

A $2^{O(n^{1-\frac{1}{d}} \log n)}$ time algorithm for d-dimensional protein folding in the HP-model

Bin Fu ^{*} Wei Wang [†]

Abstract

The protein folding problem in the HP-model is NP-hard in both 2D and 3D [4, 6]. The problem is to put a sequence, consisting of two characters H and P, on a d-dimensional grid to have the maximal number of HH contacts. We design a $2^{O(n^{1-\frac{1}{d}} \log n)}$ time algorithm for d-dimensional protein folding in the HP-model. In particular, our algorithm has $O(2^{6.145\sqrt{n} \log n})$ and $O(2^{4.306n^{\frac{2}{3}} \log n})$ computational time in 2D and 3D respectively. The algorithm is derived via our separator theorem for points on a d -dimensional grid. For example, for a set of n points P on a 2-dimensional grid, there is a separator with at most $1.129\sqrt{n}$ points that partitions P into two sides with at most $(\frac{2}{3})n$ points on each side. Our separator theorem for grid points has a greatly reduced upper bound than that for the general planar graph [2].

1. Introduction

Proteins are composed of 20 amino acids. Two amino acids can be connected via a peptide bond. A protein sequence can be generated by using peptide bonds to connect amino acids. A protein can fold into a specific 3D structure, which is uniquely determined by the sequence of amino acids. Its 3D structure determines its function. A standard procedure to determine 3D structure is to produce a pure solution with only the protein, then crystallize it followed by x-ray crystallography. This is a very time consuming process. Therefore, protein structure prediction with computational technology is one of the most significant problems in bioinformatics.

It is much easier to identify a protein's 1D sequence than its 3D structure. In order to carry out their various functions, proteins must fold into a 3D structure. By studying how proteins fold, their functions can be better understood. The study of protein folding can help answer questions such as how a protein changes to a totally different function or how the function of a protein changes with its structure

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 [14, 15]. In this 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 non-consecutive H-monomers. The folding problem in the HP-model is to find the conformation for any HP-sequence with minimal energy. This problem was proven to be NP-hard in both 2D and 3D [4, 6].

Some algorithms for this problem have been developed based on the heuristic, genetic, Monte Carlo, branch and bound methods (e.g. [26, 27, 28, 25, 19, 22, 12, 13, 21, 17, 23, 7, 3]). Although

^{*}Address: 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. Email: fu@cs.uno.edu

[†]Address: Department of Chemistry and Biochemistry, University of California at San Diego, CA 92093. Email: wwang@chem.ucsd.edu

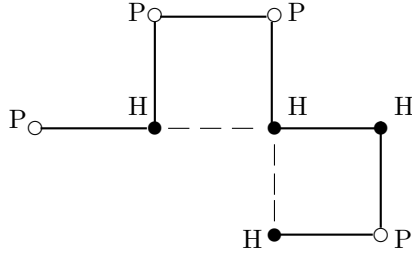


Figure 1: The sequence PHPPHHPH is put on the 2 dimensional grid. There are 2 H-H contacts marked by the dotted lines.

many experimental results were reported for testing sequences of small length, we have not seen any theoretical analysis about the computational time upper bound of the algorithms. Another approach is to develop polynomial time approximation algorithms for the protein folding in the HP model [10, 1, 18]. Hart and Istrail [10] showed a polynomial time $\frac{3}{8}$ -approximation algorithm for the 3D folding in the HP model and Newman [18] derived a polynomial time $\frac{1}{3}$ -approximation algorithm for the 2D problem, improving $\frac{1}{4}$ -approximation algorithm in [10].

If the first letter of a HP sequence is fixed at a position of 2D (3D) plane (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 (space resp.). Our algorithm computational time is bounded by $2^{O(n^{\frac{1}{2}} \log n)}$ ($2^{O(n^{\frac{2}{3}} \log n)}$) in 2D (3D resp.). As the average number of amino acids of proteins is between 400 to 600, if an algorithm could solve the the protein structure prediction with ≤ 1000 amino acids, it would be able to satisfy most of the application demand. Our effort is a theoretical step toward this target.

Our algorithm is the divide and conquer approach, which is based on our geometric separator for the points on a d -dimensional grid. Lipton and Tarjan [16] showed the well known geometric separator for planar graphs. Their result has been elaborated by many subsequent authors. The best known separator theorem for planar graphs was proved by Alon, Seymour and Thomas [2]

Theorem 1. [2] *Any planar graph of n vertices has a vertex subset of cardinality $\leq \sqrt{4.5n}$ whose removal separates the graph into two components each having $\frac{2n}{3}$ vertices.*

Some other forms of the separator theorem were applied in deriving algorithms for some geometric problems such as the planar Travelling Salesman and Steiner Tree problems (e.g. see [24]). Those problems usually have input points with fixed geometric positions in space. A set of grid points on the plane forms a planar graph by adding edges to every two grid points with distance 1. As the input of folding problem is only a sequence of letters, their locations in space are unknown and will be determined by the algorithm. We do not know if the separator theorem like Theorem 1 can be applied to the folding problem. We derive a separator theorem for the grid points with a greatly reduced upper bound for the number of points on the separator than that for the planar graph.

Theorem 2. *For a set P of n grid points on a 2-dimensional 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 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 is possible to use the separator theorem in the algorithm for the folding problem even though the locations of the letters are not known.

2. An easy separator and algorithm

We will show that there is a small set of letters with size $O(n^{1-\frac{1}{d}})$ on a hyper plane (denoted by $P_{r,a}$ for some $1 \leq r \leq d$ and integer a in the definition below) to partition the folding problem of n letters into 2 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. The separator in this section has a self-contained proof and implies an $n^{O(n^{1-\frac{1}{d}})}$ -time algorithm for the folding problem in the HP-model.

2.1. A balanced separator

Let the dimensional number d be fixed. We need the following terms:

Definition 3.

- For a set A , $|A|$ is denoted as the number of elements in A .
- The integer set is represented by $Z = \{\dots, -2, -1, 0, 1, 2, \dots\}$. For integers i and j , *integer interval* $[i, j] = \{i, i+1, \dots, j\}$. For integers x_1, \dots, x_d , (x_1, \dots, x_d) is a d -dimensional *grid point*.
- For two points p_1, p_2 with the same dimension, $\text{dist}(p_1, p_2)$ is the Euclidean distance between them.
- For a set Σ of letters, a Σ -*sequence* is a sequence of letters from Σ . For example, $P H P P H H P H$ is an $\{H, P\}$ -sequence. For a sequence S of length n and $1 \leq i \leq n$, $S[i]$ is the i -th letter of S . $S[i, j]$ denotes the subsequence $S[i]S[i+1] \cdots S[j]$. If $[i_1, j_1], [i_2, j_2], \dots, [i_t, j_t]$ are disjoint intervals inside $[1, n]$, we call $S[i_1, j_1], S[i_2, j_2], \dots, S[i_t, j_t]$ *disjoint subsequences* of S . For a set of integers $A = \{i_1 < i_2 < \dots < i_k\}$, define $S[A] = S[i_1]S[i_2] \cdots S[i_k]$.
- For a d -dimensional point (x_1, \dots, x_d) , define $\|(x_1, \dots, x_d)\| = \sum_{i=1}^d |x_i|$.
- A *self-avoiding arrangement* f for a sequence S of length n on the d -dimensional grid is a one-to-one mapping from $\{1, 2, \dots, n\}$ to Z^d such that $\|f(i) - f(i+1)\| = 1$ for $i = 1, 2, \dots, n-1$. For the disjoint subsequences $S[i_1, j_1], \dots, S[i_k, j_k]$ of S , a *partial self-avoiding arrangement* of S on $S[i_1, j_1], \dots, S[i_k, j_k]$ is a partial function f from $\{1, 2, \dots, n\}$ 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 $G_f = (1, 2, \dots, n, E)$, where the edge set $E = \{(i, j) : |i - j| > 1 \text{ and } \|f(i) - f(j)\| = 1\}$.
- A r -plane is the set $P_{r,a} = \{(x_1, \dots, x_{r-1}, a, x_{r+1}, \dots, x_d) | x_1, \dots, x_{r-1}, x_{r+1}, \dots, x_d \in Z\}$, which has all of the elements in Z^d with the r -th element of fixed value a .
- $P_{r,<a} = \{(x_1, \dots, x_{r-1}, x_r, x_{r+1}, \dots, x_d) | x_1, \dots, x_{r-1}, x_r, x_{r+1}, \dots, x_d \in Z \text{ and } x_r < a\}$.
- $P_{r,>a} = \{(x_1, \dots, x_{r-1}, x_r, x_{r+1}, \dots, x_d) | x_1, \dots, x_{r-1}, x_r, x_{r+1}, \dots, x_d \in Z \text{ and } x_r > a\}$.
- $P_{r,\leq a} = P_{r,<a} \cup P_{r,a}$, and $P_{r,\geq a} = P_{r,>a} \cup P_{r,a}$.
- For a set of points S in d -dimensional space and $1 \leq r \leq d$ and $a \in Z$, define $S(r, < a) = \{(x_1, \dots, x_d) \in S | x_r < a\}$, $S(r, = a) = \{(x_1, \dots, x_d) \in S | x_r = a\}$, and $S(r, > a) = \{(x_1, \dots, x_d) \in S | x_r > a\}$.
- For $0 < c < 1$ and a set S in d -dimensional space, a $P_{r,a}$ is a *c -balanced-separator* if $|S(r, < a)| \leq c \cdot |S|$ and $|S(r, > a)| \leq c \cdot |S|$.

- A *rectangular region* R in d -dimensional space is the intersection of a finite number of sets P_1, P_2, \dots, 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 $(i = 1, \dots, k)$. The *boundary* of a rectangular region R consists of all of those points (x_1, \dots, x_d) such that (x_1, \dots, x_d) is not in R , and for some $1 \leq r \leq d$ and $a \in \{-1, 1\}$, $(x_1, \dots, x_{r-1}, x_r + a, x_{r+1}, \dots, x_d)$ is in R .
- A rectangular region R in d -dimensional space is of *size* $m_1 \times m_2 \times \dots \times m_d$ if $m_i = \max\{x_i - x'_i | (x_1, \dots, x_d), (x'_1, \dots, x'_d) \in R\} + 1$ for $i = 1, \dots, d$.

Lemma 4. For a set S of n grid points in 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 $S = S_0 \supseteq S_1 \supseteq S_2 \supseteq \dots \supseteq S_t$ such that $t \leq d - 1$ and $|S_i| \geq \frac{1}{2}|S_{i-1}|$ for $i = 1, 2, \dots, t$. 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 the $P_{r,a}$ such that $P_{r,a}$ is a $\frac{3}{4}$ -balanced-separator for S_{r-1} . It is easy to see that 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

$$|S(r, < a)| \leq |S_{r-1}(r, < a)| + |S - S_{r-1}| \quad (1)$$

$$\leq \frac{3}{4}|S_{r-1}| + |S| - |S_{r-1}| \quad (2)$$

$$= |S| - \frac{1}{4}|S_{r-1}| \quad (3)$$

$$\leq (1 - \frac{1}{2^{r+1}})|S| \quad (4)$$

$$\leq (1 - \frac{1}{2^d})|S| \quad (5)$$

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 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 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 . Therefore, $|Q_r| \leq n^{\frac{1}{d}}$ for $1 \leq r \leq d - 1$. For every $P_{d,a}$, as $|Q_r| \leq n^{\frac{1}{d}}$ for $1 \leq r \leq d - 1$, we have

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

As $|S_{d-1}| \geq \frac{|S|}{2^{d-1}} = \frac{1}{2^{d-1}}n$, there are at least $\frac{\frac{1}{2}|S_{d-1}|}{n^{1-\frac{1}{d}}} \geq \frac{1}{2^d} \cdot n^{\frac{1}{d}}$ $P_{d,a}$ s to be $\frac{3}{4}$ -balanced-separator 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

$$|S(d, < a)| \leq |S_{d-1}(r, < a)| + |S - S_{d-1}| \quad (6)$$

$$\leq \frac{3}{4}|S_{d-1}| + |S| - |S_{d-1}| \quad (7)$$

$$= |S| - \frac{1}{4}|S_{d-1}| \quad (8)$$

$$\leq (1 - \frac{1}{2^{d+1}})|S| \quad (9)$$

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 hyper plane $P_{r,a}$, which intersects the cube, shares $n^{\frac{d-1}{d}}$ grid points with the cube. This shows it is impossible to improve the separator to $o(n^{\frac{d-1}{d}})$. The next section shows that we can improve the separator by a constant factor. This lemma indicates that the balanced separator can be found among $O(dn)$ hyper-planes.

2.2. Algorithm

As we are going to describe our algorithm recursively, we use the following term to characterize the problem. A d -dimensional **Multi-Sequence Folding Problem** F is formulated as follows:

The inputs are

- i. disjoint subsequences S_1, S_2, \dots, S_k of sequence S_0 ($S_t = S_0[i_t, j_t]$ for $t = 1, \dots, k$), and
- ii. a rectangular region R , where all of the k $\{H, P\}$ -sequences are going to be arranged, and
- iii. a series of k pairs of points in R : $(p_1, q_1), (p_2, q_2), \dots, (p_k, q_k)$, in which points $p_t \in R$ and $q_t \in R$ are the positions for putting the first and last letters of S_t respectively, and
- iv. a set of available points to put the letters from the k sequences, and
- v. a set of $\{H, P\}$ points on R , which already have letters H and P from $S_0[[1, n] - \cup_{t=1}^k [i_t, j_t]]$.

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

A hyper-plane $P_{r,a}$ partitions a multi-sequence folding problem F into two multi-sequence 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 the $P_{r,a}$ (see Figure 2). Furthermore, the available points of F_1 (F_2) are the intersection of F 's available points with $P_{r, < a}$ ($P_{r, > a}$ resp.).

Algorithm

- (a) Input d -dimensional multi-sequence folding problem F (as the definition).
- (b) For each subset S of $\leq c'(d) \cdot n^{\frac{d-1}{d}}$ letters from S_1, \dots, S_k , every plane $P_{r,a}$ (with nonempty intersection with R) and every arrangement of S in 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) Recursively solve F_1 and F_2 .
 - (g) Merge the solution 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 .

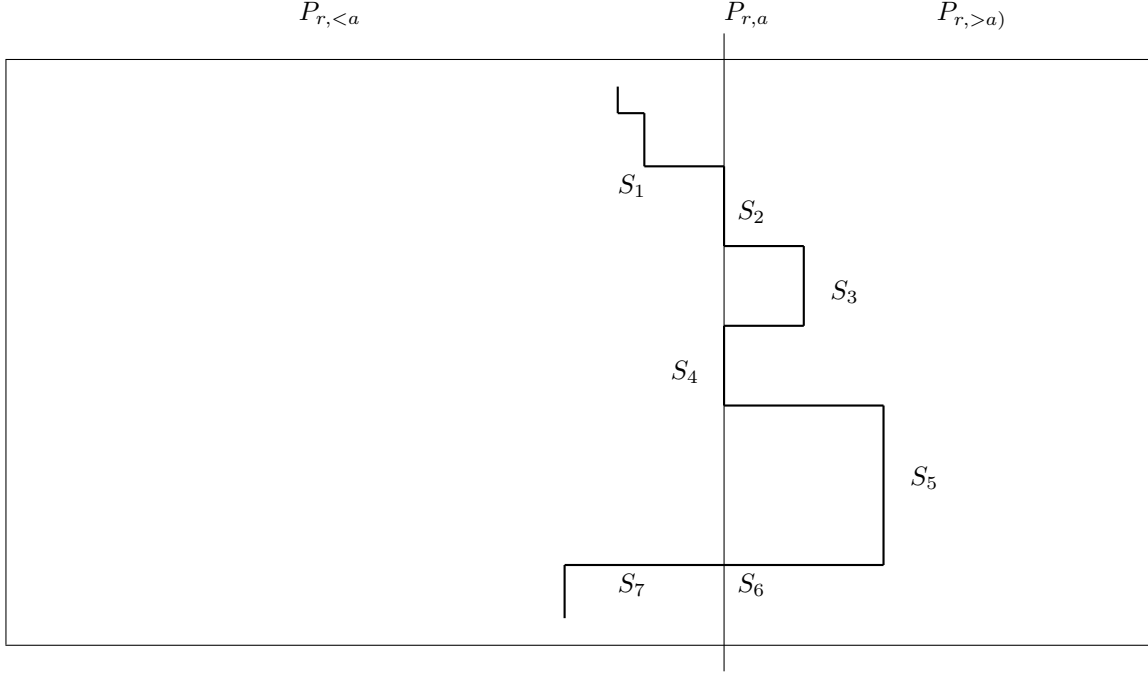


Figure 2: The hyper-plane $P_{r,a}$ partitions a sequence into 3 groups of disjoint subsequences $\{S_1, S_7\}$, $\{S_2, S_4, S_6\}$ and $\{S_3, S_5\}$ in $P_{r,<a}$, $P_{r,a}$ and $P_{r,>a}$ respectively

End of the Algorithm

Lemma 5. *There is a $(nm)^{O(n^{1-\frac{1}{d}})}$ time algorithm for the d -dimensional multi-sequence folding problem with a $m_1 \times m_2 \times \dots \times m_d$ rectangular region in the HP-model, where $m = \max\{\max\{m_i | i = 1, \dots, d\}, 2\}$ and the dimensional number d is assumed to be a constant.*

Proof: By Lemma 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 non-empty intersection with the $m_1 \times m_2 \times \dots \times m_d$ rectangular region. There are at most $d \cdot m$ ways to select the 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 letters, and put those letters on the plane. So, the loop (c)-(i) is repeated $\leq dn^t m^{(d-1)t+1}$ times.

For disjoint subsequences S_1, \dots, S_k of S_0 inside a rectangular region R , if we fix $t \leq c'(d) \cdot n^{1-\frac{1}{d}}$ letters from S_1, \dots, S_k on the hyper plane $P_{r,a}$, they are partitioned into three groups of subsequences of S_0 which are in $R \cap P_{r,<a}$, $P_{r,a}$ and $P_{r,>a}$ 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 end points under all possible cases. The sub-sequences 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 end points 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.

We have the following recursive relationship for the total time of the algorithm:

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

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

Theorem 6. *There is a $2^{O(n^{1-\frac{1}{d}} \log n)}$ time algorithm for the d -dimensional protein folding in the HP-model for fixed d .*

Proof: The folding problem can be put into a $n \times n \cdots n$ rectangular region in d -dimensional space by fixing the two middle letters in two center neighbor points in the region. By Lemma 5, we have an $n^{O(n^{1-\frac{1}{d}})} = 2^{O(n^{1-\frac{1}{d}} \log n)}$ time algorithm. \blacksquare

3. Improved separators and algorithms

The last section shows 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. Our approach is to improve the separator. The following well known fact (see [20]) will be used for deriving our new separator. Our reduced upper bound for the number of points on the separator is from on the fact below: For a set P of 2-dimensional grid points with the centerpoint o (see Lemma 7), a random line through o has the largest expected number of points of P with distance $\leq a$ to it when the points P are tightly arranged in the grid points inside a circle with the least radius. It is also true in dimension larger than 2.

Lemma 7. *For an n -element set P in 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 8. For a grid point (i, j) on 2-dimensional 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 $(i + \frac{1}{2}, j + \frac{1}{2})$. A *grid cube* is a $1 \times 1 \times 1$ cube with eight corner points $\{(i + \alpha, j + \beta, k + \gamma) | \alpha, \beta, \gamma \in \{-\frac{1}{2}, \frac{1}{2}\}\}$ for a 3-dimensional grid point (i, j, k) .

3.1. 2-dimension

Lemma 9. (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 a 2-dimensional 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 $\leq (a + \sqrt{2})(m + \sqrt{2})$.*

(5) *For every line L and fixed $a > 0$, there are at most $(2a + \sqrt{2})\sqrt{2}(n + 1)$ grid points inside a $n \times n$ 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 a 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 boundary is outside C_1 and inside C_2 . The number of grid squares intersecting C 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 $\pi r^2 - 4\sqrt{2}\pi r > n$.

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

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

Definition 10. Define $Pr_2(a, p_0, p)$ to be the probability that the point p has $\leq a$ perpendicular distance to a random line L through the point p_0 .

Lemma 11. *Let $a > 0$ be a constant and $\delta > 0$ be a small constant. Let P be a set of points on 2-dimensional grid. Assume that all 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 $p = (x, y)$ is a point of P and L is a random line passing through the center $o = (x_0, y_0)$. Let C be the circle of radius r and center o such that C covers all points in P . Let C' be the circle of radius $r' = r + \frac{\sqrt{2}}{2}$ and the same center o . It is easy to see 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 points p in P such that $\text{dist}(o, p) < r_0$. By Lemma 9, 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

$$\sum_{p \in P} Pr_2(a, o, p) = \sum_{p \in P_1} Pr_2(a, o, p) + \sum_{p \in P - P_1} Pr_2(a, o, p) \quad (10)$$

$$= \sum_{p \in P_1} 1 + \sum_{p \in P - P_1} \frac{2 \arcsin \frac{a}{\text{dist}(o,p)}}{\pi} \quad (11)$$

$$< \pi(r_0 + \frac{\sqrt{2}}{2})^2 + \sum_{p \in P - P_1} \frac{(1 + \epsilon)2a}{\pi \text{dist}(o, p)} \quad (12)$$

$$\leq \pi(r_0 + \frac{\sqrt{2}}{2})^2 + \frac{2a(1 + \epsilon)^2}{\pi} \int \int_{C'} \frac{1}{\text{dist}(o, p)} d_x d_y \quad (13)$$

$$= \frac{2a(1 + \epsilon)^2}{\pi} \int_0^{2\pi} \int_0^{r'} \frac{\rho}{\rho} d\rho d\theta + \pi(r_0 + \frac{\sqrt{2}}{2})^2 \quad (14)$$

$$= 4a(1 + \epsilon)^2 r' + \pi(r_0 + \frac{\sqrt{2}}{2})^2 \quad (15)$$

$$< 4ar + \delta r \text{ for all large } r \text{ by selecting } \epsilon \text{ small enough.} \quad (16)$$

We use the transformation $x = \rho \cos \theta + x_0, y = \rho \sin \theta + y_0$ to convert the integration at 13 to that at 14 above. ■

Lemma 12. *Let $a > 0$ be a constant and $\epsilon > 0$ be a small constant. For a set P of n grid points in a 2-dimensional 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 is at the point o (see Lemma 7). We are going to estimate the upper bound for the expected number of points in P , which have $\leq a$ distance to a random line L through o .

Let $r = \frac{1}{\sqrt{\pi}} \sqrt{n} + 4\sqrt{2}$. By Lemma 9, the circle C with center o and 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 12, $\sum_{p \in P} Pr_2(a, o, p) \leq \sum_{p \in P} Pr_2(a, o, f(p)) \leq 4ar + \delta r = \frac{4a}{\sqrt{\pi}} \sqrt{n} + \epsilon \sqrt{n}$ by selecting small δ . ■

It is easy to see that Lemma 12 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 a $m \times m$ 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 a ϵ -regular line segment. A ϵ -regular line is a line containing two ϵ -regular points.

Lemma 13. *Let $\epsilon > 0$ be a constant. Every line segment L_1 inside the $m \times m$ square has a 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 $\text{dist}(p_1, p_2) \leq \epsilon$, and for every point $q_2 \in L_2$, there is a point $q_1 \in L_1$ with $\text{dist}(q_1, q_2) \leq \epsilon$.*

Proof: Assume E_1, E_2, E_3 , and E_4 are the 4 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 closet 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 $\leq \max(\text{dist}(p_i, q_i), \text{dist}(p_j, q_j)) \leq \epsilon$, and every point q in L_2 has another point in $q' \in L_1$ with distance $\leq \max(\text{dist}(p_i, q_i), \text{dist}(p_j, q_j)) \leq \epsilon$. \blacksquare

Lemma 14. *Let a and ϵ be positive constants. Let P be a set of n points in a 2-dimensional grid. There is a ϵ -regular line L such that there are $\leq (\frac{2}{3} + \epsilon)n$ points of P on each half plane, and $\leq 4(a + \epsilon)\frac{\sqrt{n}}{\sqrt{\pi}}$ points of P to have distance $\leq a$ to L .*

Proof: Let $\delta > 0$ be a small constant. By Lemma 11, 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 has at most $\frac{2}{3}n$ points in P . By Lemma 13, 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. \blacksquare

Lemma 15. *For some constants $c_0, \epsilon > 0$, there is a $O(m^{c_0 \log n} n_0^{(6.145 - \epsilon)\sqrt{n}})$ time algorithm for the 2D Multi-Sequence Folding Problem in an $m \times m$ square, where n is the sum of lengths of 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. By the Lemma 14, there is a line L such that P has at most $d\sqrt{n}$ points to have distance $\leq 1/2$ to L , and each half plane has at most cn points from P . The letters that stay on 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 in the last section.

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 formula, we have $(d\sqrt{n})! > \frac{(d\sqrt{n})^{d\sqrt{n}}}{3^{d\sqrt{n}}}$. There are $u_2 = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{d\sqrt{n}} < d\sqrt{n} \frac{n^{d\sqrt{n}}}{(d\sqrt{n})!} < (\frac{3}{d})^{d\sqrt{n}} \cdot d\sqrt{n} \cdot n^{\frac{1}{2}d\sqrt{n}}$ ways to select the $\leq d\sqrt{n}$ letters from the n of them.

Assume fixed k ($\leq d\sqrt{n}$) letters $S_0[i_1], S_0[i_2], \dots, S_0[i_k]$ ($1 \leq i_1 < i_2 < \dots < i_k \leq n$) are from the disjoint subsequences of S_0 . By Lemma 9, there are at most $(2a + 1)\sqrt{2}(m + 1)$ positions (inside the $m \times m$ square) to put the letter $S_0[i_1]$ such that it has distance $\leq a$ to L . After the first letter position is fixed, there are at most $\prod_{j=1}^{j=k-1} (\alpha(i_{j+1} - i_j))$ ways to put the rest of them along the separation line with distance $\leq a$, where $\alpha < 2(a + \sqrt{2})$ is a constant (by Lemma 9). Since $k \leq d\sqrt{n}$ and $1 \leq i_1 < i_2 < \dots < i_k \leq n_0$, $\prod_{j=1}^{j=k-1} (\alpha(i_{j+1} - i_j)) \leq (\alpha(\frac{n_0}{k}))^k \leq (\frac{\alpha}{d})^{d\sqrt{n}} n_0^{d\sqrt{n}} n^{-\frac{1}{2}d\sqrt{n}}$ (We use the well known

fact that for positive variables y_1, \dots, y_k and fixed h with $y_1 + \dots + y_k \leq h$, the product $\prod_{t=1}^k y_t$ is maximal when $y_1 = y_2 = \dots = y_k = h/k$. The number of ways to arrange the k letters along the separation line (with distance $\leq a$ to L) is bounded by $u_3 = (2a+1)\sqrt{2}(m+1)(\frac{\alpha}{d})^{d\sqrt{n}}n_0^{d\sqrt{n}}n^{-\frac{1}{2}d\sqrt{n}}$.

We have $T(n) \leq u_1 \cdot u_2 \cdot u_3 \cdot T(cn)$. It implies that $T(n) \leq (\frac{mn}{\delta})^{c_0 \log n} 2^{c_0 \sqrt{n}} n_0^{d(\frac{1}{1-\sqrt{\epsilon}})\sqrt{n}} = O(m^{c_0 \log n} n_0^{(6.145-\epsilon)\sqrt{n}})$ by selecting constants ϵ, δ small enough, and c_0 large enough. \blacksquare

Theorem 16. *There is a $O(n^{6.145\sqrt{n}})$ time algorithm for the 2D protein folding problem in the HP-model.*

Proof: Fix the two middle letters on the two central neighbor positions of an $n \times n$ square. Let the folding be inside the $n \times n$ square, and apply Lemma 15. \blacksquare

3.2. 3-dimension

The technology used in the last section can be easily extended to the 3-dimensional grid. We give a brief proof for the case in 3-dimensional space.

Lemma 17. *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 $(\frac{3}{4\pi})^{\frac{1}{3}}n^{\frac{1}{3}} + 6a$ contains at least n grid points.*

Proof: 1) Let $r_1 = r + a$, and $r_2 = r - a$. 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 \geq the number of unit grid cubes intersecting the boundary of the sphere of radius r . 2) For $r = (\frac{3}{4\pi})^{\frac{1}{3}}n^{\frac{1}{3}} + 6a$, we have $\frac{4}{3}\pi r^3 - \frac{4}{3}\pi(6ar^2 + 2a^3) \geq n$. \blacksquare

Definition 18. 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 3-dimensional space.

Lemma 19. *Let $a > 0$ be a constant and $\delta > 0$ be a small constant. Let P be a set of points on a 3-dimensional grid. Assume that all 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 $\leq a$ to L is bounded by $4ar^2 + \delta r^2$ for all large r .*

Proof: The proof is very similar to that of Lemma 12. Let S be the sphere with radius r and center $o = (x_0, y_0, z_0)$ such that it contains all points in P . Let S' be the sphere of radius $r' = r + \frac{\sqrt{3}}{2}$ and with the same center as S . All of 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 $\frac{1}{\pi} \int \int \int_{S'} \frac{2a}{\text{dist}(a, o, p)} dx dy dz$. By the transformation $x = \rho \sin \theta \cos \alpha + x_0, y = \rho \sin \theta \sin \alpha + y_0, z = \rho \cos \theta + z_0$, we have $\frac{1}{\pi} \int \int \int_{S'} \frac{2a}{\text{dist}(a, o, p)} dx dy dz = \frac{2}{\pi} \int_0^{r'} \int_0^\pi \int_0^{2\pi} \frac{a\rho^2 \sin \theta}{\rho} d\alpha d\theta d\rho = 4ar'^2$. \blacksquare

Lemma 20. *Let $a > 0$ be a constant and $\epsilon > 0$ be a small constant. For a set P of n points in a 3-dimensional grid, there is a plane L such that P has at most $(4a(\frac{3}{4\pi})^{2/3}) \cdot n^{2/3} + \epsilon n^{2/3}$ points with distance $\leq a$ to L , and each half space divided by L has at most $\frac{3}{4}n$ points from P .*

Proof: By Lemma 17, the sphere of radius $(\frac{3}{4\pi})^{\frac{1}{3}}n^{\frac{1}{3}} + 6\sqrt{3}$ contains at least n grid points. Moving points of P into the sphere, which has center at the centerpoint of P (see Lemma 7), from the outside increases the probability to have distance $\leq a$ to a random plane through the sphere center. By Lemma 19, the expected number of points in P with distance $\leq a$ to a random plane is $(4a(\frac{3}{4\pi})^{2/3}) \cdot n^{2/3} + \epsilon n^{2/3}$ for all large n via selecting small δ . \blacksquare

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. Those points are called ϵ -regular points. Every 3 ϵ -regular points determine an ϵ -regular plane.

Lemma 21. *Let a and ϵ be positive constants. Let P be a set of n points in a 3-dimensional grid. There is an ϵ -regular plane such that there are $\leq (\frac{3}{4} + \epsilon)n$ points on each side, and $4(a + \epsilon)(\frac{3}{4\pi})^{2/3}n^{2/3}$ points to have distance $\leq a$ to it.*

Proof: Let L be the plane of Lemma 20. Let H be the area of intersection between plane L and the six faces of the $m \times m \times m$ -cube that contains all 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 non-collinear points such that p'_i has distance $\leq \delta$ to p_i for $i = 1, 2, 3$. Use the δ -plane determined by p'_1, p'_2 and p'_3 (by selecting small enough δ). ■

Lemma 22. *For some positive constant c_0 and $\epsilon > 0$, there is a $O(m^{c_0 \log n} n^{-4.407n^{2/3}} n_0^{(8.813 - \epsilon)n^{2/3}})$ time algorithm for the 3-dimensional Multi-Sequence Folding problem in an $m \times m \times m$ cube, where n is the sum of 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 $d = 4(a + \delta)(\frac{3}{4\pi})^{2/3}$. As Lemma 21, let $u_1 = \binom{8}{3}(\frac{m}{\delta})^6$, $u_2 = (\frac{3}{d})^{d\sqrt{n}} \cdot d\sqrt{n} \cdot n^{\frac{1}{2}d\sqrt{n}}$ and $u_3 = (\frac{\alpha}{d})^{2d\sqrt{n}} n^{-dn^{\frac{2}{3}}} n_0^{2dn^{\frac{2}{3}}}$. We have $T(n) \leq u_1 \cdot u_2 \cdot u_3 \cdot T(cn)$. This implies that $T(n) = (mn)^{c_0 \log n} 2^{c_0 n^{\frac{2}{3}}} n^{-\frac{2}{1-c^{2/3}}n^{2/3}} n_0^{\frac{2}{1-c^{2/3}}n^{2/3}}$ for some constant $c_0 > 0$. ■

Theorem 23. *There is a $O(n^{4.306n^{2/3}})$ time algorithm for the 3-dimensional protein folding problem in the HP-model.*

Proof: Fix the two middle letters on 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 22. ■

4. Acknowledgement

We are grateful to Mahdi Abdelguerfi, Padmanabhan Mahadevan and Seth Pincus for the helpful discussions during this research.

References

- [1] R. Agarwala, S. Batzoglou, V. Dancik, SE. Decatur, S. Hannenhalli, M. Farach, M. Muthukrishnan, S. Skiena, Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model. *Journal of Computational Biology* 4: 275-296, 1997.
- [2] N. Alon, P.Seymour, and R.Thomas, Planar Separator, *SIAM J. Discr. Math.* 7,2(1990) 184-193.
- [3] R. Backofen, Constraint techniques for solving the protein structure prediction problem, *Proceedings of 4th International conference on principle and practice of constrain programming*, 1998, *Lecture Notes in Computer Science*, 72-86, Springer-Verlag.
- [4] B. Berger and T. Leighton, Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete, *Journal of Computational Biology*, 5(1998), 27-40.
- [5] F. E. Cohen and M. J. E. Sternberg, On the prediction of protein structure: the significance of the root-mean-square deviation, *J. Mol. Biol.*, 138(1980), 321-333.
- [6] P. Crescenzi and D. Goldman and C. Papadimitriou and A. Piccolboni and M. Yannakakis, On the complexity of protein folding, *Journal of computational biology*, 5(1998), 423-465.
- [7] U.Bastolla, H. Frauenkron, E. Gerstner, P.Grassberger, and Nadler, Testing a new Monte Carlo algorithm for protein folding, *Protein: Structure, Function, and Genetics*, 32(1998), 52-66.

- [8] A. Godzik and J. Skolnick and A. Kolinski, Regularities in interaction patterns of globular proteins, *Protein Engineering*, 6(1993), 801-810.
- [9] A. Godzik and J. Skolnick and A. Kolinski, A topology fingerprint approach to inverse protein folding problem, *J. Mol. Biol.*, 227(1992), 227-238.
- [10] W. E. Hart and S. Istrail, Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal, *Proceedings 27th ACM symposium on the theory of computing*, 1995.
- [11] L. Holm and C. Sander, Mapping the protein universe, *Science*, 273(1996), 595-602.
- [12] M. Khimasia and P. Coveney, Protein structure prediction as a hard optimization problem: The genetic algorithm approach, In *Molecular Simulation*, 19(1997), 205-226.
- [13] N. Krasnogor, D. Pelta, P.M. Lopez, P. Mocchiola, and E. De la Canal, Genetic algorithms for the protein folding problem: A critical view, In C.F.E. Alpaydin, editor, *Proceedings of Engineering of Intelligent Systems*. ICSC Academic Press, 1998.
- [14] K. F. Lau and K. A. Dill, A lattice statistical mechanics model of the conformational and sequence spaces of proteins, *Macromolecules*, 22(1989), 3986-3997.
- [15] K. F. Lau and K. A. Dill, Theory for protein mutability and biogenesis, *Proc. Natl. Acad. Sci.*, 87(1990), 638-642.
- [16] R. J. Lipton and R. Tarjan, A separator theorem for planar graph, *SIAM J. Appl. Math.* 36(1979) 177-189.
- [17] F. Liang and W.H. Wong, Evolutionary Monte Carlo for Protein folding simulations, *Journal of Chemical Physics*, 115,7(2001), 3374-3380.
- [18] A. Newman, A new algorithm for protein folding in the HP model, *Proceedings 13th ACM-SIAM Symposium on Discrete Algorithms*, 2002, 876-884.
- [19] A. Patton, W.P.III, and E. Goldman, A standard ga approach to native protein conformation prediction, In *Proc 6th Intl Conf Genetic Algorithms*, Morgan Kaufman, 1995, 574-581.
- [20] J. Pach and P.K. Agarwal, *Combinatorial Geometry*, Wiley-Interscience Publication, 1995.
- [21] A. Piccolboni and G. Mauri, Application of evolutionary algorithms to protein prediction, In N. e. a. Kasabov, editor, *Proceedings of I-CONIP'97*, Springer, 1998.
- [22] A.A. Rabow and H.A. Scheraga, Improved genetic algorithm for the protein folding problem by use of a cartesian combination operator. *Protein Science*, 5(1996), 1800-1815.
- [23] R. Ramakrishnan, B. Ramachandran, and J.F. Pekney, A dynamic Monte Carlo algorithm for exploration of dense conformation spaces in heteropolymers, *Journal of Chemical Physics*, 106(1997), 2418.
- [24] W. D. Smith and N. C. Wormald, Application of geometric separator theorems, *FOCS 1998*, 232-243.
- [25] A. Sali, E. Shakhnovich, M. Karplus, How does a protein fold? *Nature*, 369(1994), 248-251.
- [26] U. Unger and J. Moult, A Genetic algorithm for three dimensional protein folding simulations, In *Proc 5th Intl Conf on Genetic Algorithms*, 1993, 581-588.
- [27] U. Unger and J. Moult, Genetic algorithms for protein folding simulations, *Journal of Molecular Biology*, 1993, 231(1),75-81
- [28] K. Yue and K. A. Dill, Sequence-structure relationships in proteins and copolymers, *Physical Review E* 48(1993), 2267-2278.