mathrm{ext}(a, b, n)$.*

Lemma 5. *Let (a, b) be a frame of a caveat-free and connected plane coloring f with $a \leq b$. Let $k = \mathrm{edge}(a, b, \mathrm{num}(f))$ and $r = \mathrm{ext}(a, b, \mathrm{num}(f))$. Then*

$$\mathrm{xsteps}(f) \leq \begin{cases} 4k + r & \text{If } 4k + r < 2(a-1) \\ 2(a-1) & \text{else} \end{cases}$$

Definition 5 (Upper bound for $\mathrm{MIC}^{n'}_{n,a,b}$). *Let n be a number and $a \leq b$ with $ab \geq n \geq (a + b) - 1$. Let $k = \mathrm{edge}(a, b, \mathrm{num}(f))$ and $r = \mathrm{ext}(a, b, \mathrm{num}(f))$, and let*

$$l = \begin{cases} 4k + r & \text{if } 4k + r < 2(a-1) \\ 2(a-1) & \text{else.} \end{cases}$$

We define $\quad max_4(a, b, n) = n + 1 - a - b \quad max_2(a, b, n) = 2a + 2b - 2l - 4$
$\quad max_3(a, b, n) = l \quad max_1(a, b, n) = l + 4.$

With $\delta_0(n)$ we denote $max(n, 0)$. Now we define

$$\begin{aligned} \mathrm{BMIC}^{n'}_{n,a,b} = & 4\min(n', max_4(a, b, n)) + 3\min(\delta_0(n' - max_4(a, b, n), max_3(a, b, n)) \\ & + 2\min(\delta_0(n' - \textstyle\sum_{i=3}^{4} max_i(a, b, n), max_2(a, b, n)) \\ & + 1\min(\delta_0(n' - \textstyle\sum_{i=2}^{4} max_i(a, b, n), max_1(a, b, n)). \end{aligned}$$

Theorem 1. *Under the condition given in the last definition, $\mathrm{BMIC}^{n'}_{n,a,b}$ is an upper bound for $\mathrm{MIC}^{n'}_{n,a,b}$, i.e., $\mathrm{MIC}^{n'}_{n,a,b} \leq \mathrm{BMIC}^{n'}_{n,a,b}$. If n is normal for (a, b), then the above bound is tight, i.e., $\mathrm{BMIC}^{n'}_{n,a,b} = \mathrm{MIC}^{n'}_{n,a,b}$.*

Proof. That $\mathrm{BMIC}^{n'}_{n,a,b}$ is an upper bound for $\mathrm{MIC}^{n'}_{n,a,b}$ follows from Lemmas 1, 3, 5 and from the fact that all plane colorings f with frame (a, b) satisfy $\mathrm{Surf}_{pl}(f) \geq 2a + 2b$. That the bound is tight if n is normal for (a, b) follows from Lemma 4.

Note that any frame (a, b) for a connected, caveat-free coloring f with $\mathrm{num}(f) = n$ will satisfy $ab \geq n \geq (a + b) - 1$, which is the reason for the bound on n in the above definition. We need to investigate properties of frames with respect to normality in greater detail. The next lemma just states that normality is kept if we either add additional colored points without changing the frame, or we switch to a smaller frame for the same number of colored points.

Lemma 6. *Let n be normal for (a, b). Then all $n \leq n' \leq ab$ are normal for (a, b). Furthermore, for all (a', b') such that $a' \leq a \wedge b' \leq b$ with $a'b' \geq n$, we have n is normal for (a', b').*

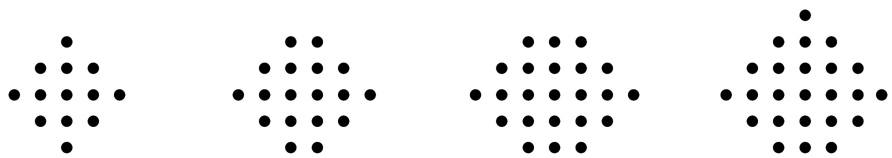
Clearly, we want to search only through the normal frames in order to find the frame (a, b) which maximizes $\mathrm{MIC}^{n'}_{n,a,b}$, given n and n'. This will be subject of Theorem 2.

Restriction to Normal Colorings For this purpose, we define colorings which have

- maximal number of x-steps for given frame (a, b) (i.e., $\mathrm{xsteps}(f) = 2\min(a, b) - 1$,
- maximal number of colored points under the above restriction.

To achieve $\mathrm{xsteps}(f) = 2\min(a, b) - 1$, we must have 2 x-steps in every line. By caveat-freeness, this implies that these maximal colorings are as given in Figure 4.

The definition of these colorings is achieved by defining maximal line number distributions (where maximal refers to maximal x-steps). Line number distributions just counts for every line (row) in the coloring, how many colored points are on that line. The maximal line number distribution for a frame (a, b) is given by $D^{a,b}_{\max3}$, and which has the property that below the line with maximal number of colored points, we add 2 points from line to line, and after the maximal line we subtract 2 points. For every line number distribution D, we have defined a canonical coloring $f_{\mathrm{can}(D)}$. $\mathrm{num}(D)$ is the number of colored points of D, which is the same as the points colored by $f_{\mathrm{can}(D)}$. The precise definitions can be found in [3]. Figure 4 gives examples of the corresponding canonical colorings with maximal number of x-steps for the frames $(5, 5)$, $(5, 6)$, $(5, 7)$ and $(6, 7)$.



$f_{\mathrm{can}(D^{5,5}_{\max3})}$ $f_{\mathrm{can}(D^{5,6}_{\max3})}$ $f_{\mathrm{can}(D^{5,7}_{\max3})}$ $f_{\mathrm{can}(D^{6,7}_{\max3})}$

Fig. 4. Canonical colorings for the elements $(5, 5)$, $(5, 6)$, $(5, 7)$ and $(6, 7)$ of M.

Now we want to find for a given n a minimal frame (a_m, b_m) such that (a_m, b_m) has maximal number of x-steps. For this purpose, we define a set of tuples

$$M = \bigcup \{\{(n, n), (n, n + 1), (n, n + 2), (n + 1, n + 2)\} \mid n \text{ odd}\}$$

Note that M is totally ordered by the lexicographic order on tuples. Hence, we can define $\mathrm{MinF}(n)$ to be the minimal element $(a, b) \in M$ such that $\mathrm{num}(D^{a,b}_{\max 3}) \geq n$.

Note that we have excluded the case (n, n) with n even in the set M. The reason is that in this case, any coloring f of this frame which has maximal number of x-steps (namely $2(n-1)$) is not maximally overlapping. This implies that we can reduce this to a smaller frame. Figure 5 shows an example.

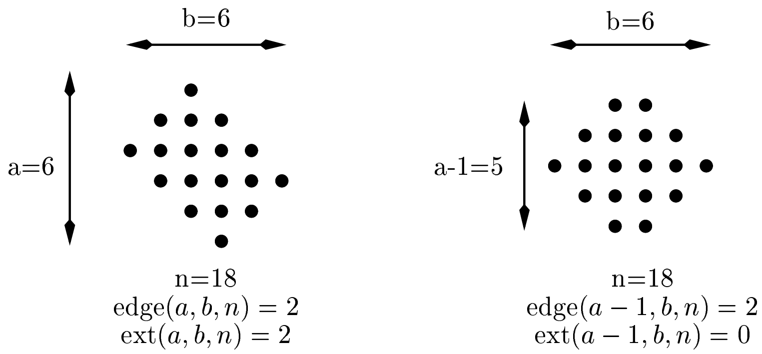


Fig. 5. Special case that a is even and $4\mathrm{edge}(a, b, n) + \mathrm{ext}(a, b, n) = 2(a-1)$. The first picture is the coloring for (a, b), the second for $(a-1, b)$.

Lemma 7. *Let n be a number and (a, b) be $\mathrm{MinF}(n)$. Then*

- *There is a plane coloring f with frame (a, b) such that $\mathrm{num}(f) = n$ and $\mathrm{xsteps}(f) = 2(a-1)$.*
- *n is normal for (a, b) or $(a, b-1)$.*

Theorem 2 (Existence of optimal normal frame). *Let n be an integer. Then for all frames (a', b') there is a frame (a, b) such that $a \leq a' \wedge b \leq b'$, n is normal for (a, b), $(a, b-1)$ or $(a-1, b)$ and $\forall n' : \mathrm{MIC}^{n'}_{n,a,b} \geq \mathrm{MIC}^{n'}_{n,a',b'}$.*

Proof (sketch). The main idea of this theorem is the following. Fix n and n'. Let (a, b) be a frame for n with maximal number of possible x-steps (i.e., there is a plane coloring f with $\mathrm{num}(f) = n$, f has frame (a, b), and $\mathrm{xsteps}(f) = 2\min(a, b) - 1$). Then we know that $\mathrm{MIC}^{n'}_{n,a+1,b} \leq \mathrm{MIC}^{n'}_{n,a,b}$ since be enlarging the frame, we loose one 4-point by lemma 1, but can win at most one x-step by Proposition 2. The same holds for $\mathrm{MIC}^{n'}_{n,a,b+1}$. Thus, it is sufficient to consider the minimal frame (a_m, b_m) which has maximal number of possible x-steps. But we can show that in this case, n is normal for (a_m, b_m), $(a_m, b_m - 1)$ or $(a_m - 1, b_m)$.

This theorem states, that we need only to consider all frames that are within distance one from a normal frame in order to find the frame (a, b) with that maximizes $\mathrm{MIC}^{n'}_{n,a,b}$ for a given n and n'. Now we are able to summarize the results.

Theorem 3. *Let f be a connected, caveat-free coloring with $f = f_1 \uplus \ldots \uplus f_k$, where f_i is a coloring of the plane $x = i$. Let $a^m_k = \lceil \sqrt{n_k} \rceil$ and $b^m_k = \lceil \frac{n_k}{a^m_k} \rceil$. Then*

$$\mathrm{con}(f) \leq \sum_{i=1}^{k-1} \max \left\{ \mathrm{BMIC}^{n_{i+1}}_{n_i,a_i,b_i} + \mathrm{LC}_{n_i,a_i,b_i} \;\middle|\; \begin{array}{l} a_i b_i \geq n_i \textit{ and } n_i \textit{ is} \\ \textit{normal for } (a_i, b_i), \\ (a_i - 1, b_i) \textit{ or } (a_i, b_i - 1) \end{array} \right\} + \mathrm{LC}_{n_k,a^m_k,b^m_k},$$

References

1. 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.
2. Rolf Backofen. Constraint techniques for solving the protein structure prediction problem. In Michael Maher and Jean-Francois Puget, editors, *Proceedings of* 4^{th} *International Conference on Principle and Practice of Constraint Programming (CP'98)*, volume 1520 of *Lecture Notes in Computer Science*, pages 72–86. Springer Verlag, 1998.
3. Rolf Backofen. *Optimization Techniques for the Protein Structure Prediction Problem.* Habilitationsschrift, University of Munich, 1999.
4. 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.
5. 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.
6. 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.
7. 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.
8. Kaizhi Yue and Ken A. Dill. Forces of tertiary structural organization in globular proteins. *Proc. Natl. Acad. Sci. USA*, 92:146 – 150, 1995.