Sublinear Time Width-Bounded Separators and Their Application to the Protein Side-Chain Packing Problem ⋆

Bin Fu[1,2] and Zhixiang Chen[3]

[1] Dept. of Computer Science, University of New Orleans, LA 70148, USA.
[2] Research Institute for Children, 200 Henry Clay Avenue
New Orleans, LA 70118. fu@cs.uno.edu
[3] Dept. of Computer Science, University of Texas - Pan American
TX 78539, USA. chen@cs.panam.edu

Abstract. Given $d > 2$ and a set of n grid points Q in $\Re^d$, we design a randomized algorithm that finds a w-wide separator, which is determined by a hyper-plane, in $O(n^{\frac{2}{d}} \log n)$ sublinear time such that Q has at most $(\frac{d}{d+1} + o(1))n$ points one either side of the hyper-plane, and at most $c_d w n^{\frac{d-1}{d}}$ points within $\frac{w}{2}$ distance to the hyper-plane, where c_d is a constant for fixed d. In particular, $c_3 = 1.209$. To our best knowledge, this is the first sublinear time algorithm for finding geometric separators. Our 3D separator is applied to derive an algorithm for the protein side-chain packing problem, which improves and simplifies the previous algorithm of Xu [26].

1 Introduction

The work in this paper aims for efficient identification of width-bounded separators for a given set of points in the d-dimensional Euclidean space and their applications to intractable practical problems. Intuitively, a width-bounded separator utilizes a simple structured hyper-plane to divide the set into two *"balanced"* subsets, while at the same time maintaining a *"low density"* of the set within a given distance to the hyper-plane. This new notion of separators was initially introduced by Fu in [11], and it was shown that these separators are very suitable in solving a number of distance-bounded geometric problems such as the protein folding problem in the HP model in [10] and some other intractable problems in [11, 6].

The main contributions of this paper are summarized as follows:

In section 5, we present an $O(n^{\frac{2}{d}} \log n)$ sublinear time randomized algorithm for finding a with-bounded separator in the dimensioinal Euclidean space $\Re^d$ for $d > 2$. To our best knowledge, this is the first sublinear time algorithm for finding geometric separators. For many other geometric problems, a higher

⋆ This research is supported by Louisiana Board of Regents fund under contract number LEQSF(2004-07)-RD-A-35, and in part by NSF Grant CNS-0521585.

dimension brings higher computational complexity. However, it is interesting to notice that the exponent of our algorithm's computational complexity is reversely propotional to the dimension of the space.

In section 6, we exhibit an application of our sublinear time separator to the protein side-chain packing problem. One of the most fundamental problems in the molecular biology is to predict a protein's 3D structure when given its 1D amino-acid sequence. Although much effort has been made for decades, this problem remains unsolved. An important component of the general protein structure prediction problem is the protein side-chain packing problem. It determines the side-chain positions onto the fixed backbone [23]. This problem has been proved to be NP-complete [1]. Recently, a $r_{ave}^{O(n^{\frac{2}{3}}\log n)}$ time algorithm was shown by Xu [26], where r_{ave} is the average number of side-chain rotamers in a protein. We apply width-bounded separators to the protein side-chain packing problem. The length of side-chain of each amino acid is small compared to the size of one protein. Two side-chains in a protein molecular do not interact with each other if their distance is slightly larger than the sum of their lengths according to models used in (e.g. [4, 5, 26]). Using our width-bounded separators, we obtain an algorithm with computational time $r_{\max}^{O(n^{\frac{2}{3}})}$, where $r_{\max}$ is the maximal number of side-chain rotamers among a protein. Since the number of rotamers is usually small, we assume both r_{ave} and r_{max} are constants, hence our new algorithm has a better complexity bound.

2 The Related Work

There have been extentive efforts on finding separators due to their critical roles in many issues of algorithm design and analysis. Because of space limit we cannot give a comprehensive review of the related work but list some representative results in this area. Lipton and Tarjan [16] proved that every n vertex planar graph has at most $\sqrt{8n}$ vertices whose removal separates the graph into two disconnected parts of size at most $\frac{2}{3}n$. Their $\frac{2}{3}$-separator has been improved by a series of papers [7, 12, 2, 8] with the best record $1.97\sqrt{n}$ by Djidjev and Venkatesan [8]. Spielman and Teng [25] showed a $\frac{3}{4}$-separator with size $1.82\sqrt{n}$ for planar graphs. Separators for more general graphs were derived in [13, 3, 22]. A planar graph can be induced by a set of non-overlapping discs on the plane such that every vertex corresponds to a disc center and each edge corresponds to a tangent relationship between two discs. The separator developed by Miller, Teng and Vavasis [17] is a generalization of planar graph separators to the d-dimensional Euclidean space. Some $O(\sqrt{k \cdot n})$ size separators for k-thick systems and the related algorithms were derived in [18, 19, 17, 24].

The study of width-bounded separators were initiated by Fu in [11] and has yielded successful applications in [10, 6]. Our width-bounded geometric separator has some interesting advantages over previous geometric separators such as the popular geometric separator by Miller, Teng and Vavasis [17]. First, the width-bounded separator has a simple linear structure as the separator is determined by a hyper-plane and a width parameter w, but Miller *et al.*'s separator is a sphere, which can be also found in linear time [9]. The linear structure is very crucial for us in deriving sublinear time algorithm in this paper. Second. the width-bounded separator has a smaller constant in its size upper bound factor than other separators. The constant factor was not clearly given in Miller *et al.*'s separator. Furthermore, their separator only has a balance condition bounded by $\frac{d+1}{d+2}n$ due to their transformation to a higher dimension, while The balance condition of the width-bounded separator is bounded by $\frac{d}{d+1}n$. Third, the width-bounded separator can be used to deal with an arbitrary set of points via using a set of grid points and weights to characterize the distribution of points from the input set.

3 Notations, Definitions, and Width-Bounded Separators

For any finite set A, $|A|$ denotes the number of elements in A. Let $\Re$ be the set of all real numbers. For two points p_1, p_2 in the d-dimensional Euclidean space $\Re^d$, $\mathrm{dist}(p_1, p_2)$ is the Euclidean distance between p_1 and p_2. For a set $A \subseteq \Re^d$, $\mathrm{dist}(p_1, A) = \min_{q\in A} \mathrm{dist}(p_1, q)$. The *diameter* of any $P \subseteq \Re^d$ is $\max_{p_1,p_2\in P} \mathrm{dist}(p_1, p_2)$. For $a > 0$ and a set A of points in $\Re^d$, if the distance between every two points in A is at least a, then A is called *a-separated.* For $\epsilon > 0$ and a set Q of points in $\Re^d$, an *ϵ-net* of Q is another set P of points in $\Re^d$ such that each point in Q has a distance $\le \epsilon$ to some point in P. We say P is a net of Q if P is an ϵ-net of Q for some constant $\epsilon > 0$ (that does not necessarily depend on the size of Q). A net set is usually a 1-separated set such as a grid point set. A weight function $w : P \to [0, \infty)$ is often used to measure the density of Q near each point in P. Let $f : \Re^d \to \Re$ be a smooth function. Its *surface* is the set $L(f) = \{v \in \Re^d | f(v) = 0\}$. A *hyper-plane* in $\Re^d$ through a fixed point $p_0 \in \Re^d$ is defined by the equation $(p - p_0) \cdot v = 0$, where v is a normal vector of the plane and " ." is the usual vector inner product. A hyper-plane in $\Re^d$ is determined by $L(f)$ for some linear function $f : \Re^d \to \Re$.

Definition 1. *Given any $Q \subseteq \Re^d$ with a net $P \subseteq \Re^d$, a constant $a > 0$, and a weight function $w : P \to [0, \infty)$, an a-wide-separator is determined by the surface $L(f)$ for some linear function $f : \Re^d \to \Re$. The separator has two measurements for its quality of separation: (1)* $\mathrm{balance}(L(f), Q) = \frac{\max(|Q_1|,|Q_2|)}{|Q|}$, *where $Q_1 = \{q \in Q | f(q) < 0\}$ and $Q_2 = \{q \in Q | f(q) > 0\}$; and (2)* $\mathrm{density}(L(f), P, \frac{a}{2}, w)$, *where in general* $\mathrm{density}(A, P, x, w) = \sum_{p\in P, \mathrm{dist}(p,A)\le x} w(p)$ *for any $A \subseteq \Re^d$ and $x > 0$. When f is fixed or no confusion arises, we use* $\mathrm{balance}(L, Q)$ *and* $\mathrm{density}(L, P, \frac{a}{2}, w)$ *to stand for* $\mathrm{balance}(L(f), Q)$ *and* $\mathrm{density}(L(f), P, \frac{a}{2}, w)$*, respectively.*

Definition 2. *A (b, c)-partition of $\Re^d$ divides the space into a disjoint union of regions $P_1, P_2, \ldots,$ such that each P_i, called a regular region, has a volume of b and a diameter $\le c$. A (b, c)-regular point set A is a set of points in $\Re^d$ with a (b, c)-partition $P_1, P_2, \ldots,$ such that each P_i contains at most one point from A.*

For two regions A and B, if $A \subseteq B$ ($A \cap B \neq \emptyset$), we say B contains (intersects resp.) A.

Let $B_d(r,o)$ be the d-dimensional ball of radius r at center o. Its volume is $V_d(r) = \frac{2^{(d+1)/2}\pi^{(d-1)/2}}{1\cdot 3 \cdots (d-2)\cdot d} r^d$ if d is odd, or $\frac{2^{d/2}\pi^{d/2}}{2\cdot 4 \cdots (d-2)\cdot d} r^d$ otherwise. Let $V_d(r) = v_d \cdot r^d$, where v_d is a constant for the fixed dimension d. In particular, $v_1 = 2, v_2 = \pi$ and $v_3 = \frac{4\pi}{3}$. We will use the following well-known fact that can be easily derived from Helly Theorem (see [21]).

Lemma 1. *For an n-element set P in the d-dimensional space $\Re^d$, 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 3. *Let $a > 0$, p and o be two points in $\Re^d$. Define $Pr_d(a, p_0, p)$ to be the probability that the point p has $\leq$ a perpendicular distance to a random hyper-plane L through the point p_0. Define function $f_{a,p,o}(L) = 1$ if p has a distance $\leq a$ to the hyper-plane L through o, or 0 otherwise. The expectation of function $f_{a,p,o}(L)$ is $E(f_{a,p,o}(L)) = Pr_d(a, o, p)$. Assume $P = \{p_1, p_2, \ldots, p_n\}$ is a set of n points in $\Re^d$ and each p_i has weight $w(p_i) \geq 0$. Define function $F_{a,P,o}(L) = \sum_{p\in P} w(p) f_{a,p,o}(L)$.*

We give an upper bound for the expectation $E(F_{a,P,o}(L))$ for $F_{a,P,o}(L)$ in the lemma below.

Lemma 2. *[11] Let $d \geq 2$. Let o be a point in $\Re^d$, $a, b, c > 0$ be constants and $\epsilon, \delta > 0$ be small constants. Assume that $P_1, P_2, \ldots,$ form a (b,c)-partition for $\Re^d$, and the weights $w_1 > \cdots > w_k > 0$ satisfy $k \cdot \max_{i=1}^{k}\{w_i\} = O(n^{\epsilon})$. Let P be a set of n weighted (b,c)-regular points in a d-dimensional plane with $w(p) \in \{w_1, \ldots, w_k\}$ for each $p \in P$. Let n_j be the number of points $p \in P$ with $w(p) = w_j$ for $j = 1, \ldots, k$. We have $E(F_{a,P,o}(L)) \leq (k_d \cdot (\frac{1}{b})^{\frac{1}{d}} + \delta) \cdot a \cdot \sum_{j=1}^{k} w_j \cdot n_j^{\frac{d-1}{d}} + O(n^{\frac{d-2}{d}+\epsilon})$, where $k_d = \frac{d \cdot h_d}{d-1} \cdot v_d^{\frac{1}{d}}$ with $h_d = \frac{2(d-1)v_{d-1}}{d \cdot v_d}$. In particular, $k_2 = \frac{4}{\sqrt{\pi}}$ and $k_3 = \frac{3}{2}\left(\frac{4\pi}{3}\right)^{\frac{1}{3}}$.*

Definition 4. *Let $a_1, \ldots, a_d > 0$ be positive constants. A $(a_1, \ldots, a_d)$-grid regular partition divides $\Re^d$ into a disjoint union of $a_1 \times \cdots \times a_d$ rectangular regions. A $(a_1, \ldots, a_d)$-grid (regular) point is a corner point of a rectangular region. Under certain translation and rotation, each $(a_1, \ldots, a_d)$-grid regular point is represented as $(a_1 t_1, \ldots, a_d t_d)$ for some integers $t_1, \ldots, t_d$. For a point $p = (x_1, \ldots, x_d) \in \Re^d$, if $x_1, \ldots, x_d$ are all integers, then p is simply called a grid point (it is a $(1, \ldots, 1)$-grid regular point). For each point q and a hyper-plane L in $\Re^d$, define $sd(q, L)$ to be the signed distance from q to L, which is $sd(q, L) = (q - q_0) \cdot v_L$, where q_0 is a point on L, and v_L is the normal vector of the plane L with the first nonzero coordinate to be positive.*

For a hyper-plane L in $\Re^d$, if L is through a point q_0 and has the normal vector v, then it has linear equation $(u - q_0) \cdot v = 0$. If $q \in \Re^d$ and $d = sd(q, L)$, then the hyper-plane L' through q and parallel to L has equation $(u-(q_0+dv))\cdot v=0$. We use $L(d)$ to represent such a hyper-plane L'.

For an interval $I \subseteq R$, $\|I\|$ is the length of I. For example, $\|[a,b)\| = b-a$. We often use $Pr(E)$ to represent the probability of an event E. For a real number x, $\lfloor x \rfloor$ is the largest integer $y \le x$, and $\lceil x \rceil$ are the least integer $z \ge x$. For an interval $[a,b] \subseteq R$, define $center([a,b])$ to be $\frac{a+b}{2}$.

Lemma 3. *Let P be a finite set of points in $\Re^d$ and q_0 be a fixed point in $\Re^d$. Then for a random hyper-plane L through q_0, $Pr(sd(p_1,L)=sd(p_2,L)$ for $p_1,p_2 \in P$ with $p_1 \neq p_2)=0$.*

Proof. A random hyper-plane L through a fixed point q_0 can be characterized by the equation $(q-q_0)\cdot v_L=0$, where v_L is the normal vector of L. Each unit vector can be considered as a point of the surface of the unit ball $B_d(1,o)$, where $o=(0,\ldots,0)$ is the origin point. The surface area size of $B_d(r,o)$ is equal to $\frac{dV_d(r)}{d_r}=dv_d r^{d-1}$. The surface area of $B_d(r,o)$ is of dimension $d-1$.

For two fixed points p_1 and p_2, if $sd(p_1,L)=sd(p_2,L)$, then $(p_1-q_0)\cdot v_L = (p_2-q_0)\cdot v_L$. It implies that $(p_1-p_4)\cdot v_L=0$. Consider the sub-area on the surface of $B(1,o)$: $\{v|(p_1-p_2)\cdot v=0$ and $v\cdot v=1\}$, which is the intersection between a plane $(p_1-p_2)\cdot v=0$ and $B_d(1,o)$, and is of dimension $d-2$. It is easy to see that it has area size 0 in the d-dimensional space. The lemma follows since the union of a finite number of areas of area size 0 still has 0 area size. □

4 An Overview of Our Techniques

Given any set Q of points in R^d with a net P, the idea of our techniques for finding an a-width-bound separator is to transform the problem from the d-dimensional space to the 1-dimensional space. By Lemma 1 and Lemma 2, we can see the existence of a hyper-plane that satisfies both the balance and the density conditions. Lemma 2 gives an upper bound on the expectation of $F_{a,P,o}(L)$. By Markov's inequality, $Pr(F_{a,P,o}(L) > (1+\alpha)E(F_{a,P,o}(L))) \le \frac{1}{1+\alpha}$. Thus, a random hyper-plane L has probability $\ge 1-\frac{1}{1+\alpha}=\frac{\alpha}{1+\alpha}$ that $F_{a,P,o}(L) \le (1+\alpha)E(F_{a,P,o}(L))$. The chance is amplified if we repeat the random selection of the hyper-plane L multiple times.

Let $n_P=|P|$ and $n_Q=|Q|$. After a hyper-plane L is fixed, we try to find another hyper-plane L' that is parallel to L. We want L' to guarantee the desired balance and density conditions. To do so, we compute signed distances for all the points in Q and P to the hyper-plane L. Those signed distances are all different for the points in Q and, respectively, for the points in P (by Lemma 3). These signed distances are all in the 1-dimensional real axis, and finding L' can be done via finding a "right position" among these distances, hence this transforms the problem from the d-dimensional space into to the 1-dimensional space as follows: Find the interval $[D_{1,d+1},D_{d,d+1}]$ such that both the left side $(-\infty,D_{1,d+1})$ and the right side $(D_{d,d+1},+\infty)$ have roughly $\frac{n_Q}{d+1}$ signed distances from Q to L. So, every hyper-plane L' (parallel to L) with a signed distance in $[D_{1,d+1},D_{d,d+1}]$

to L guarantees the balance condition. For an interval I, we compute its weight as the sum of the weights of the points of P with their signed distances in I. We then look for an interval $[x-a, x+a]$ that has $x \in [D_{1,d+1}, D_{d,d+1}]$ and the smallest weight. Finally, we let L' be a hyper-plane with a signed distance x to L. The balance boundaries $D_{1,d+1}$ and $D_{d,d+1}$ can be detected by sampling a small number of points from Q. Using the Chernoff bound, we have a high probability that there is a small fraction difference from the exact boundaries. Similarly, the desired interval can be also detected by sampling a small number of points from P.

5 The Sublinear Time Randomized Algorithm

We use the following well-known Chernoff bound (see [20] for a proof) and simplied version in Lemma 4.

Theorem 1. *[20] Let $X_1, \cdots, X_n$ be n independent random $0,1$ variables, where X_i takes 1 with probability p_i. Let $X = \sum_{i=1}^{n} X_i$, and $\mu = E[X]$. Then for any $\delta > 0$, (1) $Pr(X < (1-\delta)\mu) < e^{-\frac{1}{2}\mu\delta^2}$, and (2)$Pr(X > (1+\delta)\mu) < \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\mu}$.*

Lemma 4. *Let $X_1, \cdots, X_n$ be n independent random $0,1$ variables, where X_i takes 1 with probability p. Let $X = \sum_{i=1}^{n} X_i$. Then for any $\frac{1}{3} > \epsilon > 0$, (1) $Pr(X < pn - \epsilon n) < e^{-\frac{1}{2}n\epsilon^2}$, and (2)$Pr(X > pn + \epsilon n) < e^{-\frac{1}{3}n\epsilon^2}$.*

Theorem 2. *Let $d \geq 2$ be the fixed dimension number and v be a positive parameter. Let $a, b, c > 0$ be constants and $\delta, s_1, s_2 > 0$ be small constants. Let Q be another set of n_Q points in $\Re^d$, and P be a set of n_P (b, c)-regular points, which form a net for Q. Let $w_1 > w_2 \cdots > w_k > 0$ be positive weights with $k \cdot w_1 = O(n_P^{s_1})$, $\frac{w_1}{w_k} = o(n_P^{\frac{1}{d}})$, $\frac{k}{w_k} = O(n_P^{s_2})$, and w be a mapping from P to $\{w_1, \cdots, w_k\}$. There exists an $O(v^2 \cdot (n_P^{\frac{2}{d}+2(s_1+s_2)} \cdot \log n_P + \log n_Q))$ time randomized algorithm to find a hyper plane M with probability $\geq 1 - \frac{1}{2^v}$ such that (1) each half space has $\leq (\frac{d}{d+1} + \delta) n_Q$ points from Q, and (2) $\sum_{p \in P \text{ and } \mathrm{dist}(p,M) \leq a} w(p) \leq \left(k_d \cdot b^{\frac{-1}{d}} + \delta\right) \cdot a \cdot \sum_{j=1}^{k} w_j n_j^{\frac{d-1}{d}} + O(n_P^{\frac{d-2}{d}+s_1})$ for all large n_P, where $n_j \geq 1$ is the number of points $p \in P$ with $w(p) = w_j$ $(j = 1, \cdots, k)$.*

Proof. We use two phases to find the separator hyper-plane. The first phase determines the orientation of the hyper-plane by selecting a random hyper-plane, and finds the region of the separator hyper-plane for a balanced partition. The second phase finds the position of the separator plane with a small sum of weights for the points of the set P close to it. Without loss of generality, we assume that $0 < \delta < 1$. Since $n_j \geq 1 (j = 1, \cdots, k)$, we have $k \leq n_P$. Let $b = \prod_{i=1}^{d} a_i$. Select constant $c_0 > 0$ and let $\delta_1 = c_0 \delta$ so that $(k_d \cdot b^{\frac{-1}{d}} + 3\delta_1)(1+\delta_1)^2 \leq (k_d \cdot b^{\frac{-1}{d}} + \frac{\delta}{2})$. Let $a_1 = a(1 + \delta_1)$ and $\alpha = \delta_1$. Let c_1 be a constant such that

$$k \cdot w_1 \leq c_1 n_P^{s_1} \text{ and } \frac{k}{w_k} \leq c_1 n^{s_2}. \tag{1}$$

Let o be the center point from Lemma 1 (our algorithm does not need to find such a center point o, but will use its existence). By Lemma 2, $E(F_{a_1,P,o}) \leq (k_d \cdot b^{\frac{-1}{d}} + \delta_1) \cdot a_1 \cdot \sum_{j=1}^{k} w_j n_j^{\frac{d-1}{d}} + O(n_P^{\frac{d-2}{d}+s_1})$. By Markov inequality, $Pr(F_{a_1,P,o}(L) \geq (1+\alpha)E(F_{a_1,P,o})) \leq \frac{1}{1+\alpha}$. This tells us that a random hyper-plane L has the probability at least $1 - \frac{1}{1+\alpha}$ such that there exists a separator hyper-plane L' (it may be through o) that satisfies the conditions of the theorem and is parallel to L. We assign the values to some parameters:

$$r = c_4 v, \text{where } c_4 \text{ is a constant to be fixed later} \tag{2}$$

$$\delta_2 = \frac{\delta_1 \cdot a}{c_1} \tag{3}$$

$$\epsilon = \frac{\delta_2}{3 c_1 n_P^{\frac{1}{d}+s_1+s_2}} \tag{4}$$

$$\epsilon_0 = \frac{\delta}{6} \tag{5}$$

$$\epsilon_1 = 5\epsilon_0 \tag{6}$$

$$m_1 = \frac{3(\ln 100 + r + \log n_Q)}{\epsilon_0^2} \tag{7}$$

$$m_2 = \frac{(\ln 100 + 2\log n_P + r)}{\epsilon^2} \tag{8}$$

Phase 1 of the algorithm: The input of our algorithm is $P, Q, n_Q = |Q|$, and $n_P = |P|$. Each input point $p \in P$ has the format $< (x_1, \cdots, x_d), w(p) >$, where $p = (x_1, \cdots, x_d)$ and $w(p)$ is the weight of p. The algorithm starts with the following steps: Select a fixed point $o^* \in \Re^d$ and a random plane L through o^* (random hyper-plane can be selected via selecting a random normal vector). Select m_1 random points $q_1, \cdots, q_{m_1}$ from Q and let $Q' =< q_1, \cdots, q_{m_1} >$ represent the list of these points (one point may appear multiple times). For each $q_j \in Q'$, compute its signed distance $d_{q_i} = sd(q_i, L)$ to L. Find the $\lfloor (\frac{1}{d+1} - \epsilon_1) m_1 \rfloor$-th least point $D^*_{1,d+1} = sd(q_1^*, L)$ for $d_{q_1}, \cdots, d_{q_{m_1}}$. Find the $\lceil (\frac{d}{d+1} + \epsilon_1) m_1 \rceil$-th least point $D^*_{d,d+1} = sd(q_2^*, L)$ for $d_{q_1}, \cdots, d_{q_{m_1}}$. Select m_2 random points $p_1, \cdots, p_{m_2}$ from P and let $P' =< p_1, \cdots, p_{m_2} >$ represent the list of these points. For each $p_i \in P'$, compute $d_{p_i} = sd(p_i, L)$. It is well-known that finding the i-th element from a list takes linear steps. The computation above takes $(m_1 + m_2)$ steps. In the rest of the algorithm, we locate the position of the separator hyper-plane by finding its signed distance to L. Its position will be at the center of an interval of size $2a$. In the rest of the proof, we treat both P and Q as lists of points from $\Re^d$. Each point appears only at most once on both P and Q. Let $t_d = k_d \cdot b^{\frac{-1}{d}} + \delta$. For $q \in \Re^d$ and $A \subseteq \Re^d$, define $Pr(A, L, \leftarrow q) = \frac{|\{q' | q' \in A \text{ and } sd(q', L) \leq sd(q, L)\}|}{|A|}$. For a list of points $B =< x_1, \cdots, x_m >$ from $\Re^d$ and a point $q \in \Re^d$, define $X_{B,L,q}(i) = 1$ if $sd(q_i, L) \leq sd(q, L)$, or 0 otherwise. We also define $Y(B, L, q) = \sum_{i=1}^{m} X_{B,L,q}(i)$.

Lemma 5. *It has probability* $\geq 1 - \frac{e^{-r}}{50}$ *such that* $Pr(Q, L, \leftarrow q_1^*) \in [\frac{1}{d+1} - \delta, \frac{1}{d+1} - \frac{\delta}{6}]$ *and* $Pr(Q, L, \leftarrow q_2^*) \in [\frac{d}{d+1} + \frac{\delta}{6}, \frac{d}{d+1} + \delta]$.

Phase 2 of the algorithm: In this phase, we will find a position of L' (parallel to L) with the signed distance to L in the range $[D_{1,d+1}^*, D_{d,d+1}^*]$. Lemma 5 guarantees (with high probability) that each position in the interval $[D_{1,d+1}^*, D_{d,d+1}^*]$ gives a balance partition. We look for the position that has the small sum of weights for the points of P close to L'.

For a list $A =< x_1, \cdots, x_m >$, $|A| = m$ is denoted to be the *length of* A and $x \in A$ means that x is one of the elements in A ($x = x_i$ for some $1 \leq i \leq m$). For a real subset $J \subseteq R$ and a list A of finite points in $\Re^d$, define

$$Pr_*(A, L, J, w_j) = \frac{|\{|p|p \in A \text{ and } w(p) = w_j \text{ and } sd(p, L) \in J\}|}{|A|},$$

and $Z(A, L, J, w_j) = \sum_{p \in A} X_{L,p,J,w_j}^*$, where $X_{L,p,J,w_j}^* = 1$ if $sd(p, L) \in J$ and $w(p) = w_j$, or 0 otherwise. We also define $W(A, L, J) = \sum_{p \in A \text{ and } sd(p,L) \in J} w(p)$. By the definitions, It is easy to see that

$$W(A, L, J) = \sum_{j=1}^{k} w_j Z(A, L, J, w_j) = \sum_{j=1}^{k} w_j Pr_*(A, L, J, w_j)|A|. \tag{9}$$

Since $\sum_{j=1}^{k} n_j = n_P$, we have that $n_j \geq \frac{n_P}{k}$ for some $1 \leq j \leq k$. By (1), we have that $n_P^{s_2} \geq \frac{k}{c_1 w_k} \geq \frac{k^{\frac{d-1}{d}}}{c_1 w_j}$. This implies that $n_P^{\frac{d-1}{d} - s_2} \leq c_1 w_j (\frac{n_P}{k})^{\frac{d-1}{d}} \leq c_1 w_j n_j^{\frac{d-1}{d}}$ for some $1 \leq j \leq k$. By (3), for some $1 \leq j \leq k$,

$$\delta_2 \cdot n_P^{\frac{d-1}{d} - s_2} \leq \delta_1 \cdot a \cdot w_j n_j^{\frac{d-1}{d}}. \tag{10}$$

Lemma 6. *Let* $f \leq n_P$ *be an integer and* $H_1, H_2, \cdots, H_f \subseteq R$ *be* f *real intervals. It has probability* $\geq 1 - \frac{1}{100} e^{-r}$ *such that* $W(P, L, H_i) \in [W(P', L, H_i) \frac{n_P}{m_2} - \delta_2 n_P^{\frac{d-1}{d} - s_2}, W(P', L, H_i) \frac{n_P}{m_2} + \delta_2 n_P^{\frac{d-1}{d} - s_2}])$ *for* $i \leq f$ *and* $j \leq k$.

Case 1. $|D_{1,d+1}^* - D_{d,d+1}^*| \geq 3a n_P^{\frac{2}{d}}$. Partition $[D_{1,d+1}^*, D_{d,d+1}^*]$ into disjoint intervals $[l_1, l_2)$, $[l_2, l_3)$, $\cdots$, $[l_{u-1}, l_u)$, $[l_u, l_{u+1}]$ such that each $l_{i+1} - l_i (i = 1, \cdots, u)$ is equal to $\frac{|D_{1,d+1}^* - D_{d,d+1}^*|}{g_1(n_P)} \geq 3a$, where $g_1(n_P) = u = n_P^{\frac{2}{d}}$. Let $J_i = [l_i, l_{i+1})$ if $i < u$, and $J_u = [l_u, l_{u+1}]$. Compute $W(P', L, J_i)$ for $i = 1, \cdots, u$, which takes $O(m_2 + g_1(n_P)) = O(m_2)$ steps. The algorithm selects $J = J_{i_0}$ that has the least $W(P', L, J_{i_0})$ and let $L' = L(center(J_{i_0}))$, which takes $O(g_1(n_P)) = O(m_2)$ steps. Assume that J_{i_1} is the interval with the least $W(P, L, J_{i_1})$.

Lemma 7. *It has probability* $\geq 1 - \frac{1}{50} e^{-r}$ *such that* $W(P, L, J_{i_0}) \leq \left(k_d \cdot b^{\frac{-1}{d}} + \delta\right) \cdot a \cdot \sum_{j=1}^{k} w_j \cdot n_j^{\frac{d-1}{d}}$.

Case 2. $|D^*_{1,d+1} - D^*_{d,d+1}| < 3an_P^{\frac{2}{d}}$. Let J^* be interval such that $center(J^*) \in [D^*_{1,d+1}, D^*_{d,d+1}]$ and $|J^*| = 2a_1 = 2a(1+\delta_1)$ and $W(P, L, J^*)$ is the least.

Subcase 2.1. $|D^*_{1,d+1} - D^*_{d,d+1}| \leq \delta_1 a$. Let $J = [D^*_{1,d+1} - a, D^*_{1,d+1} + a]$ and let $L' = L(D^*_{1,d+1})$ (In other words, $L' = L(center(J))$). Clearly, $J \subseteq J^*$ and $W(P, L, J) \leq W(P, L, J^*)$.

Subcase 2.2. $\delta_1 a < |D^*_{1,d+1} - D^*_{d,d+1}| < 3an_P$. Let $g_2(n_P)$ be the least integer $v \geq 2$ such that $\frac{|D^*_{d,d+1} - D^*_{1,d+1}| + 2a}{v} \leq \frac{\delta_1 a}{3}$. Since $v \geq 2$ and $\frac{|D^*_{d,d+1} - D^*_{1,d+1}| + 2a}{v-1} > \frac{\delta_1 a}{3}$, we have $\frac{|D^*_{d,d+1} - D^*_{1,d+1}| + 2a}{v} = \frac{v-1}{v} \frac{|D^*_{d,d+1} - D^*_{1,d+1}| + 2a}{v-1} > \frac{v-1}{v} \frac{\delta_1 a}{3} \geq \frac{\delta_1 a}{6}$. Therefore, $v \leq \frac{|D^*_{d,d+1} - D^*_{1,d+1}| + 2a}{\frac{\delta_1 a}{6}} \leq \frac{3an_P^{\frac{2}{d}} + 2a}{\frac{\delta_1 a}{6}} = \frac{6(3n_P^{\frac{2}{d}} + 2)}{\delta_1} = O(n_P^{\frac{2}{d}})$. Let $s = \frac{|D^*_{d,d+1} - D^*_{1,d+1}| + 2a}{g_2(n_P)} \in [\frac{\delta_1 a}{6}, \frac{\delta_1 a}{3}]$. Partition $[D^*_{1,d+1} - a, D^*_{d,d+1} + a]$ into the union of $g_2(n_P)$ disjoint intervals of size s: $[r_1, r_2) \cup [r_2, r_3) \cup \cdots \cup [r_{v-1}, r_v) \cup [r_v, r_{v+1}]$, where $v = g_2(n_P)$ and $r_{i+1} = r_i + s$ for $i = 1, \cdots, v$. Let $I_i = [r_i, r_{i+1})$ for $i = 1, \cdots, v-1$ and $I_v = [r_v, r_{v+1}]$. Let $J^*_i = I_i \cup I_{i+1} \cdots \cup I_{i+h-1}$ for $i = 1, \ldots, v-h+1$, where h is an integer with $2a < h \cdot s < 2a + 2s$. The algorithm selects the interval $J = J^*_{i_2}$ that has the least $W(P', L, J^*_{i_2})$. Finally, the algorithm outputs $L' = L(center(J))$ for the separator hyper-plane. We analyze the algorithm for the case 2.

Lemma 8. *Assume that J is the interval output from the case 2 (either subcase 2.1 or subcase 2.2). It has probability $\geq 1 - \frac{1}{100}e^{-r}$ such that $W(P, L, J) \leq W(P, L, J^*) + 2\delta_1 \cdot aw_j n_j^{\frac{d-1}{d}}$ for some $j \leq k$.*

For a list A of finite points in $\Re^d$ and a hyper-plane M_1, define $F_1(M_1, a, A) = \sum_{p_i \in A \text{ and } \mathrm{dist}(p_i, M_1) \leq a} w(p_i)$. If M_1 and M_2 are two hyper-planes with signed distance $d_{M_1,M_2} = sd(p, M_1)$ for some point p in the M_2, then $F_1(M_2, a, A) = W(A, M_1, J)$, where J is the interval $[d_{M_1,M_2}, -a, d_{M_1,M_2} + a]$. The the hyper-plane $L(center(J^*_{i_2}))$ output by the algorithm has that $F_1(L(center(J^*_{i_2})), a, P') \leq F_1(L(center(J^*)), a_1, P') + 2\delta_1 \cdot aw_j n_j^{\frac{d-1}{d}}$ for some $j \leq k$. See the section **??** for the algorithm description in the Appendix.

Time and accuracy of the algorithm: After the hyper-plane L is selected in phase one, by Lemma 5 we have the probability at least $1 - e^{-r}$ that both $Pr(Q, L, \leftarrow q^*_1) \in [\frac{1}{d+1} - \delta, \frac{1}{d+1} - \frac{\delta}{6}]$ and $Pr(Q, L, \leftarrow q^*_2) \in [\frac{d}{d+1} + \frac{\delta}{6}, \frac{d}{d+1} + \delta]$. This means every L' (parallel to L) with the signed distance in the interval $[D^*_{1,d+1}, D^*_{d,d+1}]$, it has at most $(\frac{d}{d+1} + \delta)n_Q$ points of Q in each of the half spaces. In phase 2, we have probability at least $1 - e^{-r}$ to output the separator L' such that $F_1(L', a, P) \leq \left(k_d \cdot b^{\frac{-1}{d}} + \delta\right) \cdot a \cdot \sum_{j=1}^k w_j \cdot n_j^{\frac{d-1}{d}}$ (case 1 of phase 1, see Lemma 7) or $F_1(L', a, P)) \leq F_1(L(J^*), a_1, P) + 2\delta_2 w_j n_j^{\frac{d-1}{d}}$ (case 2 of phase 2, see Lemma 8), where J^* is the interval of length $2a_1$ with the least $F_1(L(J^*), a_1, P)$ and center between $D^*_{1,d+1}$ and $D^*_{d,d+1}$.

Assume that L is a fixed hyper-plane and L^* is a another hyper-plane that is parallel to L and $F_1(L^*, a_1, P)$ is the least. By Lemma 7 and Lemma 8, it has probability $\geq (1 - e^{-r})^2$ such that we can get another L' (parallel to L) such that $F_1(L', a, P) \le F_1(L^*, a_1, P) + 2\delta_1 w_j n_j^{\frac{d-1}{d}}$ for some $j \le k$ or $F_1(L', a, P) \le \left(k_d \cdot b^{\frac{-1}{d}} + \delta\right) \cdot a \cdot \sum_{j=1}^k w_j \cdot n_j^{\frac{d-1}{d}}$. The number of points in Q in each side of L' is $\le (\frac{d}{d+1} + \delta) n_Q$.

We have probability at most $\frac{1}{1+\alpha}$ that $F_{a_1,P,o}(L) \ge (1+\alpha)E(F_{a_1,P,o})$. If the algorithm repeats z times, let $L_1, \cdots, L_z$ be the random hyper planes selected for L. With probability $\ge (1 - (\frac{1}{1+\alpha})^z)$, one of those L_is has another hyper-plane L_i^* such that L_i^* is parallel to L_i and has $F_{a_1,P,o}(L_i^*) \le (1+\alpha)E(F_{a_1,P,o})$. Therefore, we have probability at least $(1 - (\frac{1}{\alpha+1})^z)(1 - e^{-r})^{2z}$ to find out such a L' with $F_1(L', a, P) \le (1+\alpha)E(F_{a_1,P,o}) + 2\delta_1 w_j n_j^{\frac{d-1}{d}}$ for some $j \le k$ or $F_1(L', a, P) \le \left(k_d \cdot b^{\frac{-1}{d}} + \delta\right) \cdot a \cdot \sum_{j=1}^k w_j \cdot n_j^{\frac{d-1}{d}}$. Thus, $F_1(L', a, P) \le \left(k_d \cdot b^{\frac{-1}{d}} + \delta\right) \cdot a \cdot \sum_{j=1}^k w_j n_j^{\frac{d-1}{d}} + O(n_P^{\frac{d-2}{d}+s_1})$.

Now we give a bound for the probability. Let $z = \frac{2r}{\ln(1+\alpha)}$. Then $1 - (\frac{1}{1+\alpha})^z > 1 - e^{-r}$.

$$(1 - (\frac{1}{1+\alpha})^z)(1 - e^{-r})^{2z} > (1 - e^{-r})^{2z+1} > 1 - (2z+1)e^{-r} > 1 - \frac{1}{2^v},$$

where we let $r = c_4 v$ for some constant c_4 large enough.

The phase 1 of the algorithm takes $O(m_1 + m_2)$ steps. The case 1 of phase 2 takes $O(m_2)$ steps. The case 2 of phase 2 takes $O(m_2)$ steps. Totally, it takes $O(z(m_1+m_2)) = O(v \cdot (n_P^{\frac{2}{d}+2(s_1+s_2)} \cdot (\log n_P + v) + v \log n_Q)) = O(v^2 \cdot (n_P^{\frac{2}{d}+2(s_1+s_2)} \cdot \log n_P + \log n_Q))$ steps. □

Corollary 1. *Let $d \ge 2$ be the dimension number and the parameter $v > 0$. Let $a > 0$ be a constant and $\delta > 0$ be a small constant. There exists a randomized $O(v^2 n^{\frac{2}{d}} \log n)$ time such that given a set Q of n grid points in $\Re^d$, the algorithm finds a hyper-plane L with probability at least $1 - \frac{1}{2^v}$ such that each side of L has at most $(\frac{d}{d+1} + \delta)n$ points of Q, and the number of points of Q with distance $\le a$ to L is $\le (k_d + \delta)an^{\frac{d-1}{d}}$.*

6 An Application to Protein Side-Chain Packing Problem

We follow the description of Xu [26] for the model of protein side chain packing. The side-chain prediction problem can be formulated as follows. We use a reside interaction graph $G = (V, E)$ to represent a protein resides and their interactions. Each vertex in V represents a residue of the protein. For each reside $i \in V$, $D(i)$ is the set of all possible rotamers of side chain i. There is an interaction edge $(i, j) \in E$ if and only if there are $l \in D(i)$ and $k \in D(j)$ such that there exist an atom in the rotamer l conflicts with another atom in the rotamer k. Two atoms conflict each other iff their distance is less than the sum of their radii. For each two rotamers $l \in D(i)$ and $k \in D(j)$ $(i \ne j)$, there is an associated score $P_{i,j}(l,k)$ if residue i interacts with residue j. For each rotamer $l \in D(i)$, there is a score $S_i(l)$, which characterizes the interaction energy between l and the backbone of the protein. The prediction problem is to give $A(i) \in D(i)$ to residues $i \in V$ so that the following energy value is minimized. $E(G) = \sum_{i \in V} S_i(A(i)) + \sum_{i \neq j,(I,j) \in E} P_{i,j}(A(i), A(j))$.

For more detailed description about the protein side chain packing, see (e.g. [23, 4, 26, 5]). Let d_u^* be distance such that there is no interaction between two resides if their distance is $\geq d_u^*$. Let d_l^* be the minimal distance between two amino acids. Both d_u^* and d_l^* are constants.

Theorem 3. *There exists a $r_{\max}^{O(n^{\frac{2}{3}})}$-time algorithm to find the optimal solution for the protein side chain packing problem, where $r_{\max}$ is the maximal number of rotamers of one amino acid. In other words, $r_{\max} = \max_i |D(i)|$.*

Proof. Our algorithm is based on the divide and conquer method. Let $d_0 = d_l^* \frac{\sqrt{2}}{2}$ be the unit distance. Since $d_l^* = \sqrt{2} d_0$, we consider that the minimal distance between two amino acids is $d_l = \sqrt{2}$ and the minimal distance for the interaction between two side chains is $d_u = \frac{d_u^*}{d_0}$. For a grid point $p = (x, y, z)$ (x, y, z are integers), define $cube(p) = \{(u, v, w) \in \Re^3 | x - \frac{1}{2} \leq u < x + \frac{1}{2}$ and $y - \frac{1}{2} \leq v < y + \frac{1}{2}$ and $z - \frac{1}{2} \leq w < z + \frac{1}{2}\}$. The 3D space $\Re^3$ is partitioned into many cubes: $\Re^3 = cube(p_0) \cup cube(p_1) \cup \cdots$. For different grid points $p \neq p'$, $cube(p) \cap cube(p') = \emptyset$. Each amino acid is represented by the position of its C_α. Therefore, no two amino acids can stay at the same $cube(p)$ for any grid point p. Let P be the set of all grid points p such that $cube(p)$ contains the C_α for an amino acid.

Let $w = d_u + 2\sqrt{2}$. By Corollary 1, there exists a w-wide separator L plane such that each side has at most $(\frac{3}{4} + \delta)n$ contain amino acid, and the number of grid points (with amino acids in its cube) is bounded by $1.209wn^{\frac{2}{3}}$, where $\delta > 0$ is an arbitrary small constant. The w-wide separator partitions the problem into P_1, S and P_2, where S is the separator area. Clearly, a side chain whose amino acid C_α is in $cube(p)$ with $p \in P_1$ does not interact another side chain in P_2 because of the w-wide separator between P_1 and P_2.

The number of ways to arrange the side chains in the separator area S is bounded by $r_{\max}^{1.209wn^{\frac{2}{3}}}$. We only need $O(n)$ time for computing the separator. We assume that $r_{\max} \geq 2$ (otherwise, it is trivial). Let $T(n)$ is the computational time for the protein side chain packing problem with n resides. Solving each sub-problem $P_i (i = 1, 2)$ takes $T((\frac{3}{4} + \delta)n)$ steps. We have the recursive $T(n) \leq 2(r_{\max}^{1.209wn^{\frac{2}{3}}} + O(n))T((\frac{3}{4} + \delta)n)$. This gives that $T(n) = r_{\max}^{O(n^{\frac{2}{3}})}$. □

References

1. T. Akutsu, NP-hardness results for protein side-chain packing, In S. Miyano and T. Takagi, editors, Genome Informatics 8, 1997, pp. 180-186.
2. N. Alon, P. Seymour, and R. Thomas, Planar Separator, SIAM J. Discr. Math. 7,2(1990) 184-193.

3. N. Alon, P. Seymour, and R. Thomas, A separator theorem for graphs with an excluded minor and its applications, STOC'90, pp. 293-299.
4. A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack Jr., A graph-theory algorithm for rapid protein side-chain prediction, b Protein science, 12: 2003, pp. 2001-2014.
5. B. Chazelle, C. Kingsford, and M. Singh, A semidefinite programming approach to side-chain positioning with new rounding strategies, INFORMS Journal on Computing, 2004, pp. 86-94.
6. Z. Chen, B. Fu, Y. Tang, and B. Zhu, A PTAS for a DISC covering problem using width-bounded separator, COCOON'05, 2005.
7. H. N. Djidjev, On the problem of partitioning planar graphs. SIAM journal on discrete mathematics, 3(2) June, 1982, pp. 229-240.
8. H. N. Djidjev and S. M. Venkatesan, Reduced constants for simple cycle graph separation, Acta informatica, 34(1997), pp. 231-234.
9. D. Eppstein, G. L. Miller, S. Teng, A Deterministic Linear Time Algorithm for Geometric Separators and its Applications, SOCG'93, pp. 99-108.
10. 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, ICALP'04, pp. 630-644.
11. B. Fu, Theory and application of width bounded geometric separator, Electronic Colloquium on Computational Complexity 2005, TR05-13. Accepted by STACS'06.
12. H. Gazit, An improved algorithm for separating a planar graph, manuscript, USC, 1986.
13. J. R. Gilbert, J. P. Hutchinson, and R. E. Tarjan, A separation theorem for graphs of bounded genus, Journal of algorithm, (5)1984, pp. 391-407.
14. S. Jadhar, Computing a center of a finite planar set of points in linear time, SOCG'93, pp. 83-90.
15. D. Lichtenstein, Planar formula and their uses, SIAM journal on computing, 11,2(1982), pp. 329-343.
16. R. J. Lipton and R. Tarjan, A separator theorem for planar graph, SIAM Journal on Applied Mathematics, 36(1979) 177-189.
17. G. L. Miller, S.-H. Teng, and S. A. Vavasis, An unified geometric approach to graph separators, FOCS'91, pp. 538-547.
18. G. L. Miller and W. Thurston, Separators in two and three dimensions, STOC'90, pp. 300-309.
19. G. L. Miller and S. A. Vavasis, Density graphs and separators, SODA'91, pp. 331-336.
20. R. Motwani and P. Raghavan, Randomized algorithms, Cambridge University Press, 1995.
21. J. Pach and P. K. Agarwal, Combinatorial geometry, Wiley-Interscience Publication, 1995.
22. S. Plotkin, S. Rao, and W. D. Smith, Shallow excluded minors and improved graph decomposition, SODA'90, pp. 462-470.
23. J. W. Ponter, and F. M. Richards, Tertiary templates for proteins: use of packing criteria and the enumeration of allowed sequences for different structural classes. J. molecular biology, 193, 1987, pp. 775-791.
24. W. D. Smith and N. C. Wormald, Application of geometric separator theorems, FOCS'98, pp. 232-243.
25. D. A. Spielman and S. H. Teng, Disk packings and planar separators, SOCG'96, pp.349-358.
26. J. Xu, Rapid protein side-chain packing via tree decomposition, RECOMB,05, pp. 408-422.