SIAM J. COMPUT.
Vol. 37, No. 4, pp. 1014–1029

GEOMETRIC SEPARATORS AND THEIR APPLICATIONS TO PROTEIN FOLDING IN THE HP-MODEL*

BIN FU† AND WEI WANG‡

Abstract. We develop a new method for deriving a geometric separator for a set of grid points. Our separator has a linear structure, which can effectively partition a grid graph. For example, we prove that for a grid graph G with a set of n points P in a two-dimensional grid, there is a separator with at most $1.129\sqrt{n}$ points in P that partitions G into two disconnected grid graphs each with at most $\frac{2n}{3}$ points. Our separator theorem for grid graphs has a significantly smaller upper bound than that was obtained for the general planar graphs in [H. N. Djidjev and S. M. Venkatesan, *Acta Inform.*, 34 (1997), pp. 231–234]. The protein folding problem in the HP-model is to put a sequence, consisting of two characters H and P, in a d-dimensional grid to have maximal number of HH-contacts, where an HH-contact is a pair of non-consecutive H letters that are put at two grid points of distance 1. Our separator is then applied to develop an exact algorithm for the protein-folding problem in the HP-model, which is NP-hard both in both two and three dimensions [B. Berger and T. Leighton, *J. Comput. Biol.*, 5 (1998), pp. 27–40; P. Crescenzi et al., *J. Comput. Biol.*, 5 (1998), pp. 423–465]. We design a $2^{O(n^{1-\frac{1}{d}}\log n)}$ time algorithm for the d-dimensional protein folding problem in the HP-model. In particular, our algorithm has $O(2^{6.145\sqrt{n}\log n})$ and $O(2^{6.913n^{\frac{2}{3}}\log n})$ computational time in two and three dimensions, respectively.

Key words. separator, protein folding, time complexity, algorithm

AMS subject classification. 68W01

DOI. 10.1137/S0097539704440727

1. Introduction.

Geometric separators are fundamental tools in algorithm design for solving many geometric problems. Lipton and Tarjan [24] showed a well-known geometric separator for planar graphs. Their result has been elaborated on by many authors [9, 15, 2, 10]. The following best known separator theorem for planar graphs was proved by Djidjev and Venkatesan [10].

THEOREM 1 (see [10]). *Any planar graph of n vertices has a vertex subset of cardinality $\leq 1.97\sqrt{n}$ whose removal separates the graph into two components each having at most $\frac{2n}{3}$ vertices.*

Spielman and Teng [38] showed a $\frac{3}{4}$-separator with size $1.82\sqrt{n}$ for planar graphs. Separators for more general graphs were presented in, e.g., [16, 3, 32]. Planar graph separators were applied to derive some $2^{O(\sqrt{n})}$-time algorithms for certain NP-hard problems about planar graphs by Lipton and Tarjan [25] and Ravi and Hunt [35]. Those problems include computing a maximum independent set, a minimum vertex cover, and three-colorings of a planar graph, and the number of satisfying truth assignments to a planar 3CNF formula [23].

*Received by the editors February 8, 2004; accepted for publication (in revised form) March 3, 2007; published electronically September 19, 2007. This research was supported by the Louisiana Board of Regents fund under contract LEQSF(2004-07)-RD-A-35. An earlier version of this paper was presented at ICALP 2004 [14].

http://www.siam.org/journals/sicomp/37-4/44072.html

†Department of Computer Science, University of Texas-Pan American, Edinburg, TX 78539-2999 (binfu@cs.panam.edu). Part of this research was done while the author was affiliated with the Department of Computer Science, University of New Orleans, New Orleans, LA 70148, and Research Institute for Children, 200 Henry Clay Avenue, New Orleans, LA 70118.

‡Department of Chemistry and Biochemistry, University of California at San Diego, San Diego, CA 92093 (wwang@chem.ucsd.edu).

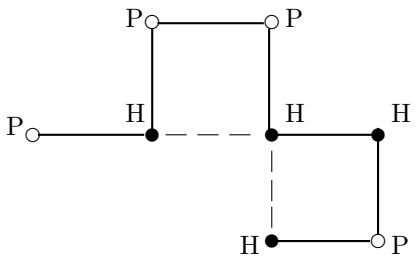

FIG. 1. *The sequence PHPPHHPH is put on the 2D grid. There are two H-H contacts marked by the dotted lines.*

Some other forms of the geometric separators were studied by Miller, Teng, and Vavasis [26], Miller and Thurston [27] and Smith and Wormald [37]. For a set of regular geometric objects such as circles, rectangles, etc., if every point on the plane is covered by at most k objects, the set of the objects is called a k-thick system. Some $O(\sqrt{k \cdot n})$ size separators for k-thick systems and the algorithms for finding them were derived in [27, 26, 37]. Smith and Wormald [37] applied their separators to develop algorithms for some geometric problems such as the planar travelling salesman and the Steiner tree problems (e.g., see [37]). Those problems usually have input points with fixed geometric positions in space.

Finding a minimal size separator, which maintains a similar balance partition condition like one of those mentioned above (e.g., each side of the separator has at most $\frac{2n}{3}$ points), for a grid graph is also an interesting combinatorial problem. Using special geometric properties of grid points, we develop a method for obtaining a separator with a smaller size for grid points via controlling the distances from the grid points to the separator line.

A set of grid points on the plane forms a grid graph by adding edges to every two grid points with distance 1. A grid graph is also a planar graph. In the protein folding of the HP-model, the 20 letter alphabet of amino acids is reduced to a two letter alphabet, namely H and P. H represents hydrophobic amino acids, whereas P represents polar or hydrophilic amino acids. Two monomers form a contact in some specific conformation if they are not consecutive, but occupy neighboring positions in the conformation (i.e., the distance vector between their positions in the conformation is a unit vector). A conformation with minimal energy is just a conformation with the maximal number of contacts between nonconsecutive H-monomers (see Figure 1). The protein folding problem in the HP-model is to find the conformation for any HP-sequence with minimal energy. The protein folding problem in the HP-model is an interesting and challenging problem that deals with a puzzle in the grids. The problem is to put a sequence, consisting of two characters H and P, in a d-dimensional grid to have the the maximal number of HH-contacts. As the input of the protein folding problem is only a sequence of letters, the locations of those letters in space are unknown and will be determined by the algorithm. For this reason, we do not know whether the separator theorems, such as Theorem 1, can be applied to the protein folding problem. We derive a separator theorem for the grid graph with a significantly smaller upper bound on the number of points on the separator than that obtained for planar graphs. Our result is stated as the following theorem.

THEOREM 2. *For a set P of n grid points on the two-dimensional (2D) plane, there is a line on the plane and a subset $Q \subseteq P$ of cardinality $\leq 1.129\sqrt{n}$ such that each half-plane determined by the line contains at most $\frac{2}{3}n$ points of P, and every two*

points $p_1, p_2 \in P$ on the different sides of the line have distance > 1, unless at least one of p_1, p_2 is in Q.

Furthermore, we also provide $O(n^2)$ possible locations to find such a line based on the folding region within a fixed $n \times n$ square. This makes it possible to use the separator theorem in the algorithm for the protein folding problem, even though the locations of the letters are not known. The approximation method in searching for the separator region highly depends on the linear structure of the separator. We derive a similar separator for three-dimensional (3D) grid points, which is also applied to the 3D protein folding problem.

The lower bounds of $1.555\sqrt{n}$ and $1.581\sqrt{n}$ for the $\frac{2}{3}$-separator like that in Theorem 2 for the planar graph were proved by Djidev and Venkatesan [10] and Smith and Wormald [37], respectively. However, we develop a new approach and obtain an upper bound on the separator for a grid graph with a size smaller than their lower bounds. This shows that our smaller-sized separator for grid graphs cannot be directly obtained from that for planar graphs.

Our development of the separator technology is motivated by finding fast exact algorithms for the protein folding problem. A protein can be folded into a specific 3D structure, which is uniquely determined by the sequence of amino acids. Its 3D structure determines its function. Protein structure prediction with computational technology is one of the most significant problems in bioinformatics.

A simplified representation of proteins is a lattice conformation, which is a self-avoiding sequence in Z^3. An important representative of lattice models is the HP-model, which was introduced in [20, 21]. This problem was proven to be NP-hard both on two and three dimensions [6, 8].

Some algorithms for this problem have been developed based on heuristic, genetic, Monte Carlo, branch, and bound methods (e.g., [39, 40, 41, 36, 30, 33, 18, 19, 31, 22, 34, 5, 4]). Although many experimental results were reported for testing sequences of small lengths, we have not seen any theoretical analysis about upper bounds on the computational time of the algorithms. Another approach is to develop polynomial time approximation algorithms for protein folding in the HP-model, [17, 1, 28]. Hart and Istrail [17] showed a polynomial time $\frac{3}{8}$-approximation algorithm for the 3D protein folding in the HP-model and Newman [28] derived a polynomial time $\frac{1}{3}$-approximation algorithm for the 2D problem, improving the $\frac{1}{4}$-approximation algorithm in [17].

If the first letter of an HP-sequence is fixed at a position of a 2D plane (or 3D space), we have at least 2^{n-1} (3^{n-1}) ways and at most $3^{n-1}(5^{n-1})$ ways to put the rest of the letters on the plane (in the space, resp.). The computational time of our algorithm is bounded by $2^{O(n^{\frac{1}{2}}\log n)}$ ($2^{O(n^{\frac{2}{3}}\log n)}$) in two dimensions (in three dimensions, resp.). As the average number of amino acids of proteins is between 400 and 600, if an algorithm could solve the protein structure prediction problem with ≤ 1000 amino acids, it would be able to satisfy most of the application demands. Our effort is a theoretical step toward this target. Our algorithm is a divide-and-conquer approach, which is based on our geometric separator for separating the points in a d-dimensional grid.

The paper is organized as follows. In sections 2 and 3, we develop the separator theory. In sections 4 and 5, we apply the separators to the protein folding problem in the HP-model. In section 2, we show a class of easy separators on a set of d-dimensional grid points. This kind of separator is used in section 4 to obtain a $2^{O(n^{1-1/d}\log n)}$-time algorithm for the d-dimensional protein folding problem in the HP-model. In section 3, we develop sharp separators in both two and three dimen-

sions. Those separators are used to obtain faster algorithms for the protein folding problem in section 5. Precisely, those separators help us reduce the constant factor in the exponents of the computational complexity of the protein folding problem.

In this paper, we only apply the separators for grid points to the protein folding problem in the HP-model. When developing algorithms for some geometric problems with the input of a set of points in the Euclidean space, we can select a set of grid points to characterize the distribution of the input points. This brings more applications of separators for grid points. A series of advances [11, 7, 13, 12] has been made along this line of the separator technology, which starts from the earlier version [14] of this paper. The method of this paper was extended and applied to a class of other NP-hard geometric problems by Fu [11], improving their exact algorithms to $2^{O(\sqrt{n})}$-time from $n^{O(\sqrt{n})}$-time. Those problems include the problems of disk covering, maximum independent set, minimum vertex cover, and minimum dominating set on disk graphs. An efficient sublinear time randomized algorithm was developed in [12] for finding separators. The method was also applied to derive an approximation algorithm for a geometric problem [7] that has application in digital image half-toning.

2. An easy separator for grid points. Given a set of n grid points P, we will show that there is a hyperplane (denoted by $P_{r,a}$ for some r with $1 \leq r \leq d$ and an integer a in the definition below), which contains $O(n^{1-\frac{1}{d}})$ grid points from P, to partition P into two parts of at most $c(d)n$ grid points on each side, where $0 < c(d) < 1$ and $c(d)$ is a constant for fixed d. The separator in this section has a self-contained proof and is used in deriving an $n^{O(n^{1-\frac{1}{d}})}$-time algorithm for the protein folding problem in the HP-model in section 4. Let the dimensional number d be fixed. We need the following terms.

DEFINITION 3.

- *For a set A, $|A|$ denotes the number of elements in A.*
- *The integer set is represented by $Z = \{\cdots, -2, -1, 0, 1, 2, \cdots\}$. For integers i and j, integer interval $[i, j] = \{i, i+1, \cdots, j\}$. For integers $x_1, \cdots, x_d$, $(x_1, \cdots, x_d)$ is a d-dimensional grid point.*
- *For two points p_1, p_2 with the same dimension,* $\mathrm{dist}(p_1, p_2)$ *is the Euclidean distance between them.*
- *An r-plane is the set $P_{r,a} = \{(x_1, \cdots, x_{r-1}, a, x_{r+1}, \cdots, x_d) | x_1, \cdots, x_{r-1}, x_{r+1}, \cdots, x_d \in Z\}$, which has all of the elements in Z^d with the rth component being a fixed value a.*
- *$P_{r,<a} = \{(x_1, \cdots, x_{r-1}, x_r, x_{r+1}, \cdots, x_d) | x_1, \cdots, x_{r-1}, x_r, x_{r+1}, \cdots, x_d \in Z$ and $x_r < a\}$.*
- *$P_{r,>a} = \{(x_1, \cdots, x_{r-1}, x_r, x_{r+1}, \cdots, x_d) | x_1, \cdots, x_{r-1}, x_r, x_{r+1}, \cdots, x_d \in Z$ and $x_r > a\}$.*
- *$P_{r,\leq a} = Pr_{r,<a} \cup P_{r,a}$, and $P_{r,\geq a} = P_{r,>a} \cup P_{r,a}$.*
- *For a set of points S in the d-dimensional space and $1 \leq r \leq d$ and $a \in Z$, define $S(r, < a) = \{(x_1, \cdots, x_d) \in S | x_r < a\}$, $S(r, = a) = \{(x_1, \cdots, x_d) \in S | x_r = a\}$, and $S(r, > a) = \{(x_1, \cdots, x_d) \in S | x_r > a\}$.*
- *For $0 < c < 1$ and a set S in the d-dimensional space, a $P_{r,a}$ is a c-balanced-separator if $|S(r, < a)| \leq c \cdot |S|$ and $|S(r, > a)| \leq c \cdot |S|$.*

THEOREM 4. *For a set S of n grid points in the d-dimensional space, there is a $c(d)$-balanced-separator P^* that contains at most $\leq c'(d)n^{1-\frac{1}{d}}$ points from S, where $0 < c(d) < 1$, $0 < c'(d)$ and both $c(d)$ and $c'(d)$ are constants for a fixed dimensional number d.*

Proof. We will construct a series of sets $S = S_0 \supseteq S_1 \supseteq S_2 \supseteq \cdots \supseteq S_t$ such that $t \leq d-1$ and $|S_i| \geq \frac{1}{2}|S_{i-1}|$ for $i = 1, 2, \ldots, t$. The construction of P^* starts from Stage 0 and can go up to Stage d.

Stage 0: Let $S_0 = S$ and $r = 1$. Enter Stage 1. **End of Stage 0.**

Stage r ($1 \leq r \leq d-1$): Let Q_r contain all of $P_{r,a}$ such that $P_{r,a}$ is a $\frac{3}{4}$-balanced-separator for S_{r-1}. At most $\frac{1}{4}$ of the elements in S_{r-1} with smallest a values (for the rth component) stay to the left of all the $\frac{3}{4}$-balanced separators and at most $\frac{1}{4}$ of the elements in S_{r-1} with largest a values (for the rth component) stay to the right of all the $\frac{3}{4}$-separators. The set $\cup_{P_{r,a} \in Q_r} P_{r,a}$ has at least $\frac{1}{2}$ of the elements from S_{r-1}. Thus, Q_r is not empty. If a $P_{r,a}$ in Q_r contains no more than $n^{1-\frac{1}{d}}$ elements from S, let $P^* = P_{r,a}$ and terminate the construction. We have $|S_{r-1}| \geq \frac{1}{2^{r-1}}|S|$ and

$$|S(r, <a)| \leq |S_{r-1}(r, <a)| + |S - S_{r-1}| \leq \frac{3}{4}|S_{r-1}| + |S| - |S_{r-1}| \tag{1}$$

$$= |S| - \frac{1}{4}|S_{r-1}| \leq \left(1 - \frac{1}{2^{r+1}}\right)|S| \leq \left(1 - \frac{1}{2^d}\right)|S|. \tag{2}$$

Similarly, $|S(r, >a)| \leq (1 - \frac{1}{2^d})|S|$.

If every $P_{r,a} \in Q_r$ has $> n^{1-\frac{1}{d}}$ elements from S, $|Q_r| \leq n^{\frac{1}{d}}$ because $|\cup_{P_{r,a} \in Q_r} (P_{r,a} \cap S)| \leq |S| = n$ and all planes in Q_r are disjoint from each other. It is easy to see that there is an integer interval $[c_1, c_2]$ such that $Q_r = \{P_{r,a} | a \in [c_1, c_2]\}$. Let $S_r = \cup_{P_{r,a} \in Q_r}(P_{r,a} \cap S_{r-1})$. We have $S_r \subseteq S_{r-1}$ and $|S_r| \geq |S_{r-1}|/2$ (because $[c_1, c_2]$ is the set of all the integers a such that $P_{r,a}$ is a $\frac{3}{4}$-balanced-separator). Let $r = r + 1$ and go to the next stage. **End of stage r.**

Stage d: Assume that for each r with $1 \leq r \leq d-1$, Q_r has no plane $P_{r,a}$ with elements $\leq n^{1-\frac{1}{d}}$ from S. Hence, $|Q_r| \leq n^{\frac{1}{d}}$ for $1 \leq r \leq d-1$. If a is fixed, every $p \in P_{r,a}$ has the rth component equal to a. Therefore, $\{x_r | x_r$ is the rth component of some $p \in P_{r,a}$ for some $P_{r,a} \in Q_r\}$ has $\leq n^{\frac{1}{d}}$ elements since $|Q_r| \leq n^{\frac{1}{d}}$ $(1 \leq r \leq d-1)$. This implies that for every $P_{d,a}$,

$$|\{p | p \in P_{r,a_r} \text{ for some } P_{r,a_r} \in Q_r\, (r = 1, \ldots, d-1) \text{ and } p \in P_{d,a}\}| \leq (n^{\frac{1}{d}})^{d-1} = n^{\frac{d-1}{d}}.$$

As $|S_{d-1}| \geq \frac{|S|}{2^{d-1}} = \frac{1}{2^{d-1}}n$, there are at least $\frac{\frac{1}{2}|S_{d-1}|}{n^{1-\frac{1}{d}}} \geq \frac{1}{2^d} \cdot n^{\frac{1}{d}}$ many $P_{d,a}$'s to be $\frac{3}{4}$-balanced-separators for S_{d-1}. One of them has at most $\frac{|S|}{\frac{1}{2^d}n^{\frac{1}{d}}} = 2^d n^{1-\frac{1}{d}}$ elements from S. Let P^* be such a $P_{d,a}$. As $|S_{d-1}| \geq \frac{1}{2^{d-1}}|S|$, we have

$$|S(d, <a)| \leq |S_{d-1}(d, <a)| + |S - S_{d-1}| \leq \frac{3}{4}|S_{d-1}| + |S| - |S_{d-1}| \tag{3}$$

$$= |S| - \frac{1}{4}|S_{d-1}| \leq \left(1 - \frac{1}{2^{d+1}}\right)|S|. \tag{4}$$

Similarly, we also have $|S(d, >a)| \leq (1 - \frac{1}{2^{d+1}})|S|$. **End of stage d.** ∎

For a d-dimensional cube that contains n grid points, its edge length is $n^{\frac{1}{d}}$. Every hyperplane $P_{r,a}$ which intersects the cube shares $n^{\frac{d-1}{d}}$ grid points with the cube. This shows that it is impossible to improve the upper bound on the number of points on the separator to $o(n^{\frac{d-1}{d}})$. In the next section, we shows that we can improve the separator by a constant factor. Theorem 4 indicates that the balanced separator can be found among $O(dn)$ axis-parallel hyperplanes.

3. Sharp separators for grid points. We will improve the quality of the separator obtained in the previous section. The following lemma is a well-known fact (see [29]) that will be used for deriving our new separator. Our reduced upper bound on the number of points on the separator is from the following fact: For a set P of 2D grid points with the centerpoint o (see Lemma 5), a random line through o has the largest expected number of points of P with distance $\leq a$ to the line when the points in P are tightly arranged in the grid points inside a circle with the least radius. This is also true in the higher dimensional space.

LEMMA 5. *For an n-element set P in the d-dimensional space, there is a point q with the property that any half-space that does not contain q covers at most $\frac{d}{d+1}n$ elements of P. (Such a point q is called a centerpoint of P.)*

DEFINITION 6. *For a grid point (i,j) on the 2D plane, its grid square is a 1×1 square with four corner points $(i-\frac{1}{2},j-\frac{1}{2}),(i-\frac{1}{2},j+\frac{1}{2}),(i+\frac{1}{2},j-\frac{1}{2})$, and $(i+\frac{1}{2},j+\frac{1}{2})$. For a 3D grid point (i,j,k), its grid cube is a $1\times 1\times 1$ cube with eight corner points in $\{(i+\alpha,j+\beta,k+\gamma)|\alpha,\beta,\gamma\in\{-\frac{1}{2},\frac{1}{2}\}\}$.*

3.1. 2D separators.

LEMMA 7. (1) *A circle of radius r contains at most $\pi(r+\frac{\sqrt{2}}{2})^2$ grid points.*

(2) *A circle of radius r on the 2D plane has at least $\pi r^2-4\sqrt{2}\pi r$ grid points inside it.*

(3) *A circle of radius $\frac{1}{\sqrt{\pi}}\sqrt{n}+4\sqrt{2}$ has at least n grid points in it.*

(4) *For every line segment L of length m, the number of grid points with distance $\leq a$ to at least one point of L is $\leq(2a+\sqrt{2})(m+2a+\sqrt{2})$.*

(5) *For every line L and fixed $a>0$, there are at most $(2a+\sqrt{2})(\sqrt{2}n+2a+\sqrt{2})$ grid points inside an $n\times n$ square with $\leq a$ distance to L.*

Proof. (1) If a grid point p is inside a circle C of radius r at center o, the 1×1 grid square with center at p is inside a circle C' of radius $r+\frac{\sqrt{2}}{2}$ at the same center o. The number of those 1×1 grid squares for the grid points inside C is no more than the area size of the circle C'.

(2) Let C_1, C, and C_2 be three circles on the plane with the same center. Their radii are $r-\sqrt{2}, r$, and $r+\sqrt{2}$, respectively. Every 1×1 grid square intersecting C's boundary is outside C_1 and inside C_2. The number of grid squares intersecting C's boundary is no more than $\pi(r+\sqrt{2})^2-\pi(r-\sqrt{2})^2=4\sqrt{2}\pi r$.

(3) Let $r=\frac{1}{\sqrt{\pi}}\sqrt{n}+4\sqrt{2}$. It is straightforward to verify that $\pi r^2-4\sqrt{2}\pi r>n$. Apply (2).

(4) If a point p has $\leq a$ distance to L, every point in the 1×1 grid square with center at p has distance $\leq a+\frac{\sqrt{2}}{2}$ to L. The number of those 1×1 squares with centers at points of distance $\leq a$ to L is no more than $2(a+\frac{\sqrt{2}}{2})(m+2a+\sqrt{2})$.

(5) The length of a line L inside an $n\times n$ square is $\leq\sqrt{2}n$. Apply (4). ∎

DEFINITION 8. *Assume that $a>0$ and that p_0,p are two points on the plane. Define $Pr_2(a,p_0,p)$ to be the probability that a point p has $\leq a$ perpendicular distance to a random line L through the point p_0.*

LEMMA 9. *Let $a>0$ be a constant and $\delta>0$ be a small constant. Let P be a set of points in a 2D grid. Assume that all the points of P are inside a circle of radius r with center at point o. For a random line passing through o, the expected number of points in P with distance $\leq a$ to L is bounded by $4ar+\delta r$ for all large r.*

Proof. Assume that $p=(x,y)$ is a point of P and that L is a random line passing through the center $o=(x_0,y_0)$. Let C be the circle of radius r at center o such that C covers all the points in P. Let C' be the circle of radius $r'=r+\frac{\sqrt{2}}{2}$ at the same

center o. It is easy to see that every unit square with center at a point in P is inside C'. The probability that a point p has distance $\leq a$ to L is $\frac{2\arcsin\frac{a}{\text{dist}(o,p)}}{\pi}$.

Let $\epsilon > 0$ be a small constant which will be determined later. Select r_0 to be large enough such that for every point p with $\text{dist}(o,p) \geq r_0$, $\arcsin\frac{a}{\text{dist}(o,p)} < (1+\epsilon)\frac{a}{\text{dist}(o,p)}$ and $\frac{1}{\text{dist}(o,p')} < \frac{1+\epsilon}{\text{dist}(o,p)}$ for every point p' with $\text{dist}(p',p) \leq \frac{\sqrt{2}}{2}$. Let P_1 be the set of all the points p in P such that $\text{dist}(o,p) < r_0$. By Lemma 7, the number of grid points in P_1 is no more than $\pi(r_0 + \frac{\sqrt{2}}{2})^2$. For each point $p \in P_1$, $Pr_2(a,o,p) \leq 1$. For every point $p \in P - P_1$, $Pr_2(a,o,p) = \frac{2\arcsin\frac{a}{\text{dist}(o,p)}}{\pi} \leq \frac{(1+\epsilon)2a}{\pi\text{dist}(o,p)}$.

The expected number of points in P with distance $\leq a$ to a random line through the point o is

$$\sum_{p\in P} Pr_2(a,o,p) = \sum_{p\in P_1} Pr_2(a,o,p) + \sum_{p\in P-P_1} Pr_2(a,o,p) \tag{5}$$

$$\leq \sum_{p\in P_1} 1 + \sum_{p\in P-P_1} \frac{2\arcsin\frac{a}{\text{dist}(o,p)}}{\pi} \tag{6}$$

$$< \pi\left(r_0 + \frac{\sqrt{2}}{2}\right)^2 + \sum_{p\in P-P_1} \frac{(1+\epsilon)2a}{\pi\text{dist}(o,p)} \tag{7}$$

$$\leq \pi\left(r_0 + \frac{\sqrt{2}}{2}\right)^2 + \frac{2a(1+\epsilon)^2}{\pi}\int\int_{C'} \frac{1}{\text{dist}(o,p)} d_x d_y \tag{8}$$

$$= \frac{2a(1+\epsilon)^2}{\pi}\int_0^{2\pi}\int_0^{r'} \frac{\rho}{\rho} d_\rho d_\theta + \pi\left(r_0 + \frac{\sqrt{2}}{2}\right)^2 \tag{9}$$

$$= 4a(1+\epsilon)^2 r' + \pi\left(r_0 + \frac{\sqrt{2}}{2}\right)^2 \tag{10}$$

$$< 4ar + \delta r \text{ for all large } r \text{ by selecting } \epsilon \text{ small enough.} \tag{11}$$

We use the transformation $x = \rho\cos\theta + x_0, y = \rho\sin\theta + y_0$ to convert the integral at (8) to that at (9) above. □

THEOREM 10. *Let $a > 0$ be a constant and $\epsilon > 0$ be a small constant. For a set P of n grid points in a 2D grid, there is a line L such that P has at most $(\frac{4a}{\sqrt{\pi}})\cdot\sqrt{n} + \epsilon\sqrt{n}$ points with distance $\leq a$ to L, and each half-plane divided by L has at most $\frac{2}{3}n$ points from P.*

Proof. Assume that the centerpoint of P is at the point o (see Lemma 5). We are going to estimate the upper bound for the expected number of points in P that have $\leq a$ distances to a random line L through o.

Let $r = \frac{1}{\sqrt{\pi}}\sqrt{n} + 4\sqrt{2}$. By Lemma 7, the circle C at center o with radius r contains at least n grid points. Let f be a one-to-one mapping from P to the set of grid points inside C such that $f(p) = p$ for every $p \in P$ with $\text{dist}(o,p) \leq r$. Therefore, f moves those points of P outside the circle C to the inside. It is easy to see that if $\text{dist}(o,p_1) \leq \text{dist}(o,p_2)$, then, $Pr_2(a,o,p_1) \geq Pr_2(a,o,p_2)$. The expected number of points in P with $\leq a$ distance to L is $\sum_{p\in P} Pr_2(a,o,p)$.

By Lemma 9, $\sum_{p\in P} Pr_2(a,o,p) \leq \sum_{p\in P} Pr_2(a,o,f(p)) \leq 4ar + \delta r = \frac{4a}{\pi}\sqrt{n} + \epsilon\sqrt{n}$ by selecting a small δ. □

3.2. 3D separators. The technology used in the previous section can be easily extended to the 3D grid. We give a brief proof for the case in the 3D space.

LEMMA 11. *Let $a = \sqrt{3}$. (1) A sphere of radius r has at least $\frac{4}{3}\pi r^3 - \frac{4}{3}\pi(6ar^2 + 2a^3)$ grid points. (2) A sphere of radius $(\frac{3}{4\pi})^{\frac{1}{3}} n^{\frac{1}{3}} + 7a$ contains at least n grid points.*

Proof. (1) Let $r_1 = r + a$, and let $r_2 = r - a$. The volume difference between the sphere of radius r_1 and the sphere of radius r_2 is $\frac{4}{3}\pi(6ar^2 + 2a^3)$, which is larger or equal to the number of unit grid cubes intersecting the boundary of the sphere of radius r. (2) For $r = (\frac{3}{4\pi})^{\frac{1}{3}} n^{\frac{1}{3}} + 7a$, we have $\frac{4}{3}\pi r^3 - \frac{4}{3}\pi(6ar^2 + 2a^3) \geq n$. ∎

DEFINITION 12. *Assume that $a > 0$ and that p_0, p are two points in the 3D Euclidean space. Define $Pr_3(a, p_0, p)$ to be the probability that the point p has $\leq a$ perpendicular distance to a random plane L through the point p_0 in the 3D space.*

Assume that both a and p_0 are fixed. We want to compute $Pr_3(a, p_0, p)$, which depends on the parameter a and the distance between p_0 and p. Without loss of generality, we assume that p_0 is the origin point $(0, 0, 0)$ and $p = (x, 0, 0)$, where $x = \text{dist}(p_0, p)$. A random plane through the origin point is uniquely determined by its normal vector (u, v, w) with $u \geq 0$. The distance between p to the plane with normal vector (u, v, w) is equal to xu. If the distance is at most a, then $u \leq \frac{a}{x}$. The set $G_{p,a} = \{(u, v, w) | u^2 + v^2 + w^2 = 1 \text{ and } 0 \leq u \leq \frac{a}{x}\}$ contains all the normal vectors of those planes (through the origin) such that p has distance at most a to each of them. The set $G_{p,a}$ is a subarea of the half-sphere $H_1 = \{(u, v, w) | u^2 + v^2 + w^2 = 1 \text{ and } 0 \leq u\}$ with center at the origin point and radius 1. If a is fixed and x is large, the area size of $G_{p,a}$ can be computed by the formula

$$\int_0^{\frac{a}{x}} 2\pi\sqrt{1 - y^2}\, d_y = \frac{2\pi a}{x} + O\left(\frac{a^2}{x^2}\right).$$

Since the area size of a half-sphere of radius 1 is 2π, the probability that p has distance at most a to a random plane through the origin is

$$Pr_3(a, p_0, p) = \frac{\text{the area size of } G_{p,a}}{\text{the area size of } H_1} = \frac{\frac{2\pi a}{x} + O(\frac{a^2}{x^2})}{2\pi} = \frac{a}{x} + O\left(\frac{a^2}{x^2}\right).$$

The above formula for computing $Pr_3(a, p_0, p)$ corrects a mistake that we made in the extended abstract of this paper [14]. It also gives a slightly different upper bound for the exact algorithm for the folding problem in the 3D space reported in [14].

LEMMA 13. *Let $a > 0$ be a constant and let $\delta > 0$ be a small constant. Let P be a set of points in a 3D grid. Assume that all the points of P are inside a sphere of radius r with center at point o. For a random plane passing through o, the expected number of points in P with distance at most a to L is bounded by $2\pi ar^2 + \delta r^2$ for all large r.*

Proof. The proof is very similar to that of Lemma 9. Let S be the sphere with radius r at center $o = (x_0, y_0, z_0)$ such that it contains all the points in P. Let S' be the sphere of radius $r' = r + \frac{\sqrt{3}}{2}$ at the same center of S. All unit cubes with center at points in P are inside S'.

The expected number of points in P with distance $\leq a$ to a random plane through o is $\sum_{p=(x,y,z)\in P} Pr_3(a, o, p)$, which has the main part $\int\int\int_{S'} \frac{a}{\text{dist}(a,o,p)} d_x d_y d_z$. By the transformation $x = \rho\sin\theta\cos\alpha + x_0, y = \rho\sin\theta\cos\alpha + y_0, z = \rho\sin\theta + z_0$, we have $\int\int\int_{S'} \frac{a}{\text{dist}(a,o,p)} d_x d_y d_z = \int_0^{r'} \int_0^{\pi} \int_0^{2\pi} \frac{a\rho^2 \sin\theta}{\rho} d_\alpha d_\theta d_\rho = 2\pi a r'^2$. ∎

THEOREM 14. *Let $a > 0$ be a constant and let $\epsilon > 0$ be a small constant. For a set P of n points in a 3D grid, there is a plane L such that P has at most*

$(2\pi a(\frac{3}{4\pi})^{2/3}) \cdot n^{2/3} + \epsilon n^{2/3}$ points with distance at most a to L, and each half-space divided by L has at most $\frac{3}{4}n$ points from P.

Proof. By Lemma 11, the sphere of radius $(\frac{3}{4\pi})^{\frac{1}{3}} n^{\frac{1}{3}} + 7\sqrt{3}$ contains at least n grid points. Moving points of P into the sphere, which has center at the centerpoint of P (see Lemma 5), from the outside increases the probability to have distance $\leq a$ to a random plane through the sphere center. By Lemma 13, the expected number of points in P with distance $\leq a$ to a random plane is $(2\pi a(\frac{3}{4\pi})^{2/3}) \cdot n^{2/3} + \epsilon n^{2/3}$ for all large n via selecting small δ. ∎

4. An application of the easy separators to the protein folding problem. We apply the easy separators to the protein folding problem in the HP-model, and obtain the first subexponential time algorithm for it. We have already shown that there is a small set of letters with size $O(n^{1-\frac{1}{d}})$ on a hyperplane to partition the folding problem of n letters into 2 smaller problems of $\leq c(d)n$ letters, where $0 < c(d) < 1$, $c(d)$ is a constant for fixed d, and n is the size of the input (the number of H and P characters). The 2 smaller problems are recursively solved, and their solutions are merged to derive the solution to the original problem. As the separator has only $O(n^{1-\frac{1}{d}})$ letters, there are at most $n^{O(n^{1-\frac{1}{d}})}$ cases to partition the problem.

DEFINITION 15.

- *For a d-dimensional point $(x_1, \cdots, x_d)$, define $||(x_1, \cdots, x_d)|| = \sum_{i=1}^{d} |x_i|$.*
- *For a set Σ of letters, a Σ-sequence is a sequence of letters from Σ. For example, $PHPPHHPH$ is an $\{H, P\}$-sequence. For a sequence S of length n and $1 \leq i \leq n$, $S[i]$ is the ith letter of S and $S[i, j]$ denotes the subsequence $S[i]S[i+1] \cdots S[j]$. If $[i_1, j_1], [i_2, j_2], \cdots, [i_t, j_t]$ are pairwise disjoint intervals inside $[1, n]$, we call $S[i_1, j_1], S[i_2, j_2], \cdots, S[i_t, j_t]$ disjoint subsequences of S. For a set of integers $A = \{i_1 < i_2 < \cdots < i_k\}$, define $S[A] = S[i_1]S[i_2] \cdots S[i_k]$.*
- *A self-avoiding arrangement f for a sequence S of length n in the d-dimensional grid is a one-to-one mapping from $\{1, 2, \cdots, n\}$ to Z^d such that $||f(i) - f(i+1)|| = 1$ for $i = 1, 2, \ldots, n-1$. For the disjoint subsequences $S[i_1, j_1], \cdots, S[i_k, j_k]$ of S, a partial self-avoiding arrangement of S on $S[i_1, j_1], \cdots, S[i_k, j_k]$ is a partial function f from $\{1, 2, \cdots, n\}$ to Z^d such that f is defined on $\cup_{t=1}^{k}[i_t, j_t]$, and f can be extended to a (full) self-avoiding arrangement of S on Z^d.*
- *For a grid self-avoiding arrangement, its contact map is the graph $G_f = (1, 2, \cdots, n, E)$, where the edge set $E = \{(i, j) : |i-j| > 1$ and $||f(i)-f(j)|| = 1\}$.*
- *A rectangular region R in a d-dimensional space is the intersection of a finite number of sets $P_1, P_2, \cdots, P_k$, where $P_i = P_{r,<a}$ or $P_i = P_{r,>a}$ with $1 \leq r \leq d$ and $a \in Z$ for $i = 1, \ldots, k$.*
- *A rectangular region R in a d-dimensional space is of size $m_1 \times m_2 \times \cdots \times m_d$ if $m_i = max\{x_i - x_i' | (x_1, \cdots, x_d), (x_1', \cdots, x_d') \in R\} + 1$ for $i = 1, \ldots, d$.*

As we are going to describe our algorithm recursively, we use the following term to characterize the problem. A d-dimensional *multisequence folding problem* F is formulated as follows.

The inputs are

1. a list of disjoint subsequences $S_1, S_2, \cdots, S_k$ of sequence S_0 ($S_t = S_0[i_t, j_t]$ for $t = 1, \ldots, k$),
2. a rectangular region R, where all of the k $\{H, P\}$-sequences are going to be arranged,

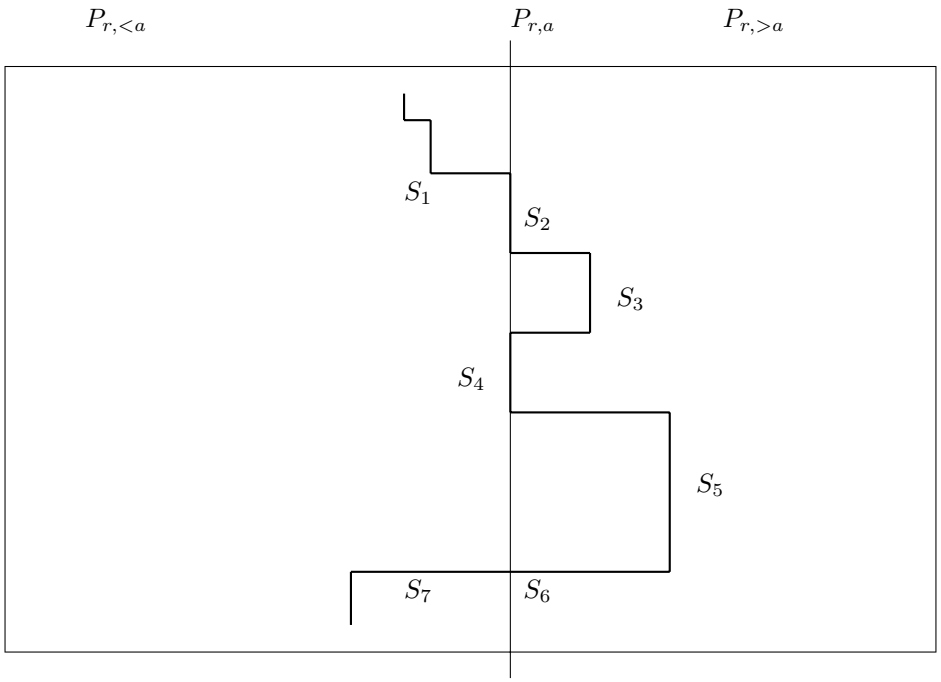

FIG. 2. *The hyperplane $P_{r,a}$ partitions a sequence into 3 groups of disjoint subsequences $\{S_1, S_7\}, \{S_2, S_4, S_6\}$, and $\{S_3, S_5\}$ in $P_{r,<a}, P_{r,a}$, and $P_{r,>a}$, respectively. (Notice that S_6 is a point that is the intersection between the two lines.)*

3. a series of k pairs of points in R: $(p_1, q_1), (p_2, q_2), \cdots, (p_k, q_k)$, in which points $p_t \in R$ and $q_t \in R$ are the positions for putting the first and the last letters of S_t, respectively,
4. a set of available points to put the letters from the k sequences, and
5. a set of $\{H, P\}$ points, which already have letters H and P from $S_0[([1, n] - \cup_{t=1}^{k}[i_t, j_t])]$.

The output is a partial self-avoiding arrangement f of S_0 on $S_1, \cdots, S_k$ in the rectangular region R that satisfies $f(i_t) = p_t, f(j_t) = q_t\,(t = 1, 2, \ldots, k)$ and has the maximal number of H-H contacts, where $f(i)$ is an available point for each $i \in \cup_{t=1}^{k}[i_t, j_t]$. H-H contacts may happen between two available neighbor positions or between an available position and a nonavailable position after the arrangement.

A hyperplane $P_{r,a}$ partitions a multisequence folding problem F into two multisequence folding problems F_1 and F_2 in regions $R \cap P_{r,\leq a}$ and $R \cap P_{r,\geq a}$, respectively, by fixing some letters on $P_{r,a}$ (see Figure 2). Furthermore, the available points of F_1 (F_2, resp.) are the intersection of F's available points with $P_{r,<a}$ ($P_{r,>a}$, resp.).

Algorithm

(a) Input a d-dimensional multisequence folding problem F (as defined above);
(b) For each subset S of $\leq c'(d) \cdot n^{\frac{d-1}{d}}$ letters from $S_1, \cdots, S_k$
For every plane $P_{r,a}$ (with nonempty intersection with R) and
For every arrangement of S at available points on $P_{r,a} \cap R$
(c) Begin
(d) For each partition (by $P_{r,a}$) making F into problems F_1 and F_2 of size $\leq c(d)n$

(e) Begin
(f) Solve F_1 and F_2 recursively using this algorithm (use the brute force method when the problem size is small);
(g) Merge the solutions to F_1 and F_2 to get a potential solution for F;
(h) End
(i) End
(j) Output the solution for F with the maximal number of H-H contacts among all of the potential solutions for F;

End of the Algorithm

LEMMA 16. *There is an $(nm)^{O(n^{1-\frac{1}{d}})}$-time algorithm for the d-dimensional multi-sequence folding problem with an $m_1 \times m_2 \times \cdots \times m_d$ rectangular region in the HP-model, where $m = \max\{\max\{m_i | i = 1, \ldots, d\}, 2\}$ and the dimension d is assumed to be a constant.*

Proof. By Theorem 4, the folding problem is partitioned into two problems with a separator of size $\leq c'(d) \cdot n^{1-\frac{1}{d}}$ elements. For each $1 \leq r \leq d$, we have at most m planes $P_{r,a}$ that have a nonempty intersection with the $m_1 \times m_2 \cdots \times m_d$ rectangular region. There are at most $d \cdot m$ ways to select the separator plane. If the plane has at most t letters, there are at most $d \cdot m \cdot n^t m^{(d-1)t} = dn^t m^{(d-1)t+1}$ ways to select the plane and the position for the letters and put those letters at the selected position on the plane. Thus, the loop (c)–(i) is repeated $\leq dn^t m^{(d-1)t+1}$ times.

For disjoint subsequences $S_1, \cdots, S_k$ of S_0 inside a rectangular region R, we fix $t \leq c'(d) \cdot n^{1-\frac{1}{d}}$ letters from $S_1, \cdots, S_k$ on the hyper plane $P_{r,a}$ and partition them into three groups of subsequences of S_0 which are in $R \cap P_{r,<a}, R \cap P_{r,=a}$, and $R \cap P_{r,>a}$, respectively (see Figure 2). For each subsequence from $R \cap P_{r,<a}$ or $R \cap P_{r,>a}$, we fix the positions for its two endpoints under all possible cases. The subsequences in $R \cap P_{r,<a}$ will not affect those in $R \cap P_{r,>a}$. We have at most 2^{t+1} ways to fix the endpoints of those sequences in $R \cap P_{r,<a}$ and $R \cap P_{r,>a}$. Therefore, the loop (e)–(h) is repeated $\leq 2^{t+1}$ times.

Let $T(m, n)$ be the computational time of our algorithm, where n is the length of S_0 and m is defined as in the lemma. We have the following recursive relationship for the total time of the algorithm:

$$T(m,n) \leq 2 \cdot d \cdot m^{(d-1)c'(d)n^{1-\frac{1}{d}}+1} \cdot n^{c'(d)n^{1-\frac{1}{d}}} \cdot 2^{c'(d)n^{1-\frac{1}{d}}+1} \cdot T(m, c(d)n),$$

where $0 < c(d) < 1$ and $0 < c'(d)$ are constants for fixed d. Expanding the inequality recursively, we have $T(m, n) = (nm)^{O(n^{1-\frac{1}{d}})}$. ∎

THEOREM 17. *There is a $2^{O(n^{1-\frac{1}{d}} \log n)}$-time algorithm for the d-dimensional protein folding problem in the HP-model for fixed d.*

Proof. The folding problem can be put into an $n \times n \times \cdots \times n$ rectangular region in the d-dimensional space by fixing the two middle letters in two central neighbor points in the region. By Lemma 16, we have an $n^{O(n^{1-\frac{1}{d}})} = 2^{O(n^{1-\frac{1}{d}} \log n)}$-time algorithm. ∎

5. Application of the sharp separators to the protein folding problem. In the previous section, we show that the d-dimensional folding problem is computable in $O(2^{e(d)n^{1-\frac{1}{d}}})$ time, where $e(d)$ is constant for fixed d. We will reduce the constant $e(d)$ in this section using the sharp separators.

5.1. 2D folding algorithm. It is easy to see that Theorem 10 implies Theorem 2 by setting $a = \frac{1}{2}$. Assume that our input HP-sequence has n_0 letters and the optimal folding is inside an $m \times m$ square. Select a parameter $\epsilon > 0$. Add some points evenly on the four edges of the $m \times m$ square, so that every two neighbor points have distance $\leq \epsilon$. Those points are called *ϵ-regular points*. Every line segment connecting two ϵ-regular points is called an *ϵ-regular line segment*. An ϵ-regular line is a line containing two ϵ-regular points.

LEMMA 18. *Let $\epsilon > 0$ be a constant. Every line segment L_1 inside the $m \times m$ square has an ϵ-regular segment L_2 such that for every point $p_1 \in L_1$, there is a point $p_2 \in L_2$ with $\mathrm{dist}(p_1, p_2) \leq \epsilon$, and for every point $q_2 \in L_2$, there is a point $q_1 \in L_1$ with $\mathrm{dist}(q_1, q_2) \leq \epsilon$.*

Proof. Assume E_1, E_2, E_3, and E_4 are the four edges of the $m \times m$ square. Assume L_1 intersects two of them inside the square at two points p_i and p_j of edges E_i and $E_j (i \neq j)$, respectively. Select the ϵ-regular point q_i closest to p_i from the edge E_i, and q_j closest to p_j from E_j. The ϵ-regular line segment L_2 results from connecting q_i and q_j. Every point p in L_1 has another point $p' \in L_2$ with distance $\leq \max(\mathrm{dist}(p_i, q_i), \mathrm{dist}(p_j, q_j)) \leq \epsilon$, and every point q in L_2 has another point in $q' \in L_1$ with distance $\leq \max(\mathrm{dist}(p_i, q_i), \mathrm{dist}(p_j, q_j)) \leq \epsilon$. ∎

LEMMA 19. *Let a and ϵ be positive constants. Let P be a set of n points in a 2D grid. There is an ϵ-regular line L such that there are $\leq (\frac{2}{3} + \epsilon)n$ points of P on each of the two half-planes, and $\leq 4(a + \epsilon)\frac{\sqrt{n}}{\sqrt{\pi}}$ points of P with distance $\leq a$ to L.*

Proof. Let $\delta > 0$ be a small constant. By Theorem 10, there is a line L such that the number of points of P with distance $a + \delta$ to it is bounded by $4(a + \delta)\frac{\sqrt{n}}{\sqrt{\pi}}$, and each side of L has at most $\frac{2}{3}n$ points in P. By Lemma 18, there is a line L' close to L such that every point in L has another point in L' with distance $\leq \delta$ and every point in L' has another point in L with distance $\leq \delta$. Every point with distance $\leq a$ to the line L' has distance $\leq a + \delta$ to L. Therefore, the number of points in P with distance $\leq a$ to L' is bounded by $4(a + \epsilon)\frac{\sqrt{n}}{\sqrt{\pi}}$, and each half-plane divided by L has at most $(\frac{2}{3} + \epsilon)n$ points in P if δ is small enough. ∎

LEMMA 20. *For some constants $c_0, \epsilon > 0$, there is an $O(m^{c_0 \log n} n_0^{(6.145-\epsilon)\sqrt{n}})$-time algorithm for the 2D multisequence folding problem F in an $m \times m$ square, where n is the sum of the lengths of the input disjoint subsequences of S_0 and n_0 is the length of S_0.*

Proof. Let $a = 1/2$, $c = 2/3 + \delta$, and $d = \frac{4(a+\delta)}{\sqrt{\pi}}$, where $\delta > 0$ is a small constant which will be fixed later. We assume $m > 1$ and n is large. Let P be an optimal arrangement for the problem F. By Lemma 19, there is a line L such that P has at most $d\sqrt{n}$ points with distance $\leq 1/2$ to L, and each half-plane has at most cn points from P. The letters that stay at those positions with $\leq \frac{1}{2}$ distance to L form a separator for P. For every two letters at different sides of L that have a contact (their distance is 1), at least one of them has distance $\leq \frac{1}{2}$ to L. The algorithm is based on such a separator and is similar to that used in the previous section to find such an optimal solution P.

The number of δ-regular points at every edge of the $m \times m$ square is bounded by $\frac{m}{\delta}$. The total number of δ-regular lines is bounded by $u_1 = \binom{4}{2}(\frac{m}{\delta})^2$. By Stirling's formula, we have $(d\sqrt{n})! > \frac{(d\sqrt{n})^{d\sqrt{n}}}{2^{d\sqrt{n}}}$. There are $u_2 = \binom{n}{0} + \binom{n}{1} + \cdots \binom{n}{d\sqrt{n}} < d\sqrt{n}\frac{n^{d\sqrt{n}}}{(d\sqrt{n})!} < (\frac{2}{d})^{d\sqrt{n}} \cdot d\sqrt{n} \cdot n^{\frac{1}{2}d\sqrt{n}}$ ways to select $\leq d\sqrt{n}$ letters from those n letters among the input disjoint subsequences of S_0.

Assume k $(\leq d\sqrt{n})$ letters $S_0[i_1], S_0[i_2], \cdots, S_0[i_k]$ $(1 \leq i_1 < i_2 < \cdots < i_k \leq n)$ are fixed from the disjoint subsequences of S_0. By Lemma 7, there are at most

$\beta = (2a+\sqrt{2})(\sqrt{2}m+2a+\sqrt{2})$ positions (inside the $m \times m$ square) to put the letter $S_0[i_1]$ such that it has distance $\leq a$ to L.

By Lemma 7, after the letter $S_0[i_j]$ is put at a grid point, there are at most $(2a+\sqrt{2})((i_{j+1}-i_j+2a)+2a+\sqrt{2}) \leq (2a+\sqrt{2})(1+4a+\sqrt{2})(i_{j+1}-i_j)$ ways to arrange $S_0[i_{j+1}]$ so that its position has at most distance a to a point in L. Let $\alpha = (2a+\sqrt{2})(1+4a+\sqrt{2})$. After the position of the letter $S_0[i_1]$ is fixed, there are at most $\prod_{j=1}^{j=k-1}(\alpha(i_{j+1}-i_j))$ ways to put $S_0[i_2], S_0[i_3], \cdots, S_0[i_k]$ along the separation line with distance $\leq a$. Since $k \leq d\sqrt{n}$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq n_0$,

$$\prod_{j=1}^{j=k-1} (\alpha\,(i_{j+1}-i_j)) \leq \left(\alpha\left(\frac{n_0}{k-1}\right)\right)^{k-1} \tag{12}$$

$$\leq \left(\alpha\left(\frac{n_0}{k}\right)\right)^{k} \tag{13}$$

$$\leq \left(\alpha\left(\frac{n_0}{d\sqrt{n}}\right)\right)^{d\sqrt{n}} \tag{14}$$

$$\leq \left(\frac{\alpha}{d}\right)^{d\sqrt{n}} n_0^{d\sqrt{n}} n^{-\frac{1}{2}d\sqrt{n}}. \tag{15}$$

The inequality (12) follows from the well-known fact that for positive variables $y_1, \cdots, y_{k-1}$ and fixed h with $y_1 + \cdots + y_{k-1} \leq h$, the product $\prod_{t=1}^{k-1} y_k$ is maximal when $y_1 = y_2 = \cdots = y_{k-1} = \frac{h}{k-1}$. The number of ways to arrange the k letters along the separation line (with distance $\leq a$ to L) is bounded by

$$u_3 = \beta\left(\frac{\alpha}{d}\right)^{d\sqrt{n}} n_0^{d\sqrt{n}} n^{-\frac{1}{2}d\sqrt{n}}.$$

We have $T(n) \leq u_1 \cdot u_2 \cdot u_3 \cdot T(cn)$. It implies that

$$T(n) \leq \left(\frac{mn}{\delta}\right)^{c_0 \log n} 2^{c_0\sqrt{n}} n_0^{d(\frac{1}{1-\sqrt{c}})\sqrt{n}} = O(m^{c_0 \log n} n_0^{(6.145-\epsilon)\sqrt{n}})$$

by selecting constants ϵ, δ small enough and c_0 large enough. ∎

THEOREM 21. *There is an $O(n^{6.145\sqrt{n}})$-time algorithm for the 2D protein folding problem in the HP-model.*

Proof. Fix the two middle letters at the two central neighbor positions of an $n \times n$ square. Let the folding be inside the $n \times n$ square, and apply Lemma 20. ∎

5.2. 3D folding algorithm. The technology used in the previous section can be easily extended to a 3D grid. We give a brief proof for the case in the 3D space.

Put some regular points on each side of the six faces of an $m \times m \times m$ cube (the folding region) so that every point on each face has $\leq \epsilon$ distance to one regular point. Recall that these points are called ϵ-regular points. Every three ϵ-regular points determine a plane, called an ϵ-regular plane.

LEMMA 22. *Let a and ϵ be positive constants. Let P be a set of n points in a 3D grid. There is an ϵ-regular plane such that there are $\leq (\frac{3}{4}+\epsilon)n$ points on each side of the plane and $2\pi(a+\epsilon)(\frac{3}{4\pi})^{2/3}n^{2/3}$ points with distance at most a to it.*

Proof. Let L be the plane of Theorem 14. Let H be the area of the intersection between plane L and the six faces of the $m \times m \times m$ cube that contains all the points in P. Let p_1 and p_2 be the two points in H with the maximal distance. Let p_3 be the point in H with the largest perpendicular distance to the line p_1p_2. Let p_1', p_2'

and p'_3 be the δ-regular noncollinear points such that p'_i has distance $\leq \delta$ to p_i for $i = 1, 2, 3$. Use the δ-regular plane determined by p'_1, p'_2, and p'_3 (by selecting δ small enough). □

LEMMA 23. *For some positive constant c_0 and $\epsilon > 0$, there exists an $O(m^{c_0 \log n} n^{-6.9128 n^{2/3}} n_0^{(13.8258-\epsilon)n^{2/3}})$-time algorithm for the 3D multisequence folding problem in an $m \times m \times m$ cube, where n is the sum of the lengths of the input disjoint subsequences of S_0 and n_0 is the length of S_0.*

Proof. Let $a = 1/2$, $c = 3/4 + \delta$, and $d = 2\pi(a + \delta)(\frac{3}{4\pi})^{2/3}$. As in the proof of Lemma 20, let $u_1 = \binom{8}{3}(\frac{m}{\delta})^6$, let $u_2 = (\frac{2}{d})^{dn^{2/3}} \cdot dn^{2/3} \cdot n^{\frac{1}{3}dn^{2/3}}$, and let $u_3 = \beta'(\frac{\alpha'}{d})^{2dn^{\frac{2}{3}}} n^{-\frac{4}{3}dn^{\frac{2}{3}}} n_0^{2dn^{\frac{2}{3}}}$, where α' and β' are similar to the α and β in the proof of Lemma 20. We have $T(n) \leq u_1 \cdot u_2 \cdot u_3 \cdot T(cn)$. This implies that $T(n) = (mn)^{c_0 \log n} 2^{c_0 n^{\frac{2}{3}}} n^{-\frac{d}{(1-c^{2/3})}n^{2/3}} n_0^{\frac{2d}{(1-c^{2/3})}n^{2/3}}$ for some constant $c_0 > 0$. □

THEOREM 24. *There is an $O(n^{6.913 n^{2/3}})$-time algorithm for the 3D protein folding problem in the HP-model.*

Proof. Fix the two middle letters at the two central neighbor positions of an $n \times n \times n$ cube. Let the folding be inside the $n \times n \times n$ cube, and apply Lemma 23. □

6. Conclusions. We develop an efficient method to obtain an effective separator for a set of grid points. For a set of 2D (3D) grid points, the separator is controlled by a line (plane, resp.) L and a distance a to L. The region of the separator consists of those points with distance at most a to L. The distance parameter a provides us with a flexible way to control the width of the separator region. The separators are used in obtaining a subexponential time algorithm for the protein folding problem in the HP-model. Using the linear structure of the separator, we can find the approximate separator region by checking $O(n^2)$ ($O(n^6)$) possible locations in developing the algorithm for the 2D (3D, resp.) protein folding problem in the HP-model. These algorithms for the protein folding problem have a nontrivial upper bound, but they are not practical enough for implementation. The separators developed in this paper have been found to have more applications in a series of recent papers [11, 7, 13, 12].

Acknowledgments. We are grateful to Mahdi Abdelguerfi, Padmanabhan Mahadevan, and Seth Pincus for helpful discussions during this research. We thank the anonymous referees for helpful comments and for pointing out an error in the earlier version. We are also grateful to Zhixiang Chen and Ming Li for their suggestions for improving the presentation of this paper. The first author would also like to thank Chanda Yadavalli for introducing him to the area of bioinformatics.

REFERENCES

[1] R. AGARWALA, S. BATZOGLOU, V. DANCIK, S. DECATUR, S. HANNENHALLI, M. FARACH, M. MUTHUKRISHNAN, AND S. SKIENA, *Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model*, J. Comput. Biol., 21 (1997), pp. 275–296.

[2] N. ALON, P. SEYMOUR, AND R. THOMAS, *Planar separators*, SIAM J. Discrete Math., 7 (1994), pp. 184–193.

[3] N. ALON, P. SEYMOUR, AND R. THOMAS, *A separator theorem for graphs with an excluded minor and its applications*, in Proceedings of the 22nd Annual ACM Symposium on Theory of Computing, 1990, pp. 293–299.

[4] R. BACKOFEN, *Constraint techniques for solving the protein structure prediction problem*, in Proceedings of the 4th International Conference on Principles and Practice of Constraint Programming, Lecture Notes in Comput. Sci. 1520, Springer, London, 1998, pp. 72–86.

[5] U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, and W. Nadler, *Testing a new Monte Carlo algorithm for protein folding*, Proteins: Structure, Function, and Genetics, 32 (1998), pp. 52–66.

[6] B. Berger and T. Leighton, *Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete*, J. Comput. Biol., 5 (1998), pp. 27–40.

[7] Z. Chen, B. Fu, Y. Tang, and B. Zhu, *A PTAS for a disc covering problem using width-bounded separators*, J. Comb. Optim., (2006), pp. 203–217.

[8] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis, *On the complexity of protein folding*, J. Comput. Biol., 5 (1998), pp. 423–465.

[9] H. N. Djidjev, *On the problem of partitioning planar graphs*, SIAM J. Alg. Disc. Meth., 3 (1982), pp. 229–240.

[10] H. N. Djidjev and S. M. Venkatesan, *Reduced constants for simple cycle graph separation*, Acta Inform., 34 (1997), pp. 231–234.

[11] B. Fu, *Theory and application of width bounded geometric separator*, in Proceedings of the 23rd International Symposium on Theoretical Aspects of Computer Science (STACS'06), Lecture Notes in Comput. Sci. 3884, Springer, Berlin, 2006, pp. 277–288.

[12] B. Fu and Z. Chen, *Sublinear-time algorithms for width-bounded geometric separators and their applications to protein side-chain packing problems*, in Proceedings of the Second International Conference on Algorithmic Aspects in Information and Management, Lecture Notes in Comput. Sci. 4041, Springer, Berlin, pp. 49–160.

[13] B. Fu, S. Oprisan, and L. Xu, *Multi-directional width-bounded geometric separator and protein folding*, in Proceedings of the 16th Annual International Symposium on Algorithms and Computation, Lecture Notes in Comput. Sci. 3827, Springer, Berlin, 2005, pp. 995–1006.

[14] B. Fu and W. Wang, *A* $2^{O(n^{1-1/d}\log n)}$ *time algorithm for d-dimensional protein folding in the HP-model*, in Proceedings of the 31st International Colloquium on Automata, Languages and Programming, Lecture Notes in Comput. Sci. 3142, Springer, Berlin, 2004, pp. 630–644.

[15] H. Gazit, *An Improved Algorithm for Separating a Planar Graph*, manuscript, University of Southern California, Los Angeles, 1986.

[16] J. R. Gilbert, J. P. Hutchinson, and R. E. Tarjan, *A separation theorem for graphs of bounded genus*, J. Algorithms, 5 (1984), pp. 391–407.

[17] W. E. Hart and S. Istrail, *Fast protein folding in the hydrophobic-hydrophilic model within three-eights of optimal*, in Proceedings of the 27th ACM Symposium on the Theory of Computing, 1995, pp. 157–168.

[18] M. Khimasia and P. Coveney, *Protein structure prediction as a hard optimization problem: The genetic algorithm approach*, Molecular Simulation, 19 (1997), pp. 205–226.

[19] N. Krasnogor, D. Pelta, P. E. Martinez Lopez, and E. De La Canal, *Genetic algorithms for the protein folding problem: A critical view*, in Engineering of Intelligent Systems, C. Fyfe and E. Alpaydin, eds., International Computer Science Conventions, ICSC Academic Press, Canada/Switzerland, 1998, pp. 353–360.

[20] K. F. Lau and K. A. Dill, *A lattice statistical mechanics model of the conformational and sequence spaces of proteins*, Macromolecules, 1989, pp. 3986–3997.

[21] K. F. Lau and K. A. Dill, *Theory for protein mutability and biogenesis*, Proc. Natl. Acad. Sci. USA, 87 (1990), pp. 638–642.

[22] F. Liang and W. Wong, *Evolutionary Monte Carlo for protein folding simulations*, J. Chem. Phys., 115 (2001), pp. 3374–3380.

[23] D. Lichtenstein, *Planar formulae and their uses*, SIAM J. Comput., 11 (1982), pp. 329–343.

[24] R. J. Lipton and R. E. Tarjan, *A separator theorem for planar graphs*, SIAM J. Appl. Math., 36 (1979), pp. 177–189.

[25] R. J. Lipton and R. E. Tarjan, *Applications of a planar separator theorem*, SIAM J. Comput., 9 (1980), pp. 615–627.

[26] G. L. Miller, S.-H. Teng, and S. A. Vavasis, *A unified geometric approach to graph separators*, in 32nd Annual Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos, CA, 1991, pp. 538–547.

[27] G. L. Miller and W. Thurston, *Separators in two and three dimensions*, in Proceedings of the 22nd Annual ACM Symposium on Theory of Computing, 1990, pp. 300–309.

[28] A. Newman, *A new algorithm for protein folding in the HP model*, in Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA, 2002, pp. 876–884.

[29] J. Pach and P. Agarwal, *Combinatorial Geometry*, John Wiley & Sons, New York, 1995.

[30] A. Patton, W.P.III, and E. Goldman, *A standard ga approach to native protein conformation prediction*, in Proceedings of the 6th International Conference on Genetic Algorithms, Morgan Kaufmann, 1995, pp. 574–581.

[31] A. Piccolboni and G. Mauri, *Application of evolutionary algorithms to protein folding prediction*, in Artificial Evolution, Third European Conference 1997, Lecture Notes in Comput. Sci. 1363, Springer, London, pp. 123–136.

[32] S. Plotkin, S. Rao, and W. D. Smith, *Shallow excluded minors and improved graph decompositions*, in Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, Arlington, VA, SIAM, 1994, pp. 462–470.

[33] A. Rabow and H. Scheraga, *Improved genetic algorithm for the protein folding problem by use of a Cartesian combination operator*, Protein Sci., 5 (1996), pp. 1800–1815.

[34] R. Ramakrishnan, B. Ramachandran, and J. Pekney, *A dynamic Monte Carlo algorithm for exploration of dense conformation spaces in heteropolymers*, J. Chem. Phys., 106 (1997), pp. 2418–2425.

[35] S. S. Ravi and H. B. Hunt, III, *Application of the planar separator theorem to computing problems*, Inform. Process. Lett., 25 (1987), pp. 317–322.

[36] A. Sali, E. Shakhnovich, and M. Karplus, *How does a protein fold?*, Nature, 369 (1994), pp. 248–251.

[37] W. D. Smith and N. C. Wormald, *Geometric separator theorems and applications*, in Proceedings of the 39th Annual IEEE Symposium on Foundations of Computer Science, 1998, pp. 232–243.

[38] D. A. Spielman and S. H. Teng, *Disk packings and planar separators*, in Proceedings of the 12th Annual ACM Symposium on Computational Geometry, 1996, pp. 349–358.

[39] U. Unger and J. Moult, *Genetic algorithm for 3D protein folding simulations*, in Proceedings of the 5th International Conference on Genetic Algorithms, 1993, pp. 581–588.

[40] U. Unger and J. Moult, *Genetic algorithms for protein folding simulations*, J. Mol. Biol., 231 (1993), pp. 75–81.

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