5-20-2005

A Multi-Dimensional Width-Bounded Geometric Separator and its Applications to Protein Folding

Sorinel Oprisan
University of New Orleans

A MULTI-DIMENSIONAL WIDTH-BOUNDED GEOMETRIC SEPARATOR AND ITS APPLICATIONS TO PROTEIN FOLDING

A Thesis

Submitted to the Graduate Faculty of the
University of New Orleans
in partial fulfillment of the
requirements for the degree of

Master of Science
in
Computer Science

by

Sorinel Adrian Oprisan

B.S. "Alexandru Ioan Cuza" University of Iasi, Romania, 1987
Ph.D. "Alexandru Ioan Cuza" University of Iasi, Romania, 1998

May 2005

Dedication

This thesis is dedicated to my wife, Ana Oprisan, to my son Andrei Oprisan, and to my daughter Andra Oprisan.

Acknowledgments

I would like to give my sincere thanks to Dr. Bin Fu for his guidance, patience and supervision during my graduate work.

I express my gratitude to the committee members Dr. Adlai DePano and Dr. Yixin Chen for their help and support.

Contents

Abstract

We used a divide-and-conquer algorithm to recursively solve the two-dimensional problem of protein folding of an HP sequence with the maximum number of H-H contacts. We derived both lower and upper bounds for the algorithmic complexity by using the newly introduced concept of multi-directional width-bounded geometric separator. We proved that for a grid graph G with n grid points P, there exists a balanced separator $A \subseteq P$ such that A has less than or equal to $1.02074\sqrt{n}$ points, and G-A has two disconnected subgraphs with less than or equal to $\frac{2}{3}n$ nodes on each subgraph. We also derive a $0.7555\sqrt{n}$ lower bound for our balanced separator. Based on our multi-directional width-bounded geometric separator, we found that there is an $O(n^{5.563\sqrt{n}})$ time algorithm for the 2D protein folding problem in the HP model. We also extended the upper bound results to rectangular and triangular lattices.

Chapter 1

Introduction

In the year 2000, a distinguished panel of renowned scientists at the U.S. National Research Council identified six fundamental challenges to the scientific community [1]: (1) Developing quantum technologies, (2) Understanding complex systems, (3) Applying physics to biology, (4) Creating new materials, (5) Exploring the Universe, and (6) Unifying the forces of Nature. For computational biology, "Current challenges include [...] the (study of) mechanical and electrical properties of DNA and the enzymes essential for cell division and all cellular processes."

The behavior of complex systems, such as the proteins, depends crucially on the molecular details and therefore it seems unlikely that the traditional reductionist way would succeed in this field. For example, small perturbations of a protein's environment such as alterations of the pH or substitutions of just one amino acid in the chain might change dramatically the folding process and the biological activity of the protein. The allosteric proteins which drastically alter their shape and properties when they link a small regulating molecule (like a vitamin) are a good example of sensitivity of the global structure to small molecular details.

1.1 Proteins and amino acids

Protein etymology comes from the Greek word "proteios", which means first. Next to water, proteins make up the second greatest portion of a person's body weight.

Proteins are substances, which makeup muscles, tendons, ligaments, organs, glands, nail, hair, vital body fluids, and bones and have the general purpose of holding together, protecting, and providing structure to the body of a multicellular organism. Besides the structural component, specific proteins such as enzymes, hormones, antibodies, and globulins have

the essential role to catalyze, regulate, and protect the cell's chemistry. For example, the hemoglobin, myoglobin and various lipoproteins are responsible for the transport of oxygen and other substances within an organism. Proteins are generally regarded as beneficial, and are a necessary part of the diet of all animals. However, some proteins such as the venoms of many snakes, and ricin (extracted from castor beans), are extremely toxic. A teaspoon of botulinum toxin A, from Clostridium botulinum, would be sufficient to kill a fifth of the world's population. The toxins produced by tetanus and diphtheria microorganisms are nearly as poisonous. Allergies are generally caused by the effect of foreign proteins on our body. Proteins that are ingested are broken down into smaller peptides and amino acids by digestive enzymes called "proteases". Allergies to foods may be caused by the inability of the body to digest specific proteins. Cooking denatures (inactivates) dietary proteins and facilitates their digestion. Allergies or poisoning may also be caused by exposure to proteins that bypass the digestive system by inhalation, absorption through mucous tissues, or injection by bites or stings.

From a chemical point of view, proteins' composition is significantly different compared to the carbohydrates and lipids. Lipids are largely hydrocarbon in nature, generally being 75-85% carbon. Carbohydrates are roughly 50% oxygen, and like lipids, usually have less than 5% nitrogen (often none at all). Proteins, on the other hand, are composed of 15-25% nitrogen and about an equal amount of oxygen.

Proteins consist of amino acids which are characterized by the -CH(NH2)COOH substructure (Figure 1.1A). Nitrogen and two hydrogen atoms comprise the amino group, $-NH_2$, and the acid entity is the carboxyl group, $-COOH$. Amino acids are the basic building blocks of enzymes, hormones, proteins, and body tissues. A peptide is a compound consisting of 2 or more amino acids. Oligopeptides have 10 or fewer amino acids. Polypeptides and proteins are chains of 10 or more amino acids, but peptides consisting of more than 50 amino acids are classified as proteins.

Amino acids link to each other when the carboxyl group of one molecule reacts with the amino group of another molecule, creating a peptide bond $-C(=O)NH-$ and releasing a molecule of water (Figure 1.1B).

There are twenty different amino acids characterized by variations in their side chain (Figure 1.2). Some amino acids are called essential because they cannot be derived from other amino acids and must be supplied in the diet (isoleucine, histidine, leucine, lysine, methionine, threonine, tryptophan, valine).

There are two broad classes of amino acids based upon whether the R-group is hydrophobic or hydrophilic. The hydrophobic amino acids tend to repel the aqueous environment and,

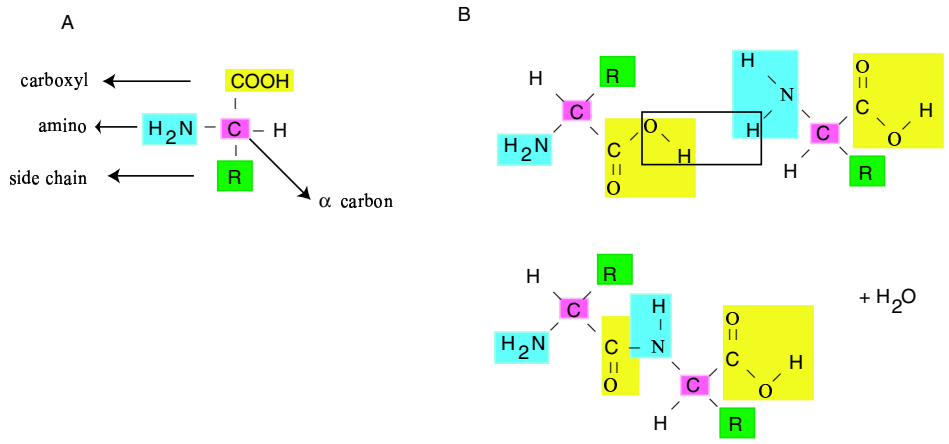

Figure 1.1: (A) Planar structure of amino acids contains an amino group ($-NH_2$), the carboxyl group (-COOH) and the α carbon. (B) Peptide bond created by interaction of amino group of one molecule with the carboxyl groups of another molecule after releasing one molecule of water.

therefore, reside predominantly in the interior of proteins. This class of amino acids does not ionize nor participate in the formation of H-bonds. The hydrophilic amino acids tend to interact with the aqueous environment, are often involved in the formation of H-bonds and are predominantly found on the exterior surfaces proteins or in the reactive centers of enzymes.

Among the well-known peptide hormones, we mention vasopressin, which contains 9 amino acids and increases the reabsorption rate of water in kidneys; insulin, which contains 51 amino acids and is involved in lowering the blood glucose level; growth hormone, which contains 191 amino acids and regulates development of the body [15, 42, 55].

1.2 Proteins structure

Proteins have multiple structural levels. The most basic structure of proteins is called the **primary structure**, which is simply the order of its amino acids. Note that by convention, the order of amino acids in a protein is always written from the amino group end to the carboxyl group end.

Secondary Structure

Proteins' secondary structure refers to certain common repeating structures found in proteins such as the alpha-helix, beta-pleated sheet, turns, and random coil. An alpha-helix is a tight helix formed out of the polypeptide chain (Figure 1.3). The polypeptide main chain

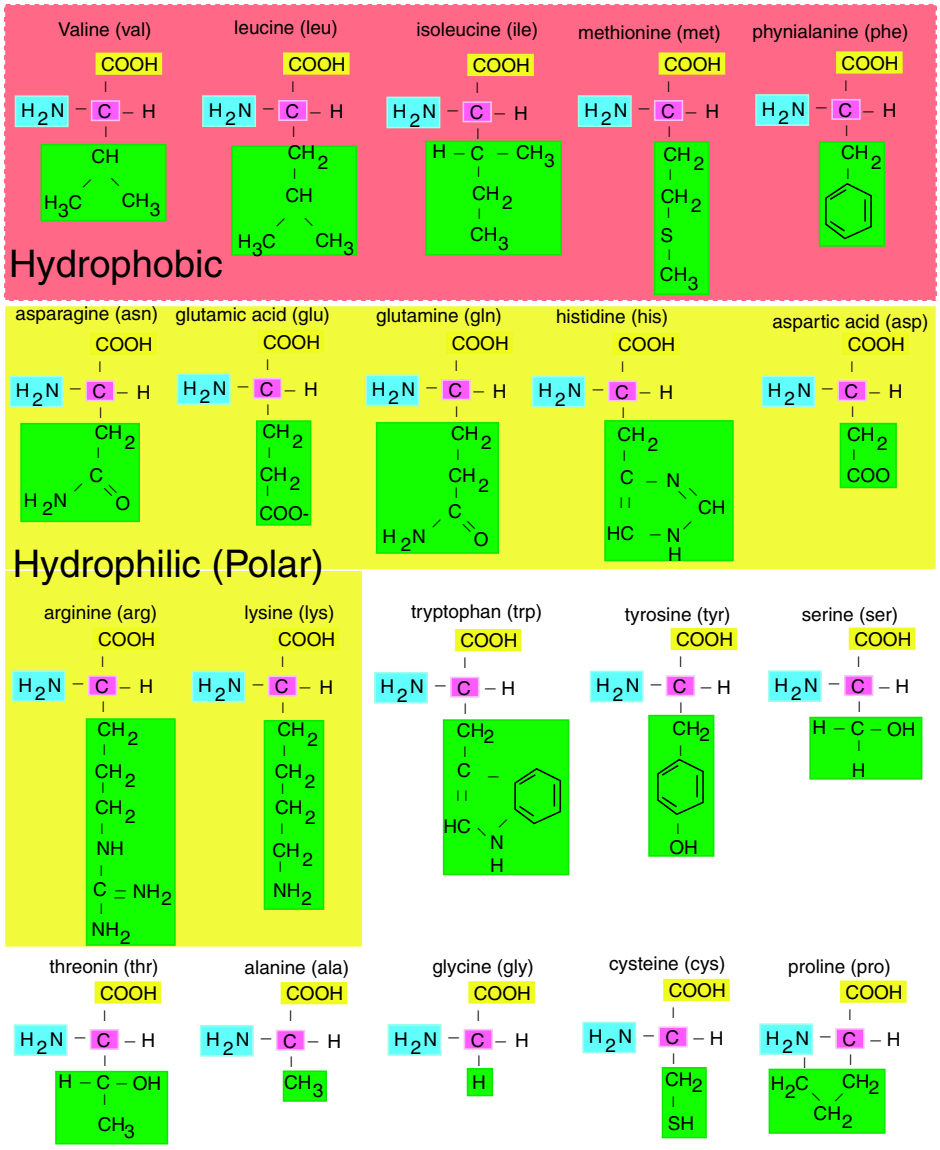

Figure 1.2: Structural formulas for the twenty natural amino acids.

makes up the central structure, and the side chains extend out and away from the helix. The carboxyl group of one amino acid (n) is hydrogen bonded to the amino group of the amino acid four residues away (n+4). Alpha-helix structure was first postulated by Linus Pauling (Nobel Prize for chemistry in 1954), Robert Corey, and Herman Branson in 1951 based on the known crystal structures of amino acids and peptides and Pauling's prediction of planar peptide bonds [24].

Beta-pleated sheets consist of two or more amino acid sequences within the same protein

1	A A S X D X S L V E V H X X V F I V P P X I L Q A V V S I A
31	T T R X D D X D S A A A S I P M V P G W V L K Q V X G S Q A
61	G S F L A I V M G G G D L E V I L I X L A G Y Q E S S I X A
91	S R S L A A S M X T T A I P S D L W G N X A X S N A A F S S
121	X E F S S X A G S V P L G F T F X E A G A K E X V I K G Q I
151	T X Q A X A F S L A X L X K L I S A M X N A X F P A G D X X
181	X X V A D I X D S H G I L X X V N Y T D A X I K M G I I F G
211	S G V N A A Y W C D S T X I A D A A D A G X X G G A G X M X
241	V C C X Q D S F R K A F P S L P Q I X Y X X T L N X X S P X
271	A X K T F E K N S X A K N X G Q S L R D V L M X Y K X X G Q
301	X H X X X A X D F X A A N V E N S S Y P A K I Q K L P H F D
331	L R X X X D L F X G D Q G I A X K T X M K X V V R R X L F L
361	I A A Y A F R L V V C X I X A I C Q K K G Y S S G H I A A X
391	G S X R D Y S G F S X N S A T X N X N I Y G W P Q S A X X S
421	K P I X I T P A I D G E G A A X X V I X S I A S S Q X X X A
451	X X S A X X X A

Table 1.1: The primary structure of the sequence of yeast hexokinase from the yeast species Saccharomyces cerevisiae (http://www.pdb.bnl.gov/pdb-bin/pdbmain). The letters represent abbreviated notations for the corresponding amino acids (see Figure 1.2).

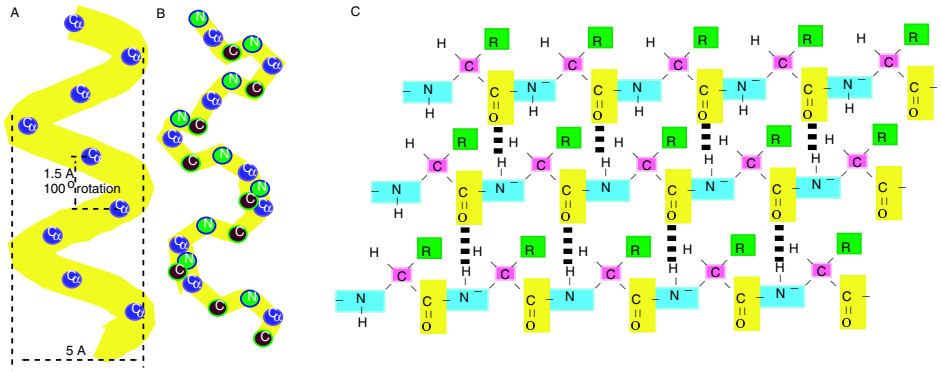

Figure 1.3: Secondary structures of proteins. (A) Alpha-helix backbone formed by α carbons and (B) a more detailed view of alpha-helix secondary structure including nitrogen atom of the amino group and carbon atom of the carboxyl group. (C) Beta-pleated sheet secondary characterized by hydrogen bonds between hydrogen atoms of amino group and the oxygen atom of carboxyl group with a periodicity of three atoms.

that are arranged adjacently and in parallel, but with alternating orientation such that hydrogen bonds can form between the two strands (Figure 1.3). The amino groups in the backbone of one strand establish hydrogen bonds with the carboxyl groups in the backbone of the adjacent, parallel strand(s).

Tertiary Structure

Tertiary structure is the full 3-dimensional folded structure of the polypeptide chain (Figure 1.4).

Quaternary Structure

Quaternary structure is only present if there is more than one polypeptide chain and represents the interconnections and organization of the peptides.

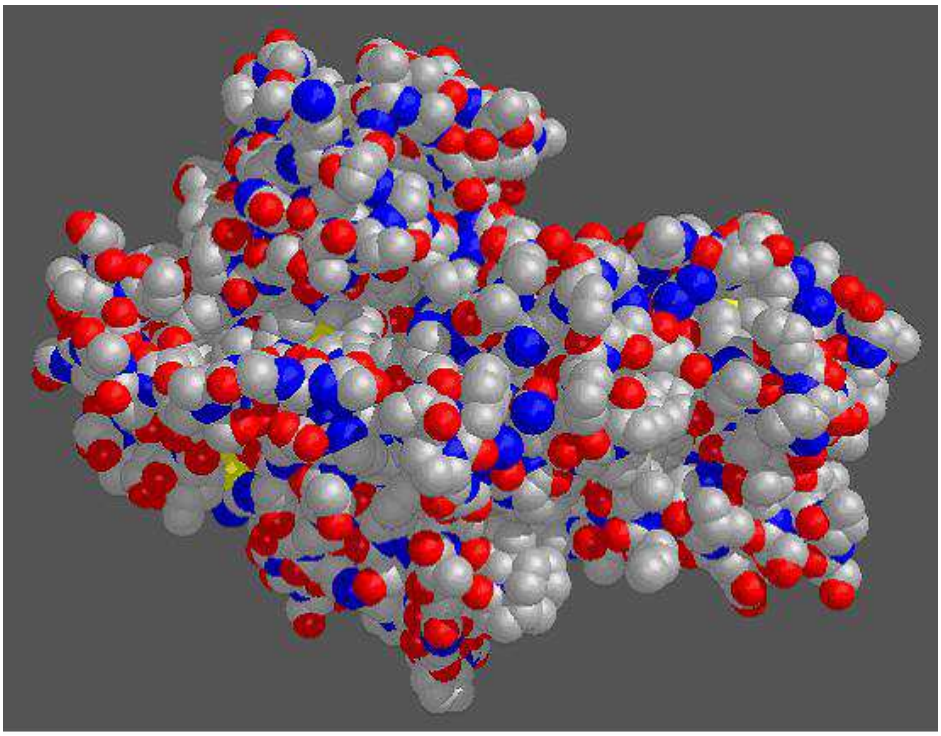

Figure 1.4: Tertiary structure of hexokinase.

1.3 Protein folding

The study of protein synthesis was for many years marked by the so-called "blind watchmaker paradox" [18] and it equates the vastness of the sequence space of polypeptides with the impossibility of ever finding a protein-like sequence. For example, a protein containing only 100 natural amino acids has $20^{100} \approx 10^{130}$ possible sequences. Therefore, the probability of observing such a protein by chance is negligible. This problem has been regarded as impossible to solve by creationists, who appeal to divine intervention and has been circumvented by evolutionists through the mechanisms of natural selection. The problem with

the "paradox" is that the probability to obtain an amino acid sequence that folds to the desired 3D shape is much higher than 10^{-130} due to an enormous degeneracy of the sequence space. It turns out that the probability to obtain a certain 3D conformation is of the order of 10^{-10} [13, 20], which is still small but is reasonable to assume that it could have happened by natural selection. Moreover, exact models [7] show that the precise information of the sequence is, most of the times, redundant. It has been found that the fold is primarily determined by the sequence written in a two-letter alphabet (hydrophobic (H) and polar (P)) rather than in the natural twenty-letter alphabet. Using this code, it was found that, if a certain sequence does fold, the sequence obtained by interchanging one hydrophobic amino acid for another hydrophobic amino acid (similarly for polar ones) will fold with a very high probability to a very similar structure. Thus, the essential features of the full $20^{100} \approx 10^{130}$ sequences space remain in the smaller space of the sequences written in the HP alphabet, which contains only $2^{100} \approx 10^{30}$ elements.

The second "paradox" of protein folding regards the folding time and is known as the Levinthal paradox [34]. If the protein scans the whole configuration space during folding then the protein will never fold to its native structure. For example, even for a small protein containing only 100 amino acids it can take up to 10 different conformations on average. This makes a total of 10^{100} different conformations for the chain. If the conformations were sampled in the shortest possible time, which is about 10^{-13} s, one would need more then 10^{77} years to sample all the conformational space. This result implies that the protein folding cannot be a completely random trial-and-error process and we must explain how the system can scan such a huge conformation space in going from the unfolded state to the native conformation in such a short time.

The goal of protein folding study is to determine how proteins so consistently fold into a stable state and to understand the complete dynamics and/or chemical changes involved in going from an unfolded linear state into a compact folded state. Although naturally posed as a numerical simulation, there are several problems of scale, including the small energy differences between folded and unfolded states, and the extremely short interval (approximately 10^{-15} seconds) for which the dynamics equations remain valid, compared to the milliseconds to seconds over which the folding takes place [16]. The thermodynamic hypothesis, first developed by Anfinsen [7], proposes that proteins fold to a minimum energy state. This motivates the attempt to predict protein folding by solving certain optimization problems. There are two main difficulties with this approach: there is as yet no scientific consensus on what the precise energy function to be minimized might be, and the functions commonly used lead to extremely difficult optimization problems [40].

1.4 HP model of protein folding

A protein can fold into a specific 3D structure, which is uniquely determined by the sequence of amino acids. One of the most important problems in molecular biology is determining a protein's 3D structure from its amino acid sequence. A protein's 3D structure determines its function. The standard procedure to determine a 3D structure is to purify the protein and crystallize it, followed by x-ray crystallography. It is a very time consuming process and not every protein can be crystallized. Therefore, protein structure prediction with computational technology is one of the most significant problems in bioinformatics.

One of the most popular models of protein folding is the hydrophobic-hydrophilic (HP) model [13, 19, 31]. In the HP model, only two types of monomers are distinguished: hydrophobic (H), which tend to bundle together to avoid surrounding water, and polar or hydrophilic (P), which are attracted to water and are frequently found on the surface of a folding [13]. These monomers are strung together in some combination to form an HP chain, either an open chain (path or arc) or a closed chain (cycle or polygon).

Usually, the proteins are folded onto the regular square lattice. More formally, a lattice embedding of a graph is a placement of vertices on distinct points of the (regular square) lattice such that each edge of the graph maps to two adjacent (unit-distance) points on the lattice. The space in which the protein folds is discretized by defining a lattice and requiring residues to lie only on lattice points. Residues which are adjacent in the primary sequence (i.e. covalently linked) must be placed at adjacent points in the lattice. A fold of a protein is a self-avoiding walk along the lattice. A contact between two residues is a topological contact if they are not covalently linked and there is an edge connecting the lattice points of the two residues. The free energy of a folded protein in the HP model is defined to be $(-1)\times$ the number of topological contacts between pairs of hydrophobic residues. The target fold for the protein is the one which has the lowest free energy. Intuitively, if a protein is folded to bring together many hydrophobic monomers (H nodes), then those monomers are hidden from the surrounding water as much as possible (Figure 1.5). An optimal embedding is one that maximizes the number of H-H contacts. This combinatorial model is attractive in its simplicity, and captures essential features of protein folding such as the tendency for the hydrophobic components to fold to the center of a globular (compactly folded) protein [13]. Unlike more sophisticated models of protein folding, the main goal of the HP model is to explore broad qualitative questions about protein folding such as whether the dominant interactions are local or global with respect to the chain [20].

The HP model was originally developed for square (2D) or cubic (3D) lattices because

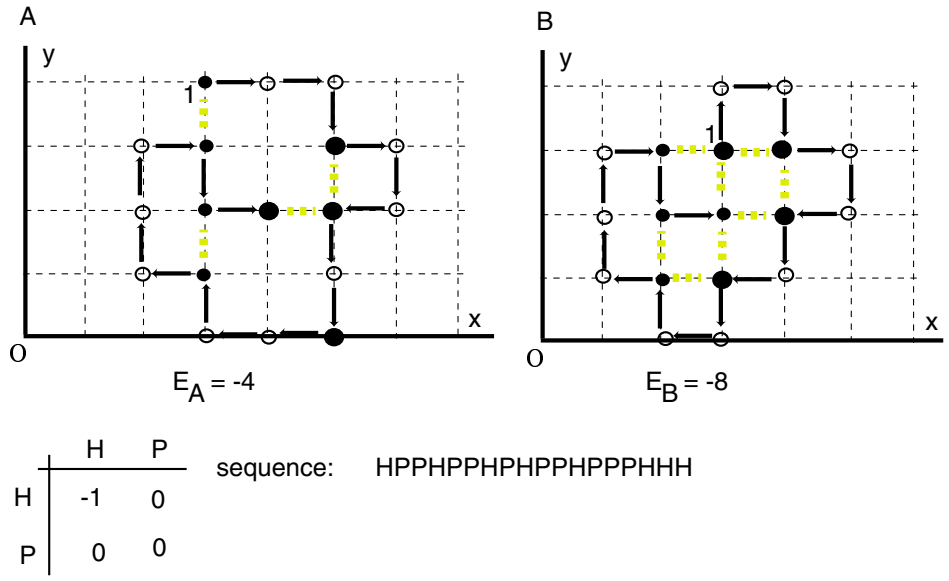

Figure 1.5: Folding of an HP sequence in a rectangular lattice. An H monomer is represented by a filled circle and a P monomer by an open circle. The free energy of an H-H topological contact (not covalently linked H monomers) is -1 and for any other topological contacts is zero. The free energy of the configuration (A) can be increased by a tighter packing of the hydrophobic core (B).

of the relative simplicity of the configuration space. However, square lattice configurations suffer from the so-called "parity problem" in which two residues of even distance from each other in the primary sequence cannot be placed in contact with each other regardless of how one arranges the intervening sequence (Figure 1.6). This parity restriction is clearly an artificial limitation introduced by the specific symmetry of the embedding and is not present when considering the real folding of proteins. For this reason, we also consider protein folding in the HP model on triangular lattices which does not exhibit the parity problem (Figure 1.6). We also note that the free energy of a configuration strongly depends on the symmetry of the embedding.

In theoretical computer science, Berger and Leighton [8, 11] proved NP-completeness of finding the optimal folding in 3D, and Crescenzi et al. [17] proved NP-completeness in 2D.

Some algorithms for this problem have been developed based on the heuristic [9, 50], genetic algorithm [32, 33, 44, 45, 47, 53, 54], Monte Carlo [10, 35, 48, 57], branch and bound methods. Although 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.

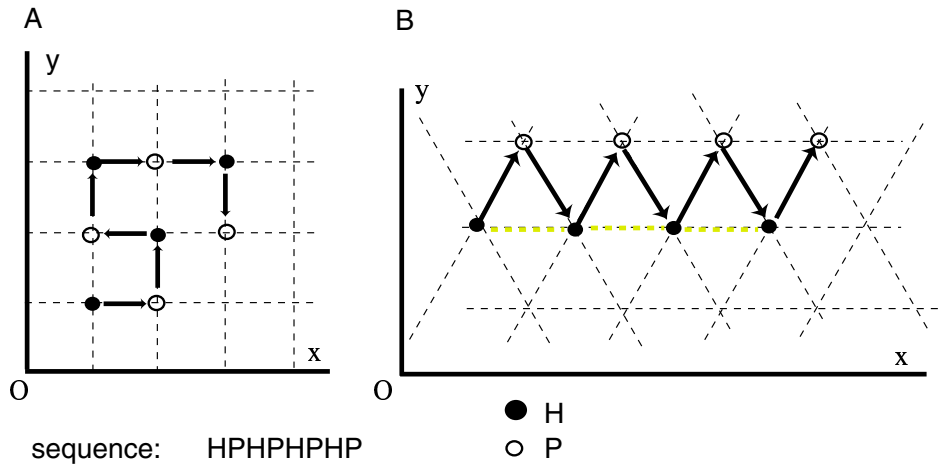

Figure 1.6: (A) Folding of an HP sequence in a rectangular lattice. For the particular case of $(HP)^n$ sequence there is no topological contact between H monomers. (B) Folding of the same HP sequence in a triangular lattice leads to a significantly lower free energy of the configuration.

Another approach is to develop polynomial time approximation algorithms for the protein folding in the HP model [2–4, 30, 41]. Hart and Istrail [30] have developed a 3/8-approximation in 3D and a 1/4- approximation in 2D of the number of H-H contacts in the HP model. Agarwala et al. [3] developed constant-factor approximation algorithms for a generalized HP model allowing multiple levels of hydrophobicity in the 2D triangular lattice and the 3D face-centered cube lattice. Newman [41] derived a polynomial time 1/3 -approximation algorithm for the 2D problem.

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). As the average number of amino acids of proteins is between 400 to 600, if an algorithm could solve the protein structure prediction with about 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 uses the divide-and-conquer approach, which is based on our geometric separator for the points on a 2D-dimensional grid. Lipton and Tarjan [36] 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 [5, 6]

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 [51].

Divide-and-conquer approach on HP model

The divide-and-conquer method is one of the most fundamental techniques in the area of algorithm design. This method divides a problem into several smaller problems. The solutions for those smaller problems are merged to obtain a solution for the larger problem. The speed of the divide-and-conquer algorithm depends on the efficiency of the problem decomposition, which is often related to separator technology. The geometric separator is a basic tool in the divide-and-conquer algorithms for many problems (e.g. [12, 14, 36, 49]). Lipton and Tarjan [36] showed that every n–vertices 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 was improved to $\sqrt{6n}$ by Djidjev [22], to $\sqrt{5n}$ by Gazit [28], to $\sqrt{4.5n}$ by Alon, Seymour and Thomas [5], and to $1.97\sqrt{n}$ by Djidjev and Venkatesan [21]. Spielman and Teng [52] found a $\frac{3}{4}$-separator with size $1.82\sqrt{n}$ for planar graphs. The separators for more general graphs were developed by Gilbert, Hutchingson, Tarjan [29], Alon, Seymour, Thomas [6], and Plotkin, Rao and Smith [46]. Some other forms of the geometric separators were studied by Miller, Teng, Thurston, and Vavasis [38, 39, 39] and by Smith and Wormald [51]. If each of n input points is covered by at most k regular geometric object such as circles, rectangles, etc, then there exist $O(\sqrt{k \cdot n})$ size separators [37–39, 51]. In particular, Smith and Wormald obtained the separator of size $4\sqrt{n}$ for the case $k = 1$. The lower bounds $1.555\sqrt{n}$ and $1.581\sqrt{n}$ for the $\frac{2}{3}$-separator for the planar graph were proven by Djidev [21], and by Smith and Wormald [51], respectively.

Each edge in a grid graph connects two grid points of distance 1 in the set of vertices. Thus a grid graph is a special planar graph. Fu and Wang [27] developed a method for deriving sharper upper bound separator for grid graphs by controlling the distance to the separator line. Their separator is determined by a straight line on the plane and the set of grid points with distance less than or equal to $\frac{1}{2}$ to the line. They proved that for an n–vertices grid graph on the plane, there is a separator that has less than or equal to $1.129\sqrt{n}$ grid points and each of two disconnected subgraphs has at most $\frac{2}{3}n$ grid points. Using this separator and their approximation to the separator line, they obtained the first $n^{O(n^{1-\frac{1}{d}})}$–time exact algorithm for the d-dimensional protein folding problem of the HP model. The method of Fu and Wang [27] was further developed and generalized by Fu [25] and applications were found to some other problems. The notion of width-bounded geometric separator was introduced by Fu [25]. For a constant $a > 0$ and a set of points Q on the plane, an a-wide separator is the region between two parallel lines of distance a that partitions Q into Q_1 (on the left side of the separator's region), S (inside the separator's region), and Q_2 (on the right side of the

separator's region).

The separator theorem for grid graph can be applied to many geometric problems [25] with arbitrary input points. Those problems, including the disk covering problem on the plane and the maximum independent set problem on disk graph, can be handled by combining the grid separator with the rounding method from arbitrary points to grid points, which merges the points in one 1×1 grid square to its top left grid point. An example of such an application is the disk-covering problem, which seeks to find the least number of fixed size discs to cover a set of points on the plane. Fu [25] derived a $2^{O(\sqrt{n})}$-time exact algorithm for it.

Chapter 2

Multi-Directional Width-Bounded Geometric Separator

For a set of points P on the plane and two vectors v_1 and v_2, the (a, b)-wide separator (along the directions v_1 and v_2) is the region of points that have no more than distance a to L along v_1 or no more than distance b to L along v_2, where L is a straight line (separator line) on the plane. The separator size is measured by the number of points from P in the region and the line L partitions the set P into two balanced subsets. In this dissertation we use this new method to improve the separator for the grid graph. The multi-directional width approach is different from that used in [25, 27], which only controls the regular distance to the middle line in the separator area. Pursuing smaller and more balanced separators is an interesting problem in combinatorics and also gives more efficient algorithms for divide-and-conquer applications. In this dissertation, we prove that for a grid graph G with n grid points P, there exists a separator subset $A \subseteq P$ such that A has up to $1.02074\sqrt{n}$ points, and $G - A$ has two disconnected subgraphs with up to $\frac{2}{3}n$ nodes on each of them. The original result we report here [26] improves the previous $1.129\sqrt{n}$ size separator for the grid graph [27]. We also prove a $0.7555\sqrt{n}$ lower bound for the size of the separators for grid graphs. Our lower bound is based on a result that the shortest curve partitioning a unit circle into two areas with ratio $1 : 2$ is a circle arc. Its length is less than that of the straight line partitioning the circle with the same ratio.

2.1. Overview of our method

Previously, Fu et al. [25, 27] controlled the distance to the separator line to derive the upper bound of the separator's size. Our current approach still uses Helly's theorem [23] derived in 1912 (see Lemma 2), which states that every line L through the center point of set P gives a balanced partition for it. If two grid neighbor points (of distance 1) are at different sides of L, one of them should have no more than $1/2$ vertical or horizontal distance to L. We compute the probability that a point p has a vertical or horizontal distance no more than $1/2$ to a random line L through the center of P. The sum of those probabilities is the expected number of points for the size of the separator, which is the upper bound of the optimal separator. We will show that the sum is maximal when grid points in P stay in the union of four circles' area (see the left of Figure 2.1). The sum is computed approximately via the integration at the four circles area and gives a smaller separator upper bound for the grid graph.

Our lower bound is based on the set of all grid points in a circle's area. If it is partitioned into two balanced areas of ratio $1 : 2$, each of the two areas is a connected grid graph if the length of the boundary surrounding the two grid graphs is minimal. This problem is converted into the problem of finding the shortest curve that partitions a circle into two areas with ratio $1 : 2$. Using the variational calculus method, we compute the length of the shortest curve with ratio $1 : 2$, which is a circle arc. Its length is less than that of the straight line to achieve $1 : 2$ partition ratio for the circle. If c_0 is the shortest length of the curve partitioning the unit circle into two areas with ratio $1 : 2$ then the lower bound of separator size can be roughly considered as $\frac{c_0}{\sqrt{2}} r_n$, where r_n is the radius of the circle C that contains n grid points. The denominator $\sqrt{2}$ corresponds to the case where the separator line goes along the diagonal direction which has the least number of grid points close to it.

With the improved separator for the grid graph, we derive an $O(n^{5.563\sqrt{n}})$ time exact algorithm for the 2D-protein folding problem in the HP model. The algorithm uses divide–and–conquer approach. The approximation line to the optimal separator is a nontrivial revision from that described in [27]. An exhaustive method is used for searching the arrangements of amino acids along the separator line and takes no more than $n^{c\sqrt{n}}$ cases, where c is proportional to the constant s such that $s\sqrt{n}$ is an upper bound of the separator size.

Section 2.2 proves a $1.0207\sqrt{n}$ size separator for the grid graph with n nodes. Section 2.3 gives a $0.7555\sqrt{n}$ lower bound for grid graph using the length of the shortest curve to partition the unit circle into two areas with ratio $1 : 2$, which is computed by the variational calculus method in section 2.4. Section 3 gives the improved exact algorithm for the 2D

protein folding problem in the HP model.

2.2. Separators upper bound for grid graphs

Definition 1. For a set A, $|A|$ denotes the number of elements in A. For two points p_1, p_2 in the d-dimensional space (R^d), $\text{dist}(p_1, p_2)$ is the Euclidean distance. For a set $A \subseteq R^d$, $\text{dist}(p_1, A) = \min_{q \in A} \text{dist}(p_1, q)$. The integer set is represented by $Z = \{\cdots, -2, -1, 0, 1, 2, \cdots\}$. For integers x_1 and x_2, (x_1, x_2) is a *grid point*. A *grid square* is an 1×1 square that has four grid points as its four corner points. For a set V of grid points on the plane, let E_V be the set of edges (v_i, v_j) (straight line segments) such that $v_i, v_j \in V$ and $\text{dist}(v_i, v_j) = 1$. Define $G = (V, E_V)$ as the *grid graph*. For $0 < \alpha < 1$, an α-separator for a grid graph $G = (V, E_V)$ is a subset $A \subseteq V$ such that $G - A$ has two disconnected areas $G_1 = (V_1, E_{V_1})$ and $G_2 = (V_2, E_{V_2})$ with $|V_1|, |V_2| \leq \alpha |V|$. For a 2D vector v, a *line* L in R^2 through a fixed point $p_0 \in R^2$ along the direction v corresponds to the equation $p = p_0 + tv$ that characterizes all the points p on L, where the parameter $t \in (-\infty, +\infty)$. For a point p_0 and a line L, the distance of p_0 to L along direction v is $\text{dist}(p_0, q)$, where q is the intersection between $p = p_0 + tv$ and L. Let $v_1, v_2, \cdots, v_k$ be k fixed vectors. A point p has distance $\leq (a_1, \cdots, a_k)$ to L along directions $v_1, v_2, \cdots, v_k$ if p has distance $\leq a_i$ along direction v_i for some $i = 1, \cdots, k$. In the rest of this paper, we use two vectors $v_1 = (1, 0)$ an $v_2 = (0, 1)$ to represent the horizontal and vertical directions, respectively. If a point p has distance $\leq (a, a)$ from a line L, it means that the point p has distance $\leq a$ from L along either direction $(1, 0)$ or $(0, 1)$ in the rest of this paper. Define $C(o, r) = \{(x, y) | \text{dist}((x, y), o) \leq r\}$, which is the disc area with center at point o and radius r. For $r > 0$, define $D(r)$ to be the union region of 4 discs $C((0, -r), r) \cup C((0, r), r) \cup C((-r, 0), r) \cup C((r, 0), r)$ (see the left of Figure 2.1). For a region R on the plane, define $G(R)$ to be the set of all grid points in the region R.

We will use the following well–known result (see [43]) to derive our width bounded separator.

Lemma 2. *(***Helly's Theorem***) For an n-element set P in a 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 center point of P).*

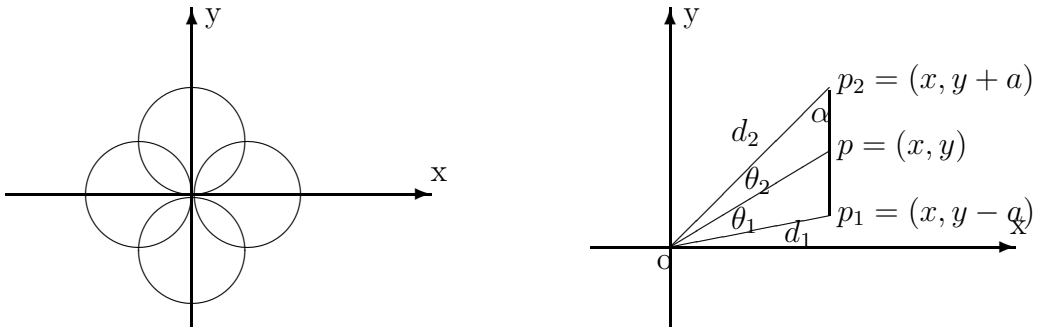

Figure 2.1: Left: Area of grid points with maximal expectation. Right: Probability analysis

Lemma 3. *Let P be a set of grid points on the plane and $(0,0) \notin P$. The sum $\sum_{p=(x,y)\in P} \max(\frac{|x|}{x^2+y^2}, \frac{|y|}{x^2+y^2})$ is maximal when $P \subseteq G(D(R))$, where R is the least radius with $|G(D(R))| \geq |P|$.*

Proof: Let L be the line segment connecting $o = (0,0)$ and $p = (x,y)$. If $p' = (x', y')$ is another point between o and p on the line L, we have $\frac{\max(|x|,|y|)}{\text{dist}(o,p)} = \frac{\max(|x'|,|y'|)}{\text{dist}(o,p')}$. Since $\text{dist}(o,p) > \text{dist}(o,p')$, we have $\frac{\max(|x|,|y|)}{\text{dist}(o,p)^2} < \frac{\max(|x'|,|y'|)}{\text{dist}(o,p')^2}$. For the constant c, let $\frac{|x|}{x^2+y^2} = c$ or $\frac{|y|}{x^2+y^2} = c$. We have $x^2 + y^2 - \frac{1}{c}|x| = 0$ or $x^2 + y^2 - \frac{1}{c}|y| = 0$. The two equations characterize the four circles of $D(\frac{1}{2c})$. All points on the external boundary $D(r)$ have the same value $\frac{\max(|x|,|y|)}{\text{dist}(o,p)^2}$. ∎

Let a be a constant > 0, p and o be two points on the plane, and P be a set of points on the plane. We define the function

$$f_{p,o,a}(L) = \begin{cases} 1 & \text{if } p \text{ has } \leq (a,a) \text{ distance to the line } L \text{ and } L \text{ is through } o; \\ \\ 0 & \text{otherwise.} \end{cases}$$

Define $F_{P,o,a}(L) = \sum_{p\in P} f_{p,o,a}(L)$, which is the number of points of P with $\leq (a,a)$ distance to L for the line L through o. The expectation $\text{E}(F_{P,o,a})$ is the expected number of points in P with distance $\leq (a,a)$ to the random line L through o.

Lemma 4. *Let $a > 0$ be a constant and $\delta > 0$ be a small constant. Let P be a set of n grid points on the plane and o be a point on the plane. Then $\text{E}(F_{P,o,a}) \leq \frac{(4\pi+8)(1+\delta)a\sqrt{n}}{\pi\sqrt{4+2\pi}}$.*

Proof: Without loss generality, we assume that $o = (0,0)$ (Notice that $F_{P,o,a}$ is invariant under translation). Let $\epsilon > 0$ be a small constant that will be fixed later. Let us consider a grid point $p = (x,y) \in P$ on the plane and let $p_1 = (x, y-a)$ and $p_2 = (x, y+a)$. The angle

between the two lines op_1 and op_2 will be estimated (Figure 2.1). Let $d = \text{dist}(o, p)$, $d_1 = \text{dist}(o, p_1)$ and $d_2 = \text{dist}(o, p_2)$. Define the angles $\theta_1 = \angle pop_1, \theta_2 = \angle pop_2$ and $\alpha = \angle op_2p$.

From $\frac{a}{\sin\theta_2} = \frac{d}{\sin\alpha}$, we have $\sin\theta_2 = \frac{a}{d} \cdot \sin\alpha = \frac{a}{d} \cdot \frac{|x|}{d_2} = \frac{a|x|}{dd_2}$. Similarly, $\sin\theta_1 = \frac{a|x|}{dd_1}$. If $d > a$, then

$$\frac{a|x|}{d(d+a)} \leq \sin\theta_1, \sin\theta_2 \leq \frac{a|x|}{d(d-a)}. \tag{2.1}$$

Let $\beta_1 = \angle poq_1$ ($\beta_2 = \angle poq_2$) be the angle between the line segments op and oq_1 (oq_2 respectively), where $q_1 = (x - a, y)$ and $q_2 = (x + a, y)$. If $d > a$, then we also have

$$\frac{a|y|}{d(d+a)} \leq \sin\beta_1, \sin\beta_2 \leq \frac{a|y|}{d(d-a)}. \tag{2.2}$$

There is a constant d_0 such that if $d > d_0$, then we have the following inequalities:

$$\begin{aligned}
\frac{a|y|}{d^2}(1-\epsilon) &\leq \beta_1, \beta_2 \leq \frac{a|y|}{d^2}(1+\epsilon), \text{ and} \\
\frac{a|x|}{d^2}(1-\epsilon) &\leq \theta_1, \theta_2 \leq \frac{a|x|}{d^2}(1+\epsilon), \text{ and} \\
\frac{(1-\epsilon)a\max(|x'|,|y'|)}{d'^2} &< \frac{a\max(|x|,|y|)}{d^2} < \frac{(1+\epsilon)a\max(|x'|,|y'|)}{d'^2} \text{ for any } (x', y') \text{ with} \\
& \quad \text{dist}((x,y),(x',y')) \leq \sqrt{2}, \text{ where } d' = \text{dist}((x',y'), o).
\end{aligned}$$

Let $Pr(o, p, a)$ be the probability that the point p has distance $\leq (a, a)$ to a random line L through o. If $d \leq d_0$, then $Pr(o, p, a) \leq 1$. Otherwise, $Pr(o, p, a) \leq \frac{\max(2\max(\beta_1,\beta_2),2\max(\theta_1,\theta_2))}{\pi} \leq \frac{2}{\pi}\max\left(\frac{a|y|}{d^2}, \frac{a|x|}{d^2}\right)(1+\epsilon)$. The number of grid points with distance $\leq d_0$ to o is $\leq \pi(d_0 + \sqrt{2})^2$.

$$\begin{aligned}
\text{E}(F_{P,o,a}) &= \sum_{p\in P} E(f_{o,p,a}) = \sum_{p\in P} Pr(o,p,a) \\
&\leq \sum_{p\in P \text{ and } \text{dist}(p,o)>d_0} Pr(o,p,a) + \sum_{p\in P \text{ and } \text{dist}(p,o)\leq d_0} Pr(o,p,a) \\
&\leq \frac{2(1+\epsilon)}{\pi} \sum_{p\in P \text{ and } \text{dist}(p,o)>d_0} \max\left(\frac{|x|}{d^2}, \frac{|y|}{d^2}\right) + \pi(d_0+\sqrt{2})^2
\end{aligned} \tag{2.3}$$

We only consider the case to make $\sum_{p\in P \text{ and } d>d_0} \max(\frac{|x|}{d^2}, \frac{|y|}{d^2})$ maximal. By Lemma 3, it is maximal when the points of P are in the area $D(R)$ with the smallest R.

For a grid point $p = (i, j)$, define $\text{grid}_1(p) = \{(x, y) | i - \frac{1}{2} < x < i + \frac{1}{2} \text{ and } j - \frac{1}{2} < y < j +$

$\frac{1}{2}\}$, and $\text{grid}_2(p) = \{(x,y)|i-\frac{1}{2} \le x \le i+\frac{1}{2}$ and $j-\frac{1}{2} \le y \le j+\frac{1}{2}\}$. If the grid point $p \notin D(R)$, then $\text{grid}_1(p) \cap D(R-\frac{\sqrt{2}}{2}) = \emptyset$. The area size of $D(R)$ is $2\pi R^2 + 4R^2$. Assume R is the minimal radius such that $D(R)$ contains at least n grid points. The region $D(R-\epsilon)$ contains $< n$ grid points for every $\epsilon > 0$. This implies $D(R-\epsilon-\frac{\sqrt{2}}{2}) \subseteq \cup_{\text{grid point } p \in D(R-\epsilon)} \text{grid}_2(p)$. Therefore, $2\pi(R-\frac{\sqrt{2}}{2}-\epsilon)^2 + 4(R-\frac{\sqrt{2}}{2}-\epsilon)^2 \le n$. Hence, $R \le \frac{\sqrt{n}}{\sqrt{4+2\pi}} + \frac{\sqrt{2}}{2} + \epsilon < \frac{\sqrt{n}}{\sqrt{4+2\pi}} + \sqrt{2}$ (the constant ϵ will be $\le \frac{\sqrt{2}}{2}$).

Let $A_1 = \{p = (x,y) \in D(R)|$the angle between op and x-axis is in $[0, \frac{\pi}{4}]\}$, which is the $\frac{1}{8}$ area of $D(R)$. The probability that a point $p(=(x,y))$ has distance $\le (a,a)$ to the random line L is $\le \frac{2(1+\epsilon)ax}{d^2}$ for p in A_1 with $\text{dist}(p,o) > d_0$. The expectation of the number of points (with distance $\le (a,a)$ to L and distance $> d_0$ to o) of P in the area A_1 is

$$\begin{aligned}
&\sum_{p \in A_1 \cap P \text{ and } \text{dist}(p,o) > d_0} Pr(o,p,a) \le \sum_{p \in A_1 \cap P \text{ and } \text{dist}(p,o) > d_0} \frac{2(1+\epsilon)ax}{\pi d^2} \\
\le & \int\int_{A_1} \frac{2(1+\epsilon)^2 ax}{\pi d^2} d_x d_y = \frac{2(1+\epsilon)^2 a}{\pi} \int_0^{\frac{\pi}{4}} \int_0^{2R\cos\theta} \frac{r\cos\theta}{r^2} \cdot r d_r d_\theta \\
= & \frac{2(1+\epsilon)^2 a}{\pi} \int_0^{\frac{\pi}{4}} \int_0^{2R\cos\theta} \cos\theta d_r d_\theta = \frac{2(1+\epsilon)^2 aR}{\pi} \int_0^{\frac{\pi}{4}} 2(\cos\theta)^2 d_\theta \\
= & \frac{2(1+\epsilon)^2 aR}{\pi} \cdot (\frac{\pi}{4} + \frac{1}{2}) = \frac{(1+\epsilon)^2 aR}{\pi} \cdot (\frac{\pi}{2} + 1) \qquad (2.4)
\end{aligned}$$

Since $R \le \frac{\sqrt{n}}{\sqrt{4+2\pi}} + \sqrt{2}$, the total expectation is

$$\begin{aligned}
\text{E}(F_{P,o,a}) &\le 8 \sum_{p \in A_1 \cap P \text{ and } \text{dist}(p,o) > d_0} Pr(o,p,a) + \pi(d_0 + \sqrt{2})^2 \\
&\le \frac{8(1+\epsilon)^2 aR}{\pi} \cdot (\frac{\pi}{2} + 1) + \pi(d_0 + \sqrt{2})^2 \\
&\le \frac{(4\pi+8)(1+3\epsilon)a\sqrt{n}}{\pi\sqrt{4+2\pi}} \le \frac{(4\pi+8)(1+\delta)a\sqrt{n}}{\pi\sqrt{4+2\pi}}
\end{aligned}$$

for all large n. We assign to the constant ϵ the value $\min(\frac{\delta}{3}, \frac{\sqrt{2}}{2})$. ∎

Theorem 5. *Let $a > 0$ be a constant and P be a set of n grid points on the plane. Let $\delta > 0$ be a small constant. There is a line L such that the number of points in P with $\le (a,a)$ distance to L is $\le \frac{(4\pi+8)(1+\delta)a\sqrt{n}}{\pi\sqrt{4+2\pi}}$, and each half plane has $\le \frac{2n}{3}$ points from P for all large n.*

Proof: Let o be the center point of set P (by Lemma 2). The theorem follows from Lemma 4. ∎

The following corollary shows that for each grid graph of n nodes, its $\frac{2}{3}$-separator size is bounded by $1.02074\sqrt{n}$. For two grid points of distance 1, if they stay on different sides of separator line L, one of them has $\leq (\frac{1}{2}, \frac{1}{2})$ distance to L.

Corollary 6. *Let P be a set of n grid points on the plane. There is a line L such that the number of points in P with $\leq (1/2, 1/2)$ distance to L is $\leq 1.02074\sqrt{n}$, and each half plane has $\leq \frac{2n}{3}$ points from P.*

Proof: By Theorem 5 with $a = \frac{1}{2}$, we have, $\frac{8(1+\epsilon)}{\pi}\frac{1}{2} \cdot (\frac{\pi}{2} + 1) \cdot \frac{1}{\sqrt{4+2\pi}} < 1.02074$ when ϵ is small enough. ∎

Theorem 7. *Let $a > 0$ be a constant, P be a set of n grid points on the plane and, o be a center point of P. Let $\delta, \epsilon > 0$ be small constants. For a random line L through the center point o, it has probability at least $\frac{\epsilon}{1+\epsilon}$ such that the number of points in P with $\leq (a, a)$ distance to L is $\leq \frac{(4\pi+8)(1+\delta)(1+\epsilon)a\sqrt{n}}{\pi\sqrt{4+2\pi}}$, and each half plane has $\leq \frac{2n}{3}$ points from P for all large n.*

Proof: Let o be the center point of set P (by Lemma 2). By Lemma 4, $\mathrm{E}(F_{P,o,a}) \leq \frac{(4\pi+8)(1+\delta)a\sqrt{n}}{\pi\sqrt{4+2\pi}}$. By Markov's inequality, Probability$(F_{P,o,a}(L) > (1+\epsilon)\mathrm{E}(F_{P,o,a})) \leq \frac{1}{1+\epsilon}$. So, Probability$(F_{P,o,a}(L) \leq (1+\epsilon)\mathrm{E}(F_{P,o,a})) \geq 1 - \frac{1}{1+\epsilon} = \frac{\epsilon}{1+\epsilon}$. ∎

2.3. Separator lower bound for grid graphs

In this section we prove the existence of a lower bound of $0.7555\sqrt{n}$ for the grid graph separator. We delay the calculation to the next section for the length of the shortest curve partitioning the unit circle into two areas with ratio 1 : 2. A simple closed curve in the plane does not cross itself. Jordan's theorem states that every simple closed curve divides the plane into two compartments, one inside the curve and one outside of it, and that it is impossible to pass continuously from one to the other without crossing the curve.

Definition 8. A graph is *connected* if there is a path between every two nodes in the graph. For a connected grid graph $G = (V, E_V)$, a *contour* of G is a circular path $C = v_1v_2 \cdots v_kv_1$ such that 1) $(v_i, v_{i+1}) \in E_V (i = 1, 2, \cdots, k-1)$ and $(v_k, v_1) \in E_V$; 2) all points of V are in the one side of C; and 3) for any $i \leq j$, $v_1 \cdots v_{i-1}v_{j+1} \cdots v_kv_1$ does not satisfy both 1) and 2). A point $v \in V$ is a *boundary point* if $d(v, u) = 1$ for some grid point $u \notin V$. A contour C *separates* w from all grid points V if every path from w to a node in V intersects C.

Example: Let V be the set of all dotted grid points in Figure 2.2. $C = v_1v_2v_3v_4v_5v_6v_7v_8v_9v_{10}$ $v_{11}v_{12}v_{13}v_{14}v_1$ is a contour for V. The condition 3) prevents $C' = v_1v_2v_{15}v_2v_3v_4v_5v_6v_7$ $v_8v_9v_{10}v_{11}v_{12}v_{13}v_{14}v_1$ from being a contour.

Lemma 9. *Let $G = (V, E_V)$ be a connected grid graph. If the grid point $v \in V$ and grid point $w \notin V$ have the distance* $\text{dist}(v, w) = 1$*, then there is a contour C such that C contains v and separates w from all grid points of V.*

Proof: Imagine that a region starting from the grid point w grows until it touches all of the reachable edges of G (but never crosses any of them). Since G is a connected grid graph, the boundary forms a contour that consists of edges of G. As $\text{dist}(w, v) = 1$, the vertex v should appear in the contour. ∎

Lemma 10. *Let $G = (V, E_V)$ be a grid graph and C be a contour of G. Let*

$U = \{u | u$ *is a grid point not in V with* $\text{dist}(u, v) = 1$ *for some $v \in V$ and C separates u from $V\}$.*

Then there is a list of grid points $u_1, u_2, \cdots, u_{m+1}$ in U such that $u_{m+1} = u_1$, $dist(u_i, u_{i+1}) \leq \sqrt{2}$ for $i = 1, 2, \cdots, m$ and all points of P are on one side of the circle path $u_1u_2 \cdots u_{m+1}$ (the edge connecting every two consecutive points u_1, u_2 is straight line).

Proof: Walking along the contour $C = v_1 \cdots v_kv_1$, we assume that only the left side has the points from V. A point v_i on C is called *special point* if $v_{i-1} = v_{i+1}$. The point v_9 is a special point at the contour $v_1v_2 \cdots v_{14}v_1$ in Figure 2.2. For each edge (v_i, v_{i+1}) in C, the grid square, which is on the right side of (v_i, v_{i+1}) and contains (v_i, v_{i+1}) as one of the four boundary edges, has at least one point not in V. Let $S_1, S_2, \cdots, S_k$ be those grid squares for $(v_1, v_2), (v_2, v_3), \cdots, (v_k, v_1)$, respectively. For each special point v_i on C, it has two special grid squares S_i' and S_i'' that share the edge (v_i, u) for some $u \in U$ with $\text{dist}(u, v_i) = 1$ and $\text{dist}(u, v_{i-1}) = 2$ (for example, S_9' and S_9'' on Figure 2.2). Insert S_i' and S_i'' between S_i and S_{i+1}. We get a new list of grid squares $H_1, H_2, \cdots, H_m$. We claim that for every two consecutive H_i and H_{i+1}, there are grid points $u_i \in H_i \cap U$ and $u_{i+1} \in H_{i+1} \cap U$ with $\text{dist}(u_i, u_{i+1}) \leq \sqrt{2}$. The lemma is verified by checking the following cases:

Case 1. $H_i = S_j$ and $H_{i+1} = S_{j+1}$ for some $j \leq k$.

Subcase 1.1. S_j and S_{j+1} share one edge $v_{j+1}u$. An example of this subcase is the grid squares S_1 and S_2 on Figure 2.2. It is easy to see that $u \in U$ since u is on the right side when walking along the cycle path C.

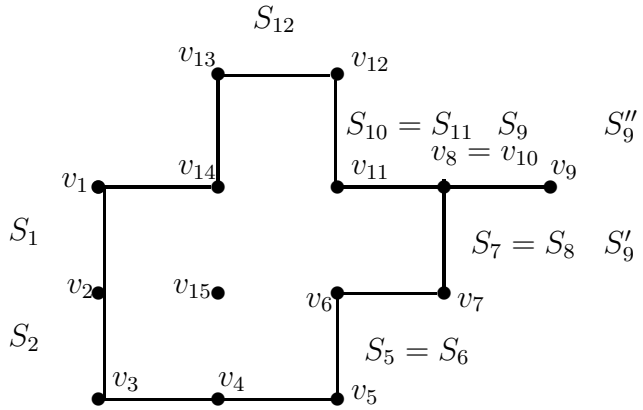

Figure 2.2: Contour $C = v_1v_2 \cdots v_{14}v_1$. The node v_9 is a special point. When walking along $v_1 \cdots v_{14}v_1$, we see that each S_i is the grid square on the right of v_iv_{i+1}

Subcase 1.2. $S_j = S_{j+1}$. An example of this subcase is the grid squares S_5 and S_6 on Figure 2.2. This is a trivial case.

Subcase 1.3. S_j and S_{j+1} only share the point v_{j+1}. An example of this subcase is the grid squares S_{11} and S_{12} on Figure 2.2. We have grid points $u_1 \in U$ and $u_2 \in U$ such that $\text{dist}(u_1, v_{j+1}) = 1$, $\text{dist}(u_2, v_{j+1}) = 1$. Furthermore, $\text{dist}(u_1, u_2) = \sqrt{2}$.

Case 2. $H_i = S''_j$ and $H_{i+1} = S_j$ for some $j < m$. An example of this subcase is the grid squares S''_9 and S_9 on Figure 2.2. The two squares share the edge v_ju for some $u \in U$.

Case 3. $H_i = S'_j$ and $H_{i+1} = S''_j$. An example of this subcase is the grid squares S'_9 and S''_9 on Figure 2.2. The two squares share the edge u_ju for some $u \in U$.

Case 4. $H_i = S_{j-1}$ and $H_i = S'_j$. An example of this subcase is the grid squares S_8 and S'_9 on Figure 2.2. The two squares share v_ju for some $u \in U$. ∎

Definition 11. For a region R on the plane, define $A(R)$ to be the area size of R. An unit circle has radius 1. For a region R in the unit circle, $L(R)$ is the length of the boundary of R inside the internal area of the unit circle. A region R inside a unit circle is *type 1* region if part of its boundary is from the unit circle boundary. Otherwise, it is called *type 2* region, which does not share any boundary with the unit circle.

Lemma 12. *Assume $s > 0$ is a constant and p_1, p_2 are two points on the plane. We have 1) the area with the shortest boundary and area size s on the plane is a circle with radius $\sqrt{\frac{s}{\pi}}$; and 2) the shortest curve that is through both p_1 and p_2, and forms an area of size s with the line segment p_1p_2 is a circle arc.*

The proof of Lemma 12 can be found in regular variational calculus textbooks (e.g. [56]). Let R be a type 1 region of area size s. Let C be the part of R boundary that is an unit

circle arc with p_1 and p_2 as two end points. Let C' be the rest of the boundary of R. Let R' be the region with the boundary C and line segment p_1p_2. Assume the length of C' is minimal. If $A(R) = A(R')$, then C' is the same as the line segment p_1p_2. If $A(R) < A(R')$, then C' is a circle arc inside R' (between C and p_1p_2). If $A(R) > A(R')$, then C' is also a circle arc outside R'. Those facts above follow from Lemma 12.

Lemma 13. *Let $s \leq \pi$ be a constant. Let $R_1, R_2, \cdots, R_k$ be k regions inside an unit circle (they may have overlaps), $\sum_{i=1}^{k} A(R_i) = s$ and $\sum_{i=1}^{k} L(R_i)$ is minimal. Then $k = 1$ and R_1 is a type 1 region.*

Proof: We consider the regions $R_1, \cdots, R_k$ that satisfy $\sum_{i=1}^{k} A(R_i) = s$ and $\sum_{i=1}^{k} L(R_i)$ is minimal for $k \geq 1$. Each $R_i (i = 1, \cdots, k)$ is either type 1 or type 2 region. The part of boundary of R_i that is also the boundary of the unit circle is called *old boundary*. Otherwise it is called type *new boundary*.

A type 2 region has to be a circle (by Lemma 12). For a type 1 region, its new boundary inside the unit circle is also a circle arc (otherwise, its length is not minimal by part 2 of Lemma 12). If we have both type 1 region R_1 and type 2 region R_2. Move R_1 to R_1^* and R_2 to R_2^* on the plane so that R_1^* and R_2^* have some intersection (not a circle) at their new boundaries. Let R_2' be the circle with the same area size as $R_1^* \cap R_2^*$. The boundary length of R_2' is less than that of $R_1^* \cap R_2^*$. So, $L(R_1) + L(R_2)$ reduces to $L(R_1^* \cup R_2^*) + L(R_2')$ if R_1 and R_2 are replaced by $R_1^* \cup R_2^*$ and R_2' (Notice that $A(R_1) + A(R_2) = A(R_1^* \cup R_2^*) + A(R_1^* \cap R_2^*) = A(R_1^* \cup R_2^*) + A(R_2')$). This contradicts that $\sum_{i=1}^{k} L(R_i)$ is minimal. Therefore, there is no type 2 region. We only have type 1 regions left. Assume that R_1 and R_2 are two type 1 regions. Let R_1 and R_2 have the unit circle arcs p_1p_2 and p_2p_3 respectively. They can merge into another type 1 region R with the unit circle arc $p_1p_2p_3$ and the same area size $A(R) = A(R_1) + A(R_2)$. Furthermore, $L(R) < L(R_1) + L(R_2)$. A contradiction again. Therefore, $k = 1$ and R_1 is a type 1 region. ∎

Definition 14. Let q be a positive real number. Partition the plane into $q \times q$ squares by the horizontal lines $y = iq$ and vertical lines $x = jq$ $(i, j \in Z)$. Each point (iq, jq) is a (q, q)-*grid point*, where $i, j \in Z$.

Lemma 15. *Let V be the set of all (q, q) grid points in the unit circle C. Let $G = (V, E_V)$ be the grid graph on V, where $E_V = \{(v_i, v_j) | \text{dist}(v_i, v_j) = 1$ and $v_i, v_j \in V\}$. Assume that l is a curve that partitions a unit circle C into P_1 and P_2 with $\frac{A(P_1)}{A(P_2)} = \frac{1}{t}$. If the minimal length of l is c_0, then every $\frac{t}{t+1}$-separator for the grid graph G has a size $\geq \frac{c_0(\sqrt{n} - \sqrt{2\pi})}{\sqrt{2}\sqrt{\pi}}$.*

Proof: Assume that the unit circle C area has n (q,q)-grid points. We have $\pi(1+q\sqrt{2})^2 \geq n \cdot q^2$. It implies $q \leq \frac{1}{\frac{\sqrt{n}}{\sqrt{\pi}}-\sqrt{2}}$. Assume $A \subseteq V$ is the smallest separator for $G=(V,E_V)$ such that $G-A$ has two disconnected subgraphs $G_1=(V_1,E_{V_1})$ and $G_2=(V_2,E_{V_2})$, which satisfy $|V_1|,|V_2| \leq \frac{tn}{t+1}$. By Corollary 6, $|A| \leq 2\sqrt{n}$. Let G_1 have connected components $F_1,\cdots,F_m$. By Lemma 10, each F_i is surrounded by a circular path H_i with grid points not from G_1. Actually, the grid points of H_i inside C are from the separator A. Let $P_1,\cdots,P_k$ be the parts of $H_1,\cdots,H_m$ inside the C. They consist of vertices in A and the distance between every two consecutive vertices in each P_i is $\leq \sqrt{2}q$ (by Lemma 10 and scaling (q,q) grid points to $(1,1)$ grid points).

The number of (q,q)-grid points with distance ≤ 2 to the unit circle boundary is also $O(\sqrt{n})$. For a (q,q)-grid point $p=(iq,jq)$, define $\text{grid}_q(p)=\{(x,y)|iq-\frac{q}{2} \leq x \leq iq+\frac{q}{2}$ and $jq-\frac{q}{2} \leq y \leq jq+\frac{q}{2}\}$. Let V_H be the set of all (q,q)-grid points in $H_1,\cdots,H_m$ and V_P be the set of all (q,q)-grid points in $P_1,\cdots,P_k$. Let $S_1=\cup_{p\in V_1}\text{grid}_q(p)$, $S_1'=\cup_{p\in V_1\cup V_H}\text{grid}_q(p)$, and $S_1''=\cup_{p\in V_1\cup V_P}\text{grid}_q(p)$. It is easy to see that $\frac{2n}{3}q^2 \geq A(S_1) \geq \frac{n}{3}q^2$ and $A(S_1')=A(S_1)+O(\sqrt{n})$ and $A(S_1'')=A(S_1)+O(\sqrt{n})$. Therefore, the sizes of S_1' and S_1'' are almost the same as that of S_1 (because $\sqrt{n} << n$). For the variable $x \geq 1$, define the function $g(x)$ to be the length of the shortest curve that partitions the unit circle into regions P_1 and P_2 with $\frac{A(P_1)}{A(P_2)}=\frac{1}{x}$. Then $g(x)$ is a decreasing continuous function (see the analysis in section 2.4).

The total length of $P_1,\cdots,P_k$ is minimal when $k=1$ by Lemma 13. Since the length of P_1 is $\geq c_0$, there are at least $\frac{c_0}{q\sqrt{2}} \geq \frac{c_0(\frac{\sqrt{n}}{\sqrt{\pi}}-\sqrt{2})}{\sqrt{2}} = \frac{c_0(\sqrt{n}-\sqrt{2\pi})}{\sqrt{2}\sqrt{\pi}}$ grid points of A along P_1. ∎

Theorem 16. *There exists a grid graph $G=(V,E_V)$ such that for any $A \subseteq V$ if $G-A$ has two disconnected graphs G_1 and G_2, and $G_i(i=1,2)$ has $\leq \frac{2|V|}{3}$ nodes, then $|A| \geq 0.7555\sqrt{n}$ when n is large.*

Proof: By Theorem 17 in the next section, the length of the shortest curve partitioning the unit circle into $1:2$ ratio is ≥ 1.8937. By Lemma 15 with $c_0=1.8937$ and $k=1$, we have $|A| \geq 0.7555\sqrt{n}$. ∎

2.4. Shortest separator of the unit circle

Let $y=f(x)$ be the function that minimizes the length of the curve connecting two arbitrary points (x_1,y_1) and (x_2,y_2) on the circle shown in Figure 2.3A, with the additional constraint

that the ratio $A1/A2$ of the two pieces is a constant k. The length of the curve connecting the two fixed points is the functional expression

$$L(x, f(x), f\prime(x)) = \int_{x_1}^{x2} \sqrt{1 + (f\prime(x))^2} dx, \tag{2.5}$$

where prime denotes the derivative with respect to x. The constant ratio of the two areas $A_1/A_2 = k$, together with $A_1 + A_2 = \pi R^2$, gives $A_1 = \pi R^2 \frac{k}{k+1}$. To determine the extremum of the functional (2.5) with the above constraint on A_1, we used the Lagrange multipliers method (see [56]). The functional whose extremum we are searching for is

$$\begin{aligned} L^*(x, f(x), f\prime(x)) &= L(x, f(x), f\prime(x)) + \lambda \left(A_1 - \pi R^2 \frac{k}{k+1} \right) = \\ &= \int_{x_1}^{x2} \sqrt{1 + (f\prime(x))^2} dx + \lambda \left(\int_{x_1}^{x2} \left(\sqrt{R^2 - x^2} - f(x) \right) dx - \pi R^2 \frac{k}{k+1} \right), \end{aligned} \tag{2.6}$$

where λ is the Lagrange multiplier and $A_1 = \int_{x_1}^{x2} \left(\sqrt{R^2 - x^2} - f(x) \right) dx$ (Figure 2.3A).

The functional that determines the extremum of (2.6) is $F(x, f(x), f\prime(x)) = \sqrt{1 + (f\prime(x))^2} + \lambda \left(\sqrt{R^2 - x^2} - f(x) \right)$. The Euler-Lagrange equation of the functional $F(x, f(x), f\prime(x))$ is $\frac{\partial F}{\partial f} - \frac{d}{dx} \frac{\partial F}{\partial f\prime} = 0$. The solution of the Euler-Lagrange equation is the minimum length separator function $y = f(x)$ with

$$f(x) = b - \sqrt{\left(\frac{1}{\lambda} \right)^2 - \left(x + \frac{a}{\lambda} \right)^2}, \tag{2.7}$$

where b is an arbitrary constant. The solution (2.7) of the variational problem (2.6) represents a circle of radius $r = \frac{1}{\lambda}$ and center at $(-\frac{a}{\lambda}, b)$ (Figure 2.3B, C).

The area of the circular region subtended by the angle θ is $R^2(\theta - \sin(\theta))/2$, and by the angle ϕ is $r^2(\phi - \sin(\phi))/2$ (Figure 2.3B). Therefore, the total area A_1 is given be the sum of the above areas

$$A_1 = \frac{R^2}{2}(\theta - \sin(\theta)) + \frac{r^2}{2}(\phi - \sin(\phi)) = \pi R^2 \frac{k}{k+1}, \tag{2.8}$$

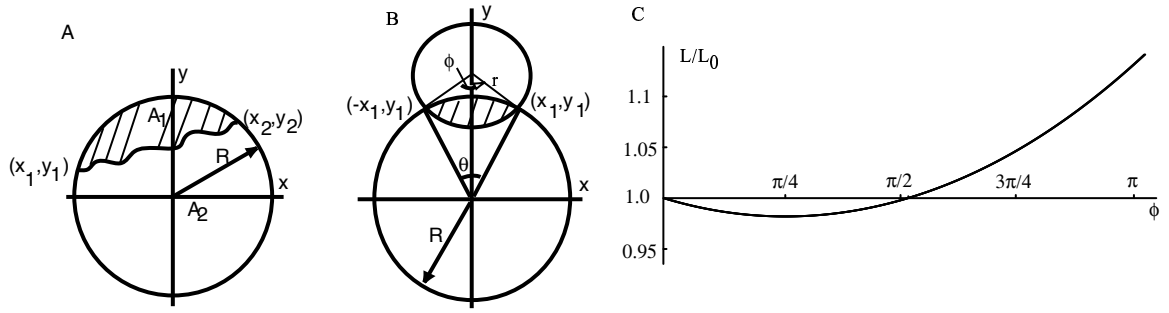

Figure 2.3: The minimum length path that divides a circle into two regions with a fixed ratio k. A. The minimum length curve connecting the points (x_1, y_1) and (x_2, y_2) is the solution of the variational problem (2.6). B. The solution of the variational problem is a circle of radius r with a subtending angle $\phi < \pi$. C. Implicit plot of the normalized arc length L/L_0 versus the subtended angle ϕ. For $k = 1/2$ the arc has a minimum length $L_{min} \approx 0.982002 L_0 = 1.8937$ for an angle $\phi_{min} \approx 0.79388$.

with the additional obvious relationship (Figure 2.3B)

$$R \sin(\theta/2) = r \sin(\phi/2). \tag{2.9}$$

The length of the separator arc that connects the two points $(-x_1, y_1)$ and (x_1, y_1) on the circle of radius r is $L = r\phi$ (Figure 2.3B). By substituting the explicit expression of θ from (2.9) into (2.8), we get an implicit relationship between the variables r and ϕ

$$2 \arcsin\left(\frac{r}{R} \sin\frac{\phi}{2}\right) - \sin\left(2 \arcsin\left(\frac{r}{R} \sin\frac{\phi}{2}\right)\right) + \frac{r^2}{R^2}(\phi - \sin(\phi)) = 2\pi \frac{k}{k+1}. \tag{2.10}$$

Using the definition of the arc length we get $r = L/\phi$, which substituted into (2.10) leads to an implicit relationship between the arc length L and the subtending angle ϕ

$$2 \arcsin\left(\frac{L}{R\phi} \sin\frac{\phi}{2}\right) - \sin\left(2 \arcsin\left(\frac{L}{R\phi} \sin\frac{\phi}{2}\right)\right) + \frac{L^2}{R^2\phi^2}(\phi - \sin(\phi)) = 2\pi \frac{k}{k+1}. \tag{2.11}$$

We numerically solved the implicit equation (2.11) for different values of $\phi \in (0, \pi)$ (Figure 2.3C). The arc length L was normalized by the arc length L_0 of the straight line that cuts the circle of radius R with the same ratio $k = A_1/A_2$ (Figure 2.3A). Based on Figure 2.3B, the length L_0 of the straight line that cuts the circle in two regions with the given ratio k is $L_0 = 2R \sin\left(\frac{\theta_0}{2}\right)$. The angle θ_0 is the solution of the constraint equation (2.8) in the limit case of $r \to \infty$ and $\phi \to 0$, which leads to $\theta_0 - \sin\theta_0 = 2\pi \frac{k}{k+1}$. For a circle of unit radius

$(R = 1)$ and $k = 1/2$ the numeric solution is $\theta_0 \approx 2.60533$ radians, and the corresponding length of the straight line separator is $L_0 \approx 1.92853$.

We numerically found that the arc separator measured along the circle of radius r is always shorter than the corresponding straight line separator ($L/L_0 \leq 1$) if the subtending angle $\phi \in (0, \pi/2)$ (Figure 2.3C). If the subtending angle $\phi < \pi/2$ (Figure 2.3B, C), then there is a value, ϕ_{min}, such that the arc length is the minimum possible and this is the optimal solution for the separator length. We numerically found that $\phi_{min} \approx 0.79388 \in (0, \pi/2)$ and the corresponding radius of the circle is $r \approx 1.23672R$. If the subtending angle $\phi > \pi/2$, according to the numerical solution of the implicit equation (2.11) shown in (Figure 2.3C), the arc is no longer the minimum length solution of the variational problem. We formulate our analysis to the theorem below:

Theorem 17. *The shortest curve that partitions a unit circle into two regions with ratio* $1 : 2$ *has length* > 1.8937.

Chapter 3

Application of multi-directional width-bounded geometric separators to protein folding in the HP model

We have shown that there is a size $O(\sqrt{n})$ separator line to partition the folding problem of n letters into 2 problems in a balanced way. 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 $O(\sqrt{n})$ letters, there are at most $n^{O(\sqrt{n})}$ cases to partition the problem. The major improvement from the algorithm in [27] is the approximation of the optimal separator line. We need the following terms:

Definition 18.

- For integers i and j, *integer interval* $[i, j] = \{i, i+1, \cdots, j\}$. 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 i-th letter of S. $S[i, j]$ denotes the subsequence $S[i]S[i+1]\cdots S[j]$. If $[i_1, j_1], [i_2, j_2], \cdots, [i_t, j_t]$ are 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 $S[A] = S[i_1]S[i_2]\cdots S[i_k]$.
- For a 2-dimensional point (x_1, x_2), define $||(x_1, x_2)|| = |x_1| + |x_2|$.
- A *self-avoiding arrangement* f for a sequence S of length n on the 2-dimensional grid is a one-to-one mapping from $\{1, 2, \cdots, n\}$ to Z^2 such that $||f(i) - f(i+1)|| = 1$ for $i = 1, 2, \cdots, n-1$. 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^2 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^2.

- For a grid self-avoiding arrangement, its *contact map* is the graph $G_f = (1, 2, \cdots, n, E)$, where the edge set $E = \{(i, j) : |i - j| > 1 \text{ and } ||f(i) - f(j)|| = 1\}$.
- For a line L with equation $f(x, y) = 0$, define $L_{<0}$ and $L_{>0}$ as the area $\{(x, y)|f(x, y) < 0\}$ and $\{(x, y)|f(x, y) > 0\}$ respectively.

Assume that our input HP sequence has n_0 letters and the optimal folding is inside an $m \times m$ square. We will 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 on the same line of the boundary have distance ϵ'. 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 19. *Let $m > 2$. Let $1 > \epsilon > 0$ and $\delta > 0$ be two small constants. Let c_1 be a constant $> \frac{(1+3\epsilon)^2}{\epsilon}$. Let L be a line, which intersects the $m \times m$ square A and has the slope s. The four boundary lines segments of A are either vertical or horizontal. Each side of L has $\geq c_1$ grid points in A and $\epsilon \leq |s| \leq \frac{1}{\epsilon}$. Then for some constant $c_2 > 0$ and every $0 < \epsilon' \leq \frac{1}{c_2 \cdot m}$, there exists an ϵ'-regular line L' such that for every grid point $q \in A$ with $\leq (a, a)$ to L' has $\leq (a + \delta, a + \delta)$ distance to L.*

Proof: Let V_1 and V_2 be the two line segments of vertical boundary of A. Let H_1 and H_2 be the two line segments of the horizontal boundary of A. Let $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$ be the two intersections of L with the boundary of A. Let $p'_1 = (x'_1, y'_1)$ and $p'_2 = (x_2, y_2)'$ be the two closest ϵ'-regular points to p_1 and p_2 respectively on the boundary of A, where ϵ' will be determined later. Let $q = (x_0, y_0)$ be a grid point in A. Let L' be the ϵ'-regular line through both (x'_1, y'_1) and (x'_2, y'_2).

Let L_v and L_h be the vertical and horizontal lines through q, respectively. The intersection between L and L_h is at the point (x, y_0), where $x = \frac{x_2 - x_1}{y_2 - y_1}(y_0 - y_1) + x_1$. The intersection between L and L_v is at the point (x_0, y), where $y = \frac{y_2 - y_1}{x_0 - x_1}(x_0 - x_1) + y_1$.

Similarly, the intersection between L' and L_h is at the point (x', y_0) where $x' = \frac{x'_2 - x'_1}{y'_2 - y'_1}(y_0 - y'_1) + x'_1$. The intersection between L' and L_v is at the point (x_0, y'), where $y' = \frac{y'_2 - y'_1}{x_0 - x'_1}(x_0 - x'_1) + y'_1$. Since s is the slope of line L that is through the points (x_1, y_1) and (x_2, y_2), we

have $s = \frac{y_2-y_1}{x_2-x_1}$. The line L', which is through (x_1', y_1') and (x_2', y_2'), has the slope $s' = \frac{y_2'-y_1'}{x_2'-x_1'}$. By the condition of this lemma, we have

$$\epsilon \leq |s| = \left|\frac{y_2 - y_1}{x_2 - x_1}\right| \leq \frac{1}{\epsilon}. \tag{3.1}$$

Case 1: both p_1 and p_2 are in $V_1 \cup V_2$. This implies that $|x_2 - x_1| = m$. Since $|s| = |\frac{y_2-y_1}{x_2-x_1}| \geq \epsilon$, we have $|y_2 - y_1| \geq \epsilon m$.

Case 2: both p_1 and p_2 are in $H_1 \cup H_2$. This implies that $|y_2 - y_1| = m$. Since $|s| = |\frac{y_2-y_1}{x_2-x_1}| \leq \frac{1}{\epsilon}$, we have $|x_2 - x_1| \geq \epsilon m$.

Case 3: p_1 and p_2 are in $V_i \cup H_j$ for some i, j. We have $(|x_2 - x_1| + 2)(|y_2 - y_1| + 2) \geq c_1$ since each side of L has at least c_1 grid points in A. Then $(|x_2 - x_1| + 2)(\frac{|x_2-x_1|}{\epsilon} + 2) > c_1$. This gives that $|x_2 - x_1| > \frac{\sqrt{4\epsilon c_1 + 4 + 4\epsilon} - (2+2\epsilon)}{2} > \sqrt{\epsilon c_1} - (1 + \epsilon)$. Since $\sqrt{\epsilon c_1} \geq \sqrt{\epsilon \frac{(1+3\epsilon)^2}{\epsilon}} = 1 + 3\epsilon$, $|x_1 - x_2| \geq \sqrt{\epsilon c_1} - (1 + \epsilon) \geq 1 + 3\epsilon - (1 + \epsilon) \geq 2\epsilon$. Similarly, $|y_2 - y_1| > 2\epsilon$. Combining the cases 1 to 3, we always have

$$|x_1 - x_2| \geq 2\epsilon \text{ and } |y_1 - y_2| \geq 2\epsilon. \tag{3.2}$$

Let $x_2' - x_1' = x_2 - x_1 + \epsilon_x$ and $y_2' - y_1' = y_2 - y_1 + \epsilon_y$. Since (x_1', y_1') is the closest ϵ'-regular point to (x_1, y_1) and (x_2', y_2') is the closest ϵ'-regular point to (x_2', y_2'), $|\epsilon_x| \leq 2\epsilon'$ and $|\epsilon_y| \leq 2\epsilon'$. Define

$$\epsilon_{0,s} = \frac{4\epsilon'}{\epsilon^2} \tag{3.3}$$

$$\begin{aligned}
|s - s'| &= \left|\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_2' - y_1'}{x_2' - x_1'}\right| = \left|\frac{(y_2 - y_1)(x_2' - x_1') - (y_2' - y_1')(x_2 - x_1)}{(x_2 - x_1)(x_2' - x_1')}\right| \\
&= \left|\frac{(y_2 - y_1)(x_2 - x_1 + \epsilon_x) - (y_2 - y_1 + \epsilon_y)(x_2 - x_1)}{(x_2 - x_1)(x_2 - x_1 + \epsilon_x)}\right| = \left|\frac{(y_2 - y_1)\epsilon_x - \epsilon_y(x_2 - x_1)}{(x_2 - x_1)(x_2 - x_1 + \epsilon_x)}\right| \\
&= \left|\frac{\epsilon_x(y_2 - y_1)}{(x_2 - x_1)(x_2 - x_1 + \epsilon_x)} - \frac{\epsilon_y}{(x_2 - x_1 + \epsilon_x)}\right| \leq \frac{|\epsilon_x||s|}{|x_2 - x_1 + \epsilon_x|} + \frac{|\epsilon_y|}{|x_2 - x_1 + \epsilon_x|} \\
&\leq \left|\frac{\epsilon_x}{\epsilon(x_2 - x_1 + \epsilon_x)}\right| + \left|\frac{\epsilon_y}{|x_1 - x_2| - |\epsilon_x|}\right| \leq \frac{2\epsilon'}{\epsilon^2} + \frac{2\epsilon'}{\epsilon} \leq \frac{4\epsilon'}{\epsilon^2} = \epsilon_{0,s}.
\end{aligned} \tag{3.4}$$

For (3.4) to (3.4), it is because $|x_1 - x_2| - |\epsilon_x| \geq 2\epsilon - \epsilon \geq \epsilon$ by (3.2). Let $x_1' = x_1 + \epsilon_{1,x}$, $y_1' = y_1 + \epsilon_{1,y}$ and $s' = s + \epsilon_s$. By (3.4) to (3.4), we have the inequality:

$$|\epsilon_s| \leq \epsilon_{0,s} \tag{3.5}$$

Since (x'_1, y'_1) is the closest ϵ'-regular point to (x_1, y_1) and (x'_2, y'_2) is the closest ϵ'-regular point to (x'_2, y'_2), $|\epsilon_{1,x}| \leq \epsilon'$ and $|\epsilon_{1,y}| \leq \epsilon'$. We consider the difference between y and y' as well as the difference between x and x'.

$$\begin{aligned}
|y - y'| &= |(s(x_0 - x_1) + y_1) - (s'(x_0 - x'_1) + y'_1)| \\
&= |(s(x_0 - x_1) + y_1) - ((s + \epsilon_s)(x_0 - x_1 - \epsilon_{1,x}) + y_1 + \epsilon_{1,y})| \\
&= |-(x_0 - x_1)\epsilon_s + \epsilon_{1,x}(s + \epsilon_s) - \epsilon_{1,y}| \\
&\leq |(x_0 - x_1)\epsilon_s| + |\epsilon_{1,x}(s + \epsilon_s)| + |\epsilon_{1,y}| \\
&\leq |\epsilon_s|m + \left|\epsilon_{1,x}(\frac{1}{\epsilon} + |\epsilon_s|)\right| + |\epsilon_{1,y}| \leq |\epsilon_s|m + \left|\frac{2\epsilon_{1,x}}{\epsilon}\right| + |\epsilon_{1,y}|. \qquad (3.6)
\end{aligned}$$

For (3.6) $\rightarrow$ (3.6), it is because the following facts: By (3.3) and (3.5), $|\epsilon_s| \leq |\epsilon_{0,s}| \leq \frac{4\epsilon'}{\epsilon^2} \leq \frac{1}{\epsilon^2} \leq \frac{1}{\epsilon}$ (the condition $4\epsilon' \leq 1$ will be satisfied when we set the constant ϵ' later).

$$\begin{aligned}
|x - x'| &= \left|\left(\frac{1}{s}(y_0 - y_1) + x_1\right) - \left(\frac{1}{s'}(y_0 - y'_1) + x'_1\right)\right| \\
&= \left|\frac{1}{s}(y_0 - y_1) - \frac{1}{s + \epsilon_s}(y_0 - y_1 - \epsilon_{1,y}) - \epsilon_{1,x}\right| \\
&= \left|(y_0 - y_1)(\frac{\epsilon_s}{s(s + \epsilon_s)}) + \frac{\epsilon_{1,y}}{s + \epsilon_s} - \epsilon_{1,x}\right| \leq \left|(y_0 - y_1)(\frac{\epsilon_s}{s(s + \epsilon_s)})\right| + \left|\frac{\epsilon_{1,y}}{s + \epsilon_s}\right| + |\epsilon_{1,x}| \\
&\leq \left|n(\frac{\epsilon_s}{s(s + \epsilon_s)})\right| + \left|\frac{\epsilon_{1,y}}{s + \epsilon_s}\right| + |\epsilon_{1,x}| \leq \left|n(\frac{2\epsilon_s}{\epsilon^2})\right| + \left|\frac{2\epsilon_{1,y}}{\epsilon}\right| + |\epsilon_{1,x}|\,. \qquad (3.7)
\end{aligned}$$

For (3.7)$\rightarrow$(3.7), it is because the following facts: By (3.5), $|\epsilon_s| \leq \epsilon_{0,s} = \frac{4\epsilon'}{\epsilon^2} \leq \frac{\epsilon}{2}$ (we will set up ϵ' so that $\epsilon' \leq \frac{\epsilon^3}{8}$). This implies that $|s + \epsilon_s| \geq |\epsilon - \frac{\epsilon}{2}| \geq \frac{\epsilon}{2}$. Therefore, $|s(s + \epsilon_s)| \geq \frac{\epsilon^2}{2}$.

We choose ϵ' so that it satisfies the following inequalities: (1)$|\epsilon_s|m \leq \frac{\delta}{3}$, (2)$\left|\frac{2\epsilon_{1,x}}{\epsilon}\right| \leq \frac{\delta}{3}$, (3)$|\epsilon_{1,y}| \leq \frac{\delta}{3}$, (4) $|m(\frac{\epsilon_s}{\epsilon^2})| \leq \frac{\delta}{3}$, (5)$|\frac{2\epsilon_{1,y}}{\epsilon}| \leq \frac{\delta}{3}$, (6)$|\epsilon_{1,x}| \leq \frac{\delta}{3}$, (7)$4\epsilon' \leq 1$, and (8) $\epsilon' \leq \frac{\epsilon^3}{8}$. Let $e' \leq \min(\frac{\delta\epsilon^2}{12m}, \frac{\delta\epsilon}{6}, \frac{\delta}{6}, \frac{\delta\epsilon^4}{12m}, \frac{\delta\epsilon}{6}, \frac{\delta}{6}, \frac{\epsilon}{2}, \frac{1}{4}, \frac{\epsilon^3}{8})$, in which each item is for the corresponding condition among (1)-(8). We let $\epsilon' = \frac{\delta\epsilon^4}{12m}$, which makes both $|x - x'| \leq \frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3} = \delta$ and $|y - y'| \leq \frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3} = \delta$. ∎

Lemma 20. *Let a and δ be positive constants. Let P be a set of n grid points in a 2-dimensional $m \times m$ square. There exist $\epsilon' = \frac{1}{c'_2 m}$ and ϵ'-regular line L' such that there are $\leq (\frac{2}{3} + \delta)n$ points of P on each half plane (divided by L'), and $\leq k_0 a(1 + \delta)\sqrt{n}$ points of P with distance $\leq (a, a)$ to L' for all large n, where $k_0 = \frac{(4\pi+8)}{\pi\sqrt{4+2\pi}}$ and c_2 is a constant > 0.*

Proof: Let $\delta_0, \delta_1, \delta_2 > 0$ be small constants with $(1+\delta) > (1+\frac{\delta_0}{a})(1+\delta_1)(1+\delta_2)$. Let o be the center of P via Lemma 2. By Theorem 7, $\mathrm{E}\,(F_{P,o,a+\delta_0}(L) \leq k_0(1+\delta_1)(1+\delta_2)(a+\delta_0)\sqrt{n}) \geq \frac{\delta_2}{1+\delta_2}$. There exists a line L that has angles $\geq \theta = \frac{1}{4}\frac{\delta_2}{1+\delta_2}$ with both x-axis and y-axis. The slope of L s satisfies $\epsilon \leq |s| \leq \frac{1}{\epsilon}$, where $\epsilon = \tan\frac{1}{4}\theta$. Each side has at most $\frac{2}{3}n$ points in P. By Lemma 19, we can select the constants $c_2 > 0$ and $\epsilon' = \frac{1}{c_2' m}$ to satisfy the conditions below: (a) $\epsilon' \leq a + \delta_0$, and (b) there exists ϵ'-regular line L' such that every grid point with distance $\leq (a, a)$ to L' has distance $\leq (a+\delta_0, a+\delta_0)$ to L. Thus, the number of points in P with distance $\leq (a, a)$ to L' is $\leq k_0(1+\delta_1)(1+\delta_2)(a+\delta_0)\sqrt{n} \leq k_0(1+\delta)a\sqrt{n}$. Let's consider the grid points between L and L'. Since the two end points of L' on the boundary of the $m \times m$ square have distance $\leq (\epsilon', \epsilon')$ to L, every point between L and L' in the $m \times m$ square has distance $\leq (\epsilon', \epsilon')$ to L. Since $\epsilon' \leq a + \delta_0$, the number of points in P with distance $\leq (\epsilon', \epsilon')$ to L is no more than the number of points of P with distance $\leq (a+\delta_0, a+\delta_0)$ to L. The number of those points is $O(\sqrt{n})$. Thus, each side of L' has $\leq (\frac{2}{3}+\delta)n$ points for all large n. Therefore, the number of points in P with distance $\leq (a, a)$ to L' is bounded by $k_0(1+\delta)a\sqrt{n}$, and each half plane divided by L' has at most $(\frac{2}{3}+\delta)n$ points in P. ∎

Let S_0 be a sequence of n_0 $\{H, P\}$ letters. As we describe our algorithm using recursion, we use the following term to characterize the problem. A 2-dimensional **Multi-Sequence Folding Problem** F is formulated as follows:

The inputs are

i. disjoint subsequences $S_1, S_2, \cdots, S_k$ of sequence S_0 ($S_t = S_0[i_t, j_t]$ for $t = 1, \cdots, k$), and

ii. a region R, where all of the k $\{H, P\}$-sequences are going to be arranged, and

iii. a series of k pairs of grid 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 last letters of S_t, respectively, and

iv. a set of available grid points, which are not occupied by H, P letters, to put the letters from the k sequences, and

v. a set of $\{H, P\}$ grid points on R, which are already occupied by the 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 $S_1, \cdots, S_k$ in the region R that satisfies $f(i_t) = p_t, f(j_t) = q_t (t = 1, 2, \cdots, k)$, has the maximal number of H-H contacts, and $f(i)$ is an available point for each $i \in \cup_{t=1}^{k}[i_t, j_t]$. Those H-H contacts may happen between

two neighbor available positions, and also between an available and an non-available position after the arrangement.

Assume that our input HP sequence has n_0 letters and the optimal folding is inside a $m \times m$ square. A line L partitions a multi-sequence folding problem F into two multi-sequence folding problems F_1 and F_2 in regions $R \cap L_{<0}$ and $R \cap L_{>0}$ respectively by fixing some letters close to L. Furthermore, the available points of F_1 (F_2) are the intersection of F's available points with $L_{<0}$ ($L_{>0}$ resp.).

Algorithm: 2D folding

Input 2-dimensional multi-sequence folding problem F and small constant $\delta > 0$.

(a) folding(F, δ) begin

(b) if n is small, then use exhaustive search to find optimal folding

(c) else

(d) begin

(e) select $\epsilon' > 0$ according to Lemma 20.

(f) For each subset S of $\leq k_0 \cdot \frac{1}{2}(1+\delta) \cdot \sqrt{n}$ letters from $S_1, \cdots, S_k$,
every ϵ'-regular line L' and
every arrangement of S in available points with $\leq (\frac{1}{2}, \frac{1}{2})$ distance to L'

(g) begin

(h) for each partition (by L') making F into problems F_1 and F_2 of size $\leq (\frac{2}{3} + \delta)n$.

(i) begin

(j) Let $M_1 =$ folding(F_1, δ) and $M_2 =$ folding(F_2, δ).

(k) Merge M_1 and M_2 to get a potential solution M for F.

(l) end

(m) end

(n) Output the solution for F with the maximal number of H-H contacts among all of the
potential solutions for F.

(o) end

(p)end

End of the Algorithm

Lemma 21. *(1) For every line segment L of length l, the number of grid points with distance $\leq a$ to at least one point of L is $\leq (2a + \sqrt{2})(l + 2a + \sqrt{2})$.*

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

Proof: (1) If a point p has $\leq a$ distance L, every point in the 1×1 square with center at p has distance $\leq a + \frac{\sqrt{2}}{2}$ to L. The number of those 1×1 squares with center at points of distance $\leq a$ to L is no more than $2(a + \frac{\sqrt{2}}{2})(l + 2a + \sqrt{2})$. (2)The length of a line L inside an $m \times m$ square is $\leq \sqrt{2}m$. Apply (1). ∎

Lemma 22. *For some constants $c_0, \epsilon > 0$, the 2D folding algorithm takes $O(m^{c_0 \log n} n_0^{(5.563-\epsilon)\sqrt{n}})$ time the 2D Multi-Sequence Folding Problem F 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 $d = k_0 a(1 + \delta)$, 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 the Lemma 20, there is an ϵ'-regular 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 $\leq (a, a)$ 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 $\leq (\frac{1}{2}, \frac{1}{2})$ distance to L.

Since the algorithm tries all the arrangements in the separator area, it is easy to verify its correctness. Let $T(n)$ be the computational time for the input with n letters. The analysis can be recursively described by $T(n) = u \cdot T(cn)$, where u is the number of cases to arrange the separators. We will determine the numbers u_1 for the number of ϵ'-regular lines to approximate the optimal separator, u_2 for the number of ways to select $\leq d\sqrt{n}$ letters from the n of them in the input, and u_3 for the number of ways to put those selected letters in the selected ϵ'-regular line. This gives $u = u_1 \cdot u_2 \cdot u_3$ as the upper bound for the number of cases in the separator area.

The number of ϵ'-regular points at every edge of the $m \times m$ square is bounded by $\frac{m}{\epsilon'}$. The total number of ϵ'-regular lines is bounded by $u_1 = \binom{4}{2}(\frac{m}{\epsilon'})^2$. By Stirling 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 the $\leq d\sqrt{n}$ letters from the n of them.

Assume fixed k ($\leq d\sqrt{n}$) letters $S_0[i_1], S_0[i_2], \cdots, S_0[i_k] (1 \leq i_1 < i_2 < \cdots < i_k \leq n)$ are chosen from the disjoint subsequences of S_0. We will select the k grid points $p_1, \cdots, p_k$ to put the k letters on them, respectively. We consider the number of cases to put them to have $\leq (a, a)$ distance to the separator line L. If a point has $\leq a$ normal distance to a line L, it also has $\leq (a, a)$ distance to L. Assume that L_j is the line through the

point $p_j (j = 1, \cdots, k)$ and is vertical to the line L. Let q_j be the intersection between L and L_j. It is easy to see that $\text{dist}(q_j, q_{j+1}) \leq i_{j+1} - i_j$. After the letter $S_0[i_j]$ has been put on a grid point p_j, there are at most $(\alpha(i_{j+1} - i_j))$ ways to select the grid point p_{j+1}, which should have $\leq (a, a)$ distance to L. By Lemma 21, 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 $\leq (a, a)$ distance 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 L with distance $\leq (a, a)$ to L, where $\alpha = (2a + \sqrt{2})(1 + 2a + \sqrt{2})$ is a constant (by Lemma 21). Since $k \leq d\sqrt{n}$, $1 \leq i_1 < i_2 < \cdots < i_k \leq n_0$ and $(i_2 - i_1) + (i_3 - i_2) + \cdots + (i_k - i_{k-1}) = i_k - i_1 \leq n_0$, $\prod_{j=1}^{j=k-1}(\alpha(i_{j+1} - i_j)) \leq (\alpha(\frac{n_0}{k-1}))^{k-1} \leq (\frac{\alpha}{d})^{d\sqrt{n}} n_0^{d\sqrt{n}} n^{-\frac{1}{2}d\sqrt{n}}$ (We use the well known fact that for positive variables $y_1, \cdots, y_{k-1}$ and fixed h with $y_1 + \cdots + y_{k-1} \leq h$, the product $\prod_{t=1}^{k-1} y_{k-1}$ is maximal when $y_1 = y_2 = \cdots = y_k = \frac{h}{k-1}$). The number of ways to arrange the k letters along the separation line (with distance $\leq (a, a)$ to L) is bounded by

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

We have $T(n) \leq u_1 \cdot u_2 \cdot u_3 \cdot T(cn)$. It implies that $T(n) \leq (\frac{mn}{\delta})^{c_0 \log n} 2^{c_0\sqrt{n}} n_0^{d(\frac{1}{1-\sqrt{c}})\sqrt{n}} = O(m^{c_0 \log n} n_0^{(5.563-\epsilon)\sqrt{n}})$ by selecting constants ϵ, δ small enough, and c_0 large enough. ∎

Theorem 23. *There is a $O(n^{5.563\sqrt{n}})$ 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 22. ∎

Chapter 4

Upper bounds for multi-directional width-bounded geometric separators in rectangular and triangular lattices

4.1. Two-dimensional rectangular lattice

The space for protein folding was considered to be a two-dimensional rectangular lattice with the characteristic lattice lengths (a_x, a_y). The spatial position of a point P (Figure 4.1A) was defined by two integer numbers (i, j) that determine the corresponding distances along the two orthogonal axes $x = ia_x$ and respectively, $y = ia_y$. The distance between the origin O of the reference frame and the point P is $d = \sqrt{x^2 + y^2} = \sqrt{(ia_x)^2 + (ja_y)^2}$. Any line through the origin O that passes at a distance smaller than $\pm a_y$ from the point P must have a slope in the range bound by the lines OP_1 and OP_2 (Figure 4.1B). From the triangle OPP_1 we get $\frac{a_y}{\sin\theta_2} = \frac{d}{\sin\alpha}$, which leads to $\sin\theta_2 = \frac{a_y}{d}\sin\alpha = \frac{a_y}{d}\frac{|x|}{d_2} = \frac{a_x a_y |i|}{d d_2}$. Since $d_2 > d - a_y$, the aforementioned equality could be easily transformed into $\sin\theta_2 < \frac{a_x a_y |i|}{d(d-a_y)}$ (Figure 4.1B). Similarly, we derived the relationship $\sin\theta_1 = \frac{a_y}{d}\frac{|x|}{d_1} = \frac{a_x a_y |i|}{d d_1}$, which leads to the inequality $\sin\theta_1 > \frac{a_x a_y |i|}{d(d+a_y)}$. In conclusion, any line through the origin 0 whose slope is bound by the two angels θ_1 and θ_2 crosses the vertical grid through the arbitrary point (x, y) at a distance smaller than the vertical characteristic length of the grid, which is a_y. In other words, the allowed angles are in the range $\frac{a_x a_y |i|}{d(d+a_y)} < \sin\theta_1$ and $\sin\theta_2 < \frac{a_x a_y |i|}{d(d-a_y)}$.

In a similar manner, any line through the origin O that crosses the horizontal grid of the lattice at a distance smaller than $\pm a_x$ from the point P has a slope in the range bound by the lines OP_3 and OP_4 (Figure 4.1C). The angular deviations from the line OP

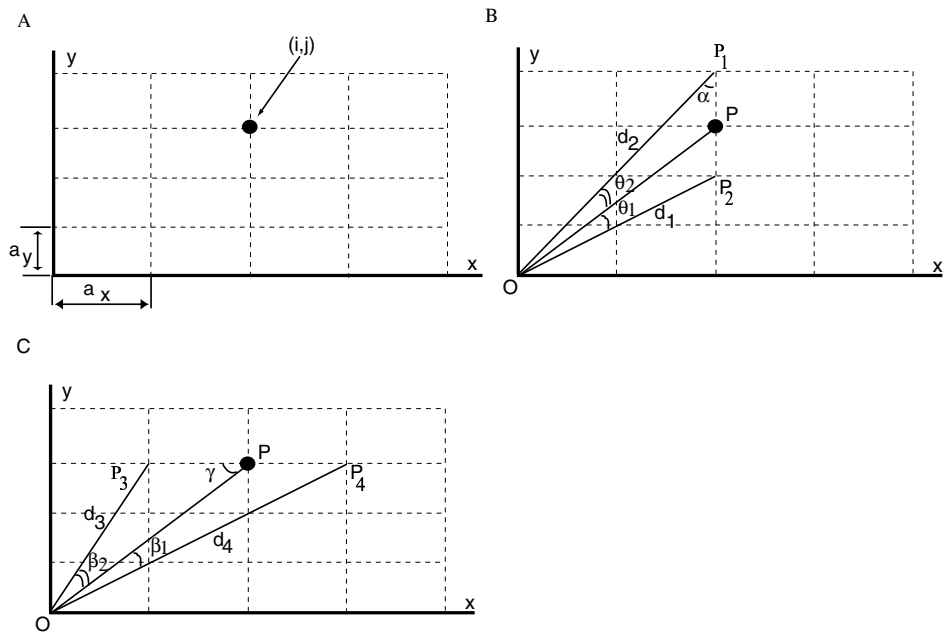

Figure 4.1: (A) Two-dimensional lattice space with characteristic lengths (a_x, a_y). (B) A line with a slope constrained by the two angles θ_1 and θ_2 passes at a distance smaller than a_y. (C) The range of slopes constrained by the two angles β_1 and β_2 determines a crossing with a distance from P smaller than a_x.

is $\sin\beta_1 = \frac{a_x}{d}\frac{|y|}{d_4} = \frac{a_x a_y |j|}{d d_4}$, which leads to the inequality $\sin\beta_1 < \frac{a_x a_y |j|}{d(d-a_x)}$. Similarly, the angular deviation of the point P_3 from the line OP (Figure 4.1C) is $\sin\beta_2 = \frac{a_x}{d}\frac{|y|}{d_3} = \frac{a_x a_y |j|}{d d_3}$, which leads to the inequality $\sin\beta_2 > \frac{a_x a_y |j|}{d(d+a_x)}$. Therefore, the allowed angles for a horizontal crossing at a distance smaller than the characteristic length of the grid (a_x) are in the range $\frac{a_x a_y |j|}{d(d+a_x)} < \sin\beta_1$ and $\sin\beta_2 < \frac{a_x a_y |j|}{d(d-a_x)}$.

In conclusion, any line through the origin O that has the slope in the region bound by $2\max\left(\max\left(\theta_1,\theta_2\right), \max\left(\beta_1,\beta_2\right)\right)$ passes either at a distance smaller than a_y or at a distance smaller than a_x from the arbitrary selected point P. The factor of 2 is necessary to take into account the fact that the angular range is symmetric on both sides of the line OP.

Assuming that $d << a_x$ and $d << a_y$, then $\sin\theta_1 \approx \sin\theta_2 \approx \frac{a_y|x|}{d^2}$ and $\sin\beta_1 \approx \sin\beta_2 \approx \frac{a_x|y|}{d^2}$. Therefore, the range of the angles that determine a crossing at a distance smaller than a_x or a_y from point P is $2\max(\frac{a_y|x|}{d^2}, \frac{a_x|y|}{d^2})$. The geometric locus of all points (x, y) with the property that any arbitrary line through the origin O passes at a distance smaller than a_y from that point is given by the expression $\frac{a_y x}{d^2} = c$, which reduces to $x^2 + y^2 - \frac{a_y}{c}x = 0$. The aforementioned locus is a circle with the center $C_1\left(\frac{a_y}{2c}, 0\right)$ and radius $R_1 = \frac{a_y}{2c}$. Similarly, the geometric locus of all points (x, y) with the property that any arbitrary line through

the origin O passes at a distance smaller than a_x from that point is given by the expression $\frac{a_x y}{d^2} = c$, which reduces to $x^2 + y^2 - \frac{a_x}{c} y = 0$. In this case, the locus is a circle with the center $C_2\left(0, \frac{a_x}{2c}\right)$ and radius $R_2 = \frac{a_x}{2c}$. The arbitrary constant is in the range $-1 < c < +1$ because $sinx \leq 1$ for every x. In summary, the region $A(R_1, R_2)$ determined by the four circles $C_1(R_1, 0)$, $C_1'(-R_1, 0)$, $C_2(0, R_2)$, $C_2'(0, -R_2)$ covers all points in the two-dimensional lattice with the property that any line through the origin O passes either at a distance smaller than a_x or at a distance smaller than a_y (Figure 4.2). The intersection of the two circles C_1 and C_2 is given by the solution of the following equations

$$\begin{aligned} (x - R_1)^2 + y^2 &= R_1^2, \\ x^2 + (y - R_2)^2 &= R_2^2, \end{aligned} \tag{4.1}$$

which determines the intersection point $P_0\left(x_0 = \frac{2R_1R_2^2}{R_1^2+R_2^2}, y_0 = \frac{2R_1^2R_2}{R_1^2+R_2^2}\right)$. The length L of the segment connecting the origin O and the point P_0 is $L = \frac{2R_1R_2}{\sqrt{R_1^2+R_2^2}}$. The angle γ_1 is given by $\cos\gamma_1 = \frac{R_2}{\sqrt{R_1^2+R_2^2}}$ and the subtended angle $\cos\alpha_1 = \frac{R_1^2-R_2^2}{R_1^2+R_2^2}$. In a similar manner, we found the corresponding angles α_2 and γ_2 for the circle C_2. Since $\cos\alpha_2 = -\frac{R_1^2-R_2^2}{R_1^2+R_2^2}$, it results that $\alpha_1 + \alpha_2 = \pi$.

Determining the arbitrary constant c

The arbitrary constant c that determines the geometric locus $A(R_1, R_2)$ has such a value that the four circles cover exactly n grid points. The total area $A(R_1, R_2)$ covered by the four circles is

$$\begin{aligned} A_{total} &= 2\pi R_1^2 + 2\pi R_2^2 - 2R_1^2(\alpha_1 - \sin\alpha_1) - 2R_2^2(\alpha_2 - \sin\alpha_2) \\ &= 2R_1^2(\pi - \alpha_1 + \sin\alpha_1) + 2R_2^2(\pi - \alpha_2 + \sin\alpha_2) \\ &= 2R_1^2(\pi - \alpha_1 + \sin\alpha_1) + 2R_2^2(\alpha_1 + \sin\alpha_1). \end{aligned} \tag{4.2}$$

Let $m = \frac{a_y}{a_x} \in \mathcal{R}^+$ be the ratio of the two characteristic lengths. The total area covered by the four circles becomes

$$\begin{aligned} A_{total} &= 2R_1^2(\pi - \alpha_1 + \sin\alpha_1) + 2R_2^2(\alpha_1 + \sin\alpha_1) \\ &= 2R_1^2(\pi + (m^2 - 1)\alpha_1 + (1 + m^2)\sin\alpha_1) \\ &= 2R_1^2\left(\pi + (m^2 - 1)\arccos\left(\frac{1 - m^2}{1 + m^2}\right) + 2m\right). \end{aligned} \tag{4.3}$$

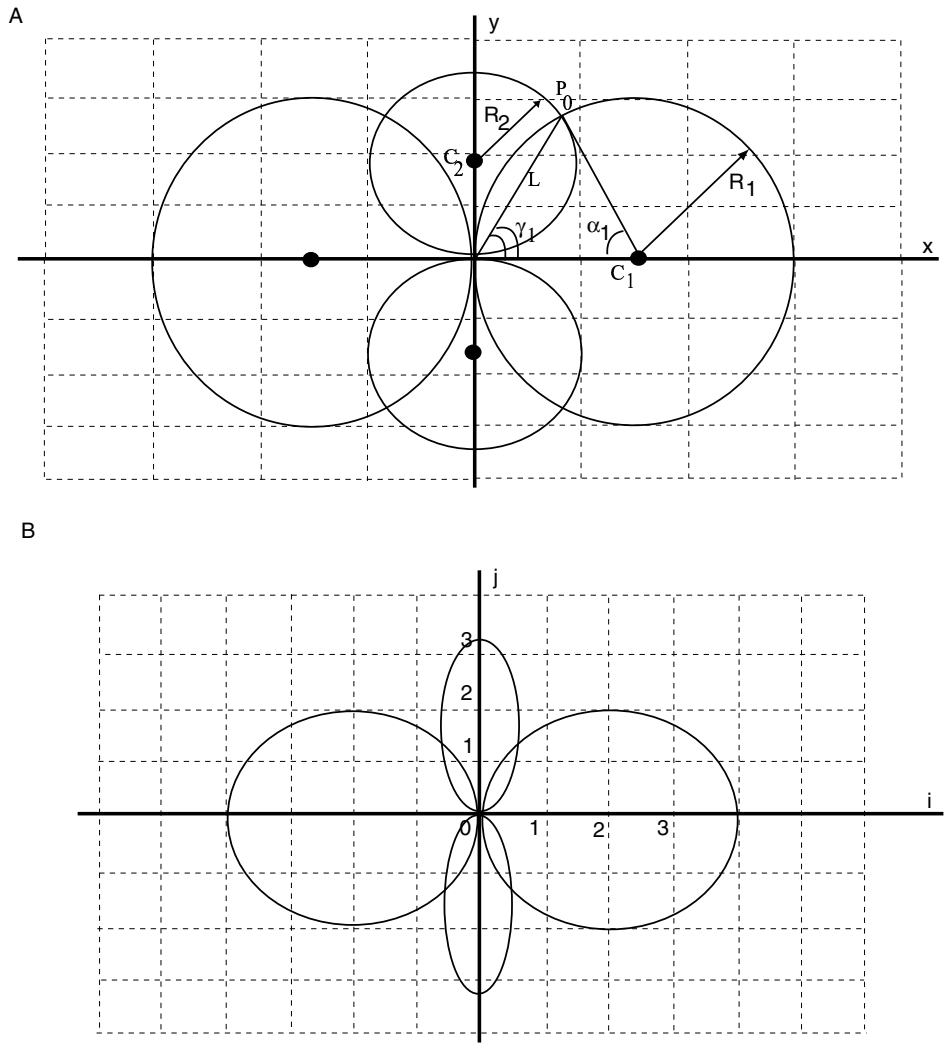

Figure 4.2: (A) Geometric locus of all points in the two-dimensional lattice space with the property that any arbitrary line through the origin passes either at a distance smaller than a_x or a_y from that point. (B) The same geometric locus is the intersection of four ellipses in the space of integer coordinates i versus j. The total number of lattice points included by the four circles in the space x versus y as well as in the i versus j space.

Let us attach to every lattice point a rectangle with the area $S_0 = a_x \times a_y$ centered on that point. The area covered by the associated rectangles for n lattice points is $S = nS_0 = na_xa_y$ (Figure 4.3). For a very large number of lattice points, the area of the region $A(R_1, R_2)$ is smaller compared to the corresponding rectangular coverage. Therefore, $na_xa_y \geq A_{total} = 2\ R_1^2 (\ \pi + (m^2 - 1) \arccos (\ \frac{1-m^2}{1+m^2}\ ) + 2\ m\ )$. The above relationship determines the value of the arbitrary constant $c \geq \sqrt{\frac{\pi+(m^2-1)\arccos\frac{1-m^2}{1+m^2}+2m}{2nm}}$.

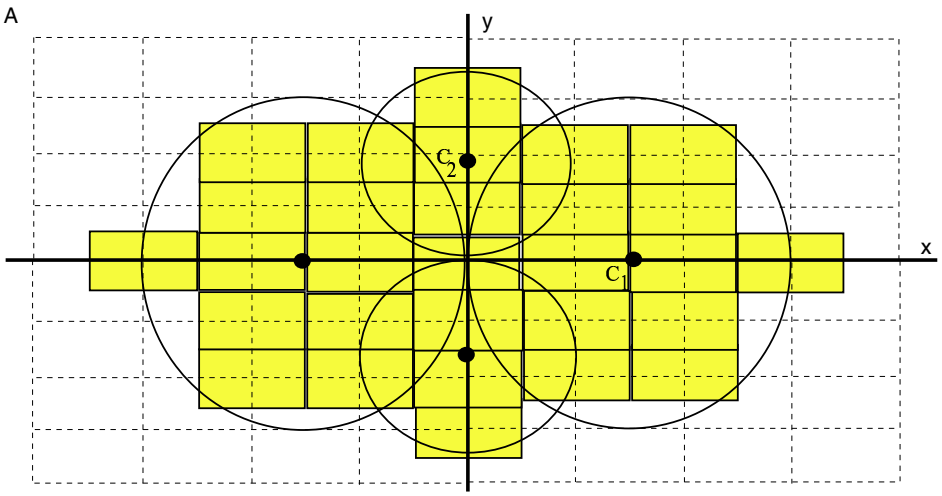

Figure 4.3: Area covered by elementary rectangles $a_x \times a_y$ centered at any grid point is larger than the area covered for the corresponding circles. The circular surfaces cover exactly n grid points and determines the value of constant c.

The probability for crossing the two-dimensional grid at distances smaller than a_x or a_y from an arbitrary point

We already showed that any line through the origin O that has the slope constrained by $2\max\left(\max\left(\theta_1,\theta_2\right),\max\left(\beta_1,\beta_2\right)\right)$ passes either at a distance smaller than a_y or smaller than a_x from the arbitrarily selected point P. In other words, the probability for an arbitrary line through the origin O to pass either at a distance smaller than a_y or smaller than a_x from the arbitrary selected point P is $Prob(O,P,a_x,a_y)=\frac{2}{\pi}\sum\limits_{(x,y)}\max(\frac{a_y|x|}{d^2},\frac{a_x|y|}{d^2})$. The above summation over all points (x,y) inside region $A(R_1,R_2)$ could be replaced by the integral

$$\begin{aligned} S_1 &= \sum_{(x,y)} 2\frac{a_y|x|}{\pi d^2} \approx \frac{2a_y}{\pi}\int_0^{\gamma_1}\int_0^{2R_1\cos\theta}\frac{r\cos\theta}{r^2}rdrd\theta \\ &= \frac{2a_yR_1}{\pi}\int_0^{\gamma_1}2(\cos\theta)^2d\theta = \frac{2a_yR_1}{\pi}\int_0^{\gamma_1}(1+\cos(2\theta))d\theta \\ &= \frac{2a_yR_1}{\pi}\left(\gamma_1+\frac{\sin(2\gamma_1)}{2}\right). \end{aligned} \tag{4.4}$$

Similarly, the sum over all ys could be replaced by an appropriate integral

$$\begin{aligned} S_2 &= \sum_{(x,y)} 2\frac{a_x|y|}{\pi d^2} \approx \frac{2a_x}{\pi}\int_0^{\gamma_2}\int_0^{2R_2\cos\theta} \frac{r\cos\theta}{r^2} r dr d\theta \\ &= \frac{2a_xR_2}{\pi}\left(\gamma_2 + \frac{\sin(2\gamma_2)}{2}\right). \end{aligned} \tag{4.5}$$

Since $\gamma_1 + \gamma_2 = \pi/2$ and $\sin(2\gamma_1) = \sin(\pi - 2\gamma_1) = \sin(2\gamma_1)$, the sum of the two regions with both x and y positive is

$$\begin{aligned} S_1 + S_2 &\approx \frac{2a_xa_y}{c\pi}\left(\frac{\pi}{2} + \sin(2\gamma_1)\right) = \frac{2a_xa_y}{c\pi}\left(\frac{\pi}{2} + \sin(2\gamma_1)\right) \\ &= \frac{2a_xa_y}{c\pi}\left(\frac{\pi}{2} + \frac{2m}{1+m^2}\right). \end{aligned} \tag{4.6}$$

In deriving the above sum $S_1 + S_2$ we used $\cos\gamma_1 = \frac{R_2}{\sqrt{R_1^2+R_2^2}} = \frac{m1}{\sqrt{1+m^2}}$ and $\sin\gamma_1 = \frac{1}{\sqrt{1+m^2}}$.

Since the area covered by all positive values of x and y is only one fourth of the total area $A(R_1, R_2)$ then the expectation is given by

$$\begin{aligned} 4(S_1 + S_2) &\approx \frac{8a_xa_y}{c\pi}\left(\frac{\pi}{2} + \frac{2m}{1+m^2}\right) \\ &\leq \frac{8S_0}{\pi}\left(\frac{\pi}{2} + \frac{2m}{1+m^2}\right)\sqrt{\frac{2m}{\pi + (m^2-1)\arccos(\frac{1-m^2}{1+m^2}) + 2m}}\sqrt{n}, \end{aligned} \tag{4.7}$$

where $S_0 = a_x \times a_y$ is the area of a unit square.

In conclusion, based on the above result, we stated that there exists a line that separates n grid points on the plane such that the number of points with distances less than $(a_x/2, a_y/2)$ to that line is

$$\frac{2}{\pi}\left(\frac{\pi}{2} + \frac{2m}{1+m^2}\right)\sqrt{\frac{2m}{\pi + (m^2-1)\arccos(\frac{1-m^2}{1+m^2}) + 2m}}\sqrt{n} = f(m)\sqrt{n}, \tag{4.8}$$

and each half plane contains less than 2n/3 points. The above result is a corollary of (4.8) for $\frac{S_0}{a_xa_y} = 1/4$.

The function $f(m)$ that determines the upper bound for the linear separator in 2D has its maximum value at $m = 1$ (Figure 4.4). The anisotropy of the lattice space has the advantage of forcing the protein folding along the shortest characteristic length axis. As a result, the upper bound of the 2D separator for the isotropic lattice with $a_x = a_y = a$ (or $m = 1$) is the

maximum possible value. Any other separator in rectangular lattices with $a_x \neq a_y$ gives a lower value for the upper bound than $\approx 1.2074\sqrt{n}$.

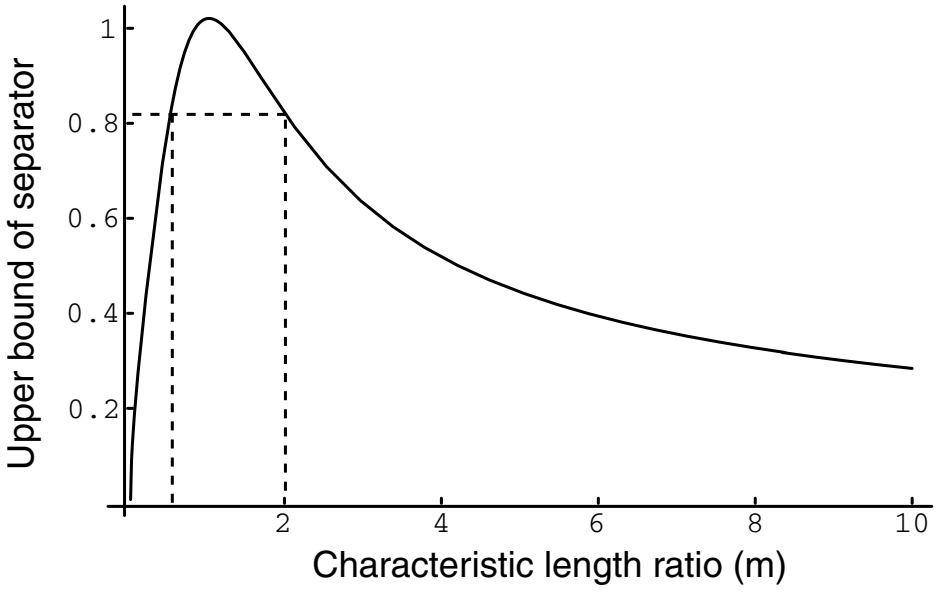

Figure 4.4: The upper bound for 2D grid points separator. For isotropic lattices with $a_x = a_y = a$, or $m = \frac{a_y}{a_x} = 1$, the value of the upper bound is maximum and equal to $\frac{\sqrt{4+2\pi}}{\pi}$. Swapping the x and y axes results from the transformation $m \to 1/m$ that does not change the upper bound. Dashed lines show that $f(1/2) = f(2) \approx 0.813035$

As expected, there is a 90^o rotational invariance which means that the two axes Ox and Oy could be interchanged by changing m to $1/m$.

4.2. Two-dimensional triangular lattice

Let us consider the proteins' folding space to be a two-dimensional triangular lattice with the characteristic lattice length l (Figure 4.5). The spatial position of a point P (Figure 4.5) was defined by the two distances x and y along the two orthogonal axes. The distance between the origin O of the reference frame and the point P is $d = \sqrt{x^2 + y^2}$. In a triangular lattice, any point has six nearest neighbors at a distance l and the corresponding coordinates of the point P_i are

$$\begin{aligned} x_i &= x + l\cos\alpha_i, \\ y_i &= y + l\sin\alpha_i, \end{aligned} \tag{4.9}$$

where l is the characteristic length of the triangular lattice and α_i is the angle between the horizontal axis through the arbitrary point P and the point P_i (Figure 4.5). For a triangular lattice, the angles $\alpha_i = i\frac{\pi}{3}$ with $i \in \{0, 1, 2, 3, 4, 5\}$. Any line through the origin O that

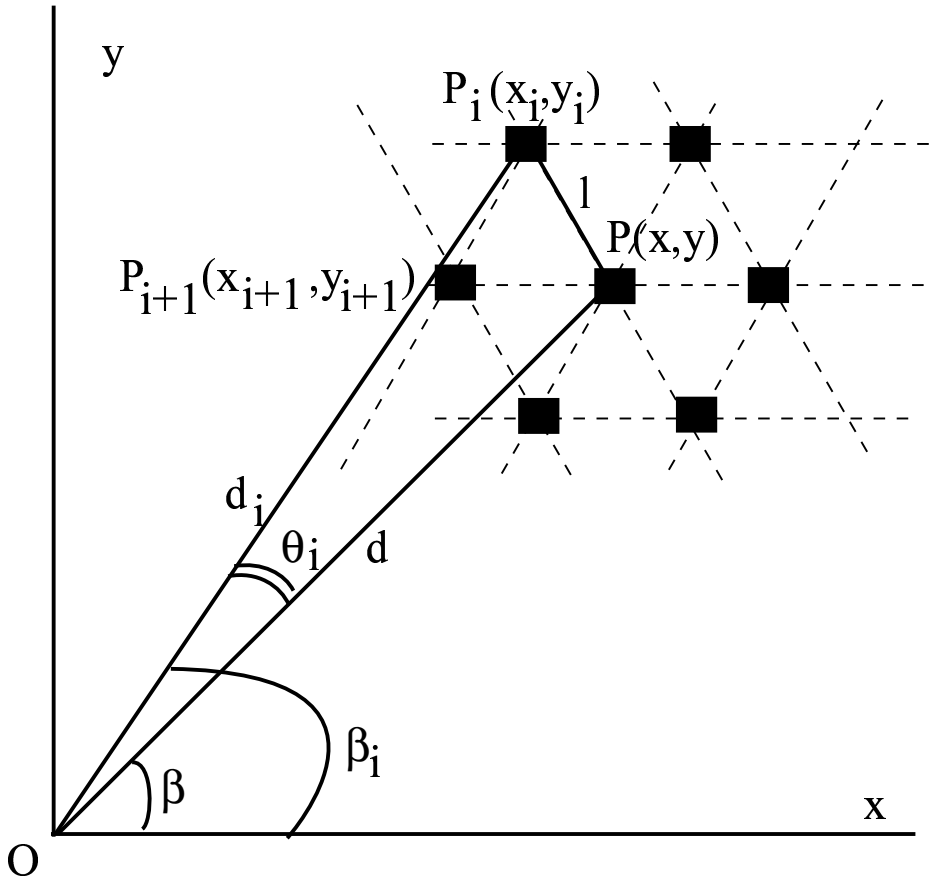

Figure 4.5: Two-dimensional triangular lattice space with characteristic lengths l. A line with a slope constrained by the two angles β and β_1 passes at a distance smaller than l between the two points P and P_i.

passes at a distance smaller than l from the point P must have a slope in the range bound by the lines OP and OP_1 (Figure 4.5). The slope of the line OP is $\tan\beta = \frac{y}{x}$ and the slope of the line OP_i is $\tan\beta_i = \frac{y+l\sin\alpha_i}{x+l\cos\alpha_i}$. As a result, the slope of a line passing between the two lines OP and OP_i is

$$\tan\theta_i = \frac{\frac{y+l\sin\alpha_i}{x+l\cos\alpha_i} - \frac{y}{x}}{1 + \frac{y}{x}\frac{y+l\sin\alpha_i}{x+l\cos\alpha_i}} = \frac{l(x\sin\alpha_i - y\cos\alpha_i)}{x^2 + y^2 + l(x\cos\alpha_i + y\sin\alpha_i)}. \tag{4.10}$$

For very small characteristic length l compared to the distance d form the origin O, the tangent and the angle itself are very close so we can safely assume $\tan\theta_1 \approx \theta_1$.

The geometric locus of all points (x, y) with the property that any arbitrary line through the origin O passes at a distance smaller than l from that point is given by the expression

$$\frac{l(x\sin\alpha_i - y\cos\alpha_i)}{x^2 + y^2 + l(x\cos\alpha_i + y\sin\alpha_i)} = c, \tag{4.11}$$

which reduces to

$$\left(x + \frac{l}{2}\left(\cos\alpha_i - \frac{\sin\alpha_i}{c}\right)\right)^2 + \left(y + \frac{l}{2}\left(\sin\alpha_i + \frac{\cos\alpha_i}{c}\right)\right)^2 = \left(\frac{l}{2}\right)^2\left(1 + \frac{1}{c^2}\right). \tag{4.12}$$

The equation (4.12) describes a circle with the center $C_i\left(-\frac{l}{2}\left(\cos\alpha_i - \frac{\sin\alpha_i}{c}\right), -\frac{l}{2}\left(\sin\alpha_i + \frac{\cos\alpha_i}{c}\right)\right)$ and the radius $R = \left(\frac{l}{2}\right)\sqrt{1 + \frac{1}{c^2}}$.

The six circles corresponding to the six points P_i that are the nearest neighbors of the arbitrary point P have their centers at the same distance R from the origin O and the angle between successive radii is $\pi/3$. There is significant overlapping between the adjacent geometric loci. For example, the geometric locus around the point P_3, which is covered by the circle C_3 (Figure 4.6) significantly overlaps with the geometric locus of the adjacent point P_4. Any point in the area marked A_1 (Figure 4.6) corresponds to an angle θ in Figure 4.5 with the property that a line through the origin O crosses both segments PP_3 and PP_4 at a distance smaller than the characteristic length l. In addition to the aforementioned overlapping between any two adjacent circles C_i and C_{i+1}, there is an overlapping between C_i and C_{i+2}. For example, the intersection between the circles C_3 and C_5 is given by the region marked A_2 in Figure 4.6. Any point in the region A_2 belongs to three different geometric loci and it means that the corresponding line with the slope θ in Figure 4.5 crosses all three sides PP_3, PP_4, and PP_5 at a distance smaller than the characteristic length l. The intersection of the two adjacent circles, for example C_1 and C_2, is given by the solution of the following equations

$$\begin{aligned} (x - x_1)^2 + (y - y_1)^2 &= R^2, \\ (x - x_2)^2 + (y - y_2)^2 &= R^2. \end{aligned} \tag{4.13}$$

Due to the symmetry of the problem and based on the fact that the slopes of successive radii OC_i (Figure 4.6) are separated by $\pi/3$, the intersection between the circle C_i and C_{i+2} is the center of the circle C_{i+1}. From symmetry considerations we also concluded that the

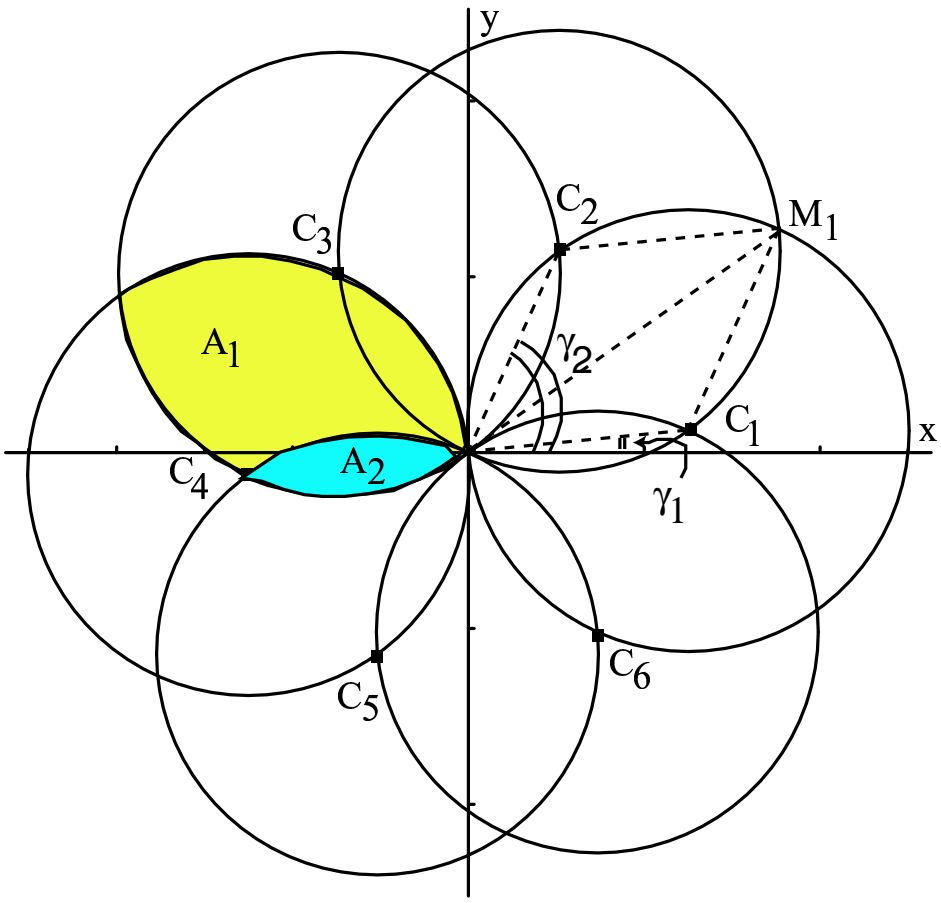

Figure 4.6: Geometric locus of all points in the two-dimensional triangular lattice space with the property that any arbitrary line through the origin crosses the side *Pepsi* at a distance smaller than the characteristic length l. The overlapping between the geometric loci corresponding to different points P_i show possible multiple crossings of different sides by the same arbitrary line OP.

intersection between the circles C_i and C_{i+1} is on the bisector line of the angle C_iOC_{i+1}.

Determining the arbitrary constant c

The arbitrary constant c that determines the geometric locus $A = \bigcup_{i=0,5} C_i$ has such a value that the six circles cover exactly n grid points. The total area A covered by the six circles is

$$\begin{aligned} A_{total} &= 6\pi R^2 - 6A_1 + 6A_2 \\ &= 6\pi R^2 - 6\frac{R^2}{2}\left(\frac{2\pi}{3} - \sin\frac{2\pi}{3}\right) + 6\frac{R^2}{2}\left(\frac{\pi}{3} - \sin\frac{\pi}{3}\right) = 5\pi R^2. \end{aligned} \tag{4.14}$$

Let us attach to every lattice point a triangle with the side equal to the characteristic length l and the corresponding area $S_0 = \frac{l^2\sqrt{3}}{4}$ centered on that point. The area covered by the associated rectangles for n lattice points is $S = nS_0 = n\frac{l^2\sqrt{3}}{4}$ (Figure 4.7). For a very

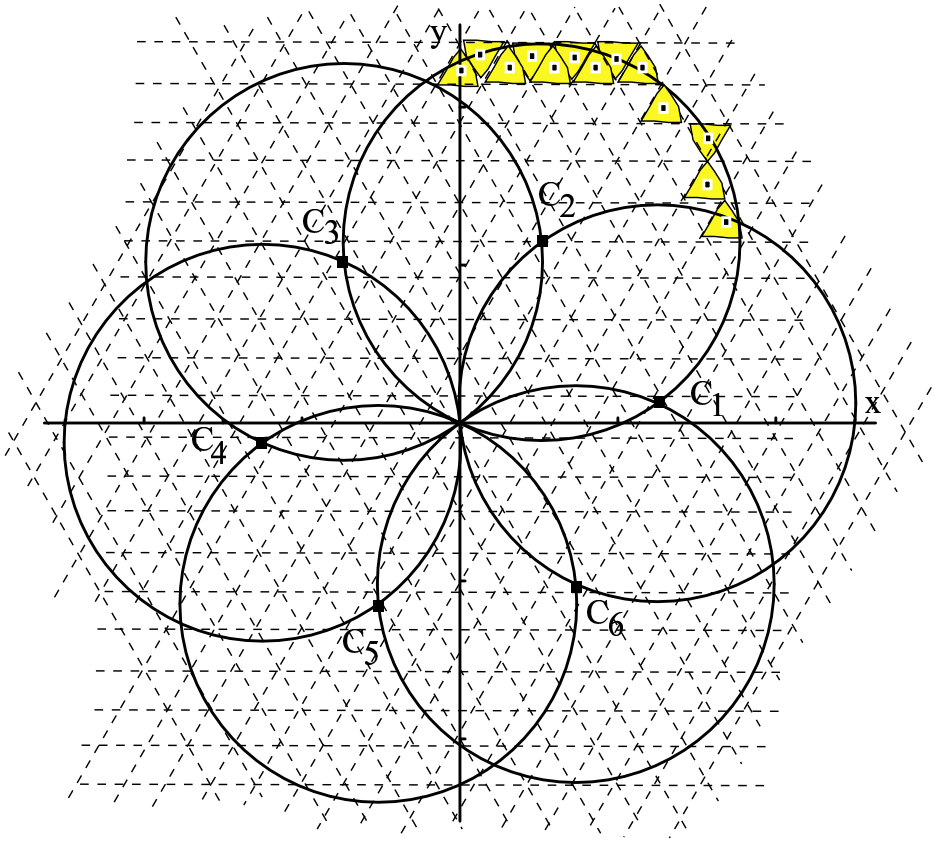

Figure 4.7: Area covered by the elementary triangles is larger than the area covered for the corresponding circles. The circular surfaces cover exactly n grid points and determines the value of constant c.

large number of lattice points, the area of the region $A = \bigcup_{i=0,5} C_i$ is smaller compared to the corresponding triangle coverage. Therefore, $n\frac{l^2\sqrt{3}}{4} \geq A_{total} = 5\pi R^2$, which determines the value of the arbitrary constant $1 + \frac{1}{c^2} \leq \frac{\sqrt{3}}{5\pi}n$.

The probability for crossing the two-dimensional triangular grid at a distance smaller than the characteristic length l from an arbitrary point

We showed that any line through the origin O that has the slope smaller than θ_i crosses the side PP_i at a distance smaller than the characteristic length l from the arbitrarily selected point P (Figure 4.5). In other words, the probability for an arbitrary line through the origin O to pass at a distance smaller than then the characteristic length l from the arbitrary selected point P is

$$Prob(O,P,l) \approx \frac{1}{\pi}\frac{l(x\sin\alpha_i - y\cos\alpha_i)}{x^2+y^2+l(x\cos\alpha_i + y\sin\alpha_i)}, \tag{4.15}$$

where the angles $\alpha_i = i\frac{\pi}{3}$ with $i \in \{0,1,2,3,4,5\}$ refers to the six nearest neighbors of the arbitrary selected point P. In deriving the above expression for the probability we used (4.10) under the assumption that $\tan\theta_i \approx \theta_i$ for points far away from the origin O compared to the characteristic length of the triangular lattice.

In order to get the expectation, a summation over all possible positions (x,y) inside region A is required. A good estimation of the sum is the integral

$$\begin{aligned}
S_1 &= \sum_{(x,y)} \frac{1}{\pi}\frac{l(x\sin\alpha_i - y\cos\alpha_i)}{x^2+y^2+l(x\cos\alpha_i + y\sin\alpha_i)} \\
&= \frac{4}{l\pi\sqrt{3}} \int_{\gamma_i}^{\gamma_i+\pi/6}\int_0^{2R\cos\phi} \frac{r\cos\phi\sin\alpha_i - r\sin\phi\cos\alpha_i}{r^2+l(r\cos\phi\cos\alpha_i + r\sin\phi\sin\alpha_i)} r dr d\phi \\
&= \frac{4}{l\pi\sqrt{3}} \int_{\gamma_i}^{\gamma_i+\pi/6}\int_0^{2R\cos\phi} \frac{r(\cos\phi\sin\alpha_i - \sin\phi\cos\alpha_i)}{r+l(\cos\phi\cos\alpha_i + \sin\phi\sin\alpha_i)} dr d\phi \\
&= \frac{4}{l\pi\sqrt{3}} \int_{\gamma_i}^{\gamma_i+\pi/6}\int_0^{2R\cos\phi} \frac{r\sin(\alpha_i-\phi)}{r+l\cos(\alpha_i-\phi)} dr d\phi \\
&= \frac{4}{l\pi\sqrt{3}} \int_{\gamma_i}^{\gamma_i+\pi/6} \sin(\alpha_i-\phi) \int_0^{2R\cos\phi} \left(1 - \frac{l\cos(\alpha_i-\phi)}{r+l\cos(\alpha_i-\phi)}\right) dr d\phi \\
&= \frac{4}{l\pi\sqrt{3}} \int_{\gamma_i}^{\gamma_i+\pi/6} \sin(\alpha_i-\phi)(r - l\cos(\alpha_i-\phi)\ln(r+l\cos(\alpha_i-\phi)))_0^{2R\cos\phi} d\phi
\end{aligned} \tag{4.16}$$

$$= \frac{4}{l\pi\sqrt{3}} \int_{\gamma_i}^{\gamma_i+\pi/6} \sin(\alpha_i-\phi)\left(2R\cos\phi - l\cos(\alpha_i-\phi)\ln\left(\frac{2R\cos\phi + l\cos(\alpha_i-\phi)}{l\cos(\alpha_i-\phi)}\right)\right)d\phi,$$

where we used the polar coordinates $x = r\cos\phi$, $y = r\sin\phi$, replaced the area element $dxdy$ by $rdrd\phi$, and normalized the integrals by the area $\frac{l^2\sqrt{3}}{4}$ which is the area of the elementary triangle. The last integral can be decomposed into two distinct integrals. The first integral is

$$\begin{aligned} I_1 &= \frac{8R}{l\pi\sqrt{3}} \int_{\gamma_i}^{\gamma_i+\pi/6} \sin(\alpha_i-\phi)\cos\phi d\phi \\ &= \frac{8R}{l\pi\sqrt{3}}\frac{1}{12}(\pi\sin\alpha_i - 3\sin(\alpha_i - 2\gamma_i - \pi/6)). \end{aligned} \tag{4.17}$$

In order to find a numerical value for the integral (4.17) we considered $\alpha_i = 0$ and the corresponding angle γ_i was found based on Figure 4.6 as being $\tan\gamma_i = \frac{y_{C1}}{x_{C1}} = \frac{\sin\alpha_i + \frac{\cos\alpha_i}{c}}{\cos\alpha_i - \frac{\sin\alpha_i}{c}} = \frac{1}{c}$. By substituting the aforementioned values into (4.17) we get

$$I1 = \frac{2R}{l\pi\sqrt{3}}\sin(2\gamma_1+\pi/6) = \frac{R}{l\pi\sqrt{3}}(\sqrt{3}\sin 2\gamma_1 + \cos 2\gamma_1). \tag{4.18}$$

Since $\sin 2\gamma_1 = 2\frac{\tan\gamma_1}{1+(\tan\gamma_1)^2} = 2\frac{1/c}{1+1/c^2}$ and $\cos 2\gamma_1 = 2(\cos\gamma_1)^2 - 1 = \frac{2}{1+1/c^2} - 1$, which gives us $\sqrt{3}\sin 2\gamma_1 + \cos 2\gamma_1 = 2\sqrt{3}\frac{1/c}{1+1/c^2} + \frac{2}{1+1/c^2} - 1$. By taking into consideration the fact that the arbitrary constant c was previously determined from $1 + \frac{1}{c^2} \leq \frac{\sqrt{3}}{5\pi}n$ we easily find that both limits $\lim_{n\to\infty}\frac{1/c}{1+1/c^2}$ and $\lim_{n\to\infty}\frac{1}{1+1/c^2}$ vanishe. This means that for a large number n of lattice points, the harmonic part of the integral (4.17) is simply $\lim_{n\to\infty}\sqrt{3}\sin 2\gamma_1 + \cos 2\gamma_1 = -1$. Therefore, for large n the integral (4.17) reduces to $|I1| = \frac{R}{l\pi\sqrt{3}} = \frac{1}{2\pi\sqrt{3}}\sqrt{1+1/c^2} = O(\sqrt{n})$.

We also solved the second integral in (4.17)

$$I_2 = -\frac{2}{\pi\sqrt{3}} \int_{\gamma_i}^{\gamma_i+\pi/6} \sin 2(\alpha_i-\phi)\ln\left(\frac{2R\cos\phi}{l\cos(\alpha_i-\phi)} + 1\right)d\phi. \tag{4.19}$$

The simplest possible form of I_2 can be obtained for $\alpha_i = 0$ and leads to

$$I_2 = \frac{2}{\pi\sqrt{3}}\ln\left(\frac{2R}{l}+1\right)\int_{\gamma_i}^{\gamma_i+\pi/6}\sin(2\phi)d\phi = -\frac{l^2}{4\pi}\ln\left(\frac{2R}{l}+1\right)\sin(2\gamma_1+\pi/6)$$

$$
\begin{aligned}
&= -\frac{1}{2\pi\sqrt{3}} \ln\left(\frac{2R}{l}+1\right)\left(\sqrt{3}\frac{2\tan\gamma_1}{1+(\tan\gamma_1)^2}+\frac{2}{1+(\tan\gamma_1)^2}-1\right) \\
&= -\frac{1}{2\pi\sqrt{3}} \ln\left(\frac{2R}{l}+1\right)\left(\sqrt{3}\frac{2/c}{1+1/c^2}+\frac{2}{1+1/c^2}-1\right) \qquad (4.20)
\end{aligned}
$$

Based on the fact that the radius R is $R=\frac{l}{2}\sqrt{1+\frac{1}{c^2}}$ and $1+\frac{1}{c^2}\leq\frac{\sqrt{3}}{5\pi}n$ we can easily determine the limits of the three terms in (4.20) for very large number of grid points. For example, the first term $-\frac{2}{\pi\sqrt{3}}\ln\left(\frac{2R}{l}+1\right)\left(\sqrt{3}\frac{2/c}{1+1/c^2}\right)=-\frac{1}{2\pi\sqrt{3}}\left(\sqrt{3}\frac{2/c}{1+1/c^2}\right)\ln\left(\sqrt{1+1/c^2}+1\right)$.

Since $1+1/c^2=O(n)$ then $\lim_{n\to\infty}=-\frac{l^2\sqrt{3}}{4\pi}\frac{O(\sqrt{n})}{O(n)}\ln(O(n))=0$. The second term in (4.20) also vanishes in the limit of large n. Only the third term in (4.20) diverges because $\frac{1}{2\pi\sqrt{3}}\ln\left(\frac{2R}{l}+1\right)=\frac{1}{2\pi\sqrt{3}}\ln\left(\sqrt{1+1/c^2}+1\right)=O(\ln(n))$.

In conclusion, the dominant term in the sum S_1 of (4.17) is I_1, which is of the order of $O(\sqrt{n})$. Therefore, for very large n we can simply approximate $S_1\approx\frac{1}{2\pi\sqrt{3}}\sqrt{1+1/c^2}\leq\frac{1}{2\pi\sqrt{3}}\sqrt{\frac{\sqrt{3}}{5\pi}}\sqrt{n}$. Since we need to consider 12 identical regions and the fact that the separator line must pass at a distance $l/2$ from the arbitrary lattice point considered then the total area is $S_{total}=24S_1=\frac{12}{\pi\sqrt{5\pi\sqrt{3}}}\sqrt{n}=0.7323\sqrt{n}$.

Chapter 5

Conclusions

We used the divide and conquer method to solve recursively the two-dimensional problem of optimal folding of a HP sequence with the maximum number of H-H contacts. Fu [25] introduced the concept of width-bounded geometric separator and found improved bounds for a line separator of square gird graphs. The present work introduced the concept of multi-directional width bounded separators and improved the bounds for the grid graph separator problem. For a grid graph G with n grid points P, there exists a separator $A \subseteq P$ such that A has less than or equal to $1.02074\sqrt{n}$ points, and $G - A$ has two disconnected subgraphs with less than or equal to $\frac{2}{3}n$ nodes on each of them. We also found a $0.7555\sqrt{n}$ lower bound for such a separator on grid graph. Once we determined a balanced separator for the grid graph then the analysis can be recursively described in terms of computational time by $T(n) = u \cdot T(cn)$, where u is the number of cases to arrange the separators and $0 < c < 1$ is the fraction of the original HP sequence in each subgraph. Based on our multi-directional width-bounded geometric separator, we found that there is an $O(n^{5.563\sqrt{n}})$ time algorithm for the 2D protein folding problem in the HP model. We also derive $0.7555\sqrt{n}$ lower bound for such a separator on grid graph and extended the upper bound results for the line separator in 2D to rectangular and triangular lattices.

Bibliography

[1] Nrc report: Physics in a new era, overview. 2001.

[2] R. Agarwala, S. Batzoglou, V. Dančík, S. Decatur, S. Hannenhalli, M. Farach, S. Muthukrishnan, and S. Skiena. Local rules for protein folding on a triangular lattice and generalized hydrophobicity in HP model. In *Proceedings of the First Annual International Conference on Computational Molecular Biology*, pages 1–2, 1997.

[3] R. Agarwala, S. Batzoglou, V. Dančík, S. Decatur, S. Hannenhalli, M. Farach, S. Muthukrishnan, and S. Skiena. Local rules for protein folding on a triangular lattice and generalized hydrophobicity in HP model. *Journal of Computational Biology*, 4(2):275–296, 1997.

[4] R. Agarwala, S. Batzoglou, V. Dančík, S. Decatur, S. Hannenhalli, M. Farach, and S. Skiena. Local rules for protein folding on a triangular lattice and generalized hydrophobicity in HP model. In *8th Annual ACM-SIAM Symposium on Discrete Algorithms, Proceedings*, pages 390–399, 1997.

[5] N. Alon, P. Seymour, and R. Thomas. Planar separator. *SIAM J. Discr. Math.*, 7(2):184–193, 1990.

[6] Noga Alon, Paul Seymour, and Robin Thomas. A separator theorem for graphs with an excluded minor and its applications. In *22nd annual ACM symposium on theory of computing*, pages 293–299, 1990.

[7] G. B. Anfinsen and H. A. Scheraga. Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.*, 29:205–300, 1975.

[8] T. Leighton B. Berger. Protein folding in the hp model is np-complete. In *Proceedings of the Second International Conference on Computational Molecular Biology (RECOMB)*, pages 30–39. ACM Press, New York, 1998.

[9] R. Backofen. Constraint techniques for solving the protein structure prediction problem. In *4th International conference on principle and practice of constrain programming, Lecture Notes in Computer Science*, pages 72–86. Springer-Verlag, 1998.

[10] U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, and Nadler. Testing a new monte carlo algorithm for protein folding. *Protein: Structure, Function, and Genetics*, 32:52–66, 1998.

[11] B. Berger and T. Leighton. Protein folding in the hydrophobic-hydrophilic (hp) model is np-complete. *Journal of Computational Biology*, 5(1):27–40, 1998.

[12] S.N. Bhatt and F.T. Leighton. A framework for solving vlsi graph layout problems. *Journal of computer and systems science*, 28(2):300–343, 1982.

[13] H.S. Chan and K.A. Dill. The protein folding problem. *Physics Today*, page 1993, 1993.

[14] N. Chiba, T. Nishizeki, and T. Saito. Applications of the lipton and tarjan planar separator theorem. *Journal of information processing letters*, 4(4):203–207, 1981.

[15] E. Ciszak and G.D Smith. Crystallographic evidence for dual coordination around zinc in the t3r3 human insulin hexamer. *Biochemistry*, 33(6):1512–1517, 1994.

[16] P. Clote and R. Backofen. *Computational Molecular Biology: an Introduction*. John Wiley and Sons Ltd., 2000. Wiley Series in Mathematical and Computational Biology.

[17] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. *Journal of Computational Biology*, 5(3):409–422, 1998.

[18] R. Dawkins. *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe without Design*. USA Publisher: W. W. Norton and Company, 1986.

[19] K.A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990.

[20] K.A. Dill, S. Bromberg, K. Yue, K.N. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. Principles of protein folding a perspective from simple exact models. *Protein Science*, 4:561–602, 1995.

[21] H.N. Djidjev and S.M. Venkatesan. Reduced constants for simple cycle graph separation. *Acta informatica*, 34:231–234, 1997.

[22] Hristo Nicolov Djidjev. On the problem of partitioning planar graphs. *SIAM J. ALG. DISC. METH.*, 3(4):229–240, 1982.

[23] J. Eckhoff. *Handbook of Convex Geometry*, chapter 2.1 Helly, Radon, and Caratheodory Type Theorems, pages 389–448. North-Holland, Amsterdam, Netherlands, 1993.

[24] David Eisenberg. The discovery of the α-helix and β-sheet, the principal structural features of proteins. In *Proceedings of the National Academy of Sciences USA*, volume 100, pages 11207–11210, 2003.

[25] Bin Fu. Theory and application of width bounded geometric separator. *Electronic Colloquium on Computational Complexity*, 13:1–14, 2005.

[26] Bin Fu, Sorinel Adrian Oprisan, and Lizhe Xu. Multi-directional width-bounded geometric separator and protein folding. (submitted to publication), 2005.

[27] Bin Fu and Wei Wang. A $2^{O(n^{1-1/d}\log n)}$-time algorithm for d-dimensional protein folding in the hp-model. In *International Colloquium on Automata, Languages and Programming (ICALP'2004), Lecture Notes in Computer Science*, volume 3142, pages 630–644. Springer, 2004.

[28] H. Gazit. An improved algorithm for separating a planar graph. preprint, 1986.

[29] J.R. Gilbert, J.P. Hutchinson, and R.E. Tarjan. A separation theorem for graphs of bounded genus. *Journal of algorithm*, 5:391–407, 1984.

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

[31] B. Hayes. Computing science: Prototeins. *American Scientist*, 86(3):216–221, 1998.

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

[33] N. Krasnogor, D. Pelta, P.M. Lopez, P. Mocciola, 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.

[34] C. Levinthal. Mossbauer spectroscopy in biological systems. In Munck E. DeBrunner J. T. P., editor, *Proceedings of a meeting held at Allerton House*, pages 22–24, 1969.

[35] F. Liang and W.H. Wong. Evolutionary monte carlo for protein folding simulations. *Journal of Chemical Physics*, 115(7):3374–3380, 2001.

[36] R.J. Lipton and R. Tarjan. A separator theorem for planar graph. *SIAM journal on computing*, 9(3):615–627, 1979.

[37] G.L. Miller, S.-H. Teng, and S.A. Vavasis. A unified geometric approach to graph separators. In *Proc. 32nd IEEE Symposium on Foundations of Computer Science (FOCS'91)*, pages 538–547. IEEE Computer Society Press, 1991.

[38] G.L. Miller and W. Thurston. Separators in two and three dimensions. In *22nd Annual ACM symposium on theory of computing*, pages 538–547, 1991.

[39] G.L. Miller and S.A. Vavasis. Density graphs and separators. In *Second annual ACM-SIAM symposium on discrete algorithms, ACM-SIAM*, pages 331–336, 1991.

[40] A. Neumaier. Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM Review*, 39(3):407 – 460, 1997.

[41] A. Newman. A new algorithm for protein folding in the hp model. In *13th ACMSIAM Symposium on Discrete Algorithms*, pages 876–884, 2002.

[42] M.W. Nirenberg, J.H. Matthaei, O.W. Jones, R.G. Martin, and S.H. Barondes. Approximation of genetic code via cell-free protein synthesis directed by template rna. *Fed Proc.*, 22:55–61, 1963 Jan-Feb.

[43] J. Pach and P.K. Agarwal. *Combinatorial Geometry*. Wiley-Interscience Publication, 1995.

[44] A. W.P.III Patton and Goldman E. A standard ga approach to native protein conformation prediction. In *6th Intl Conf Genetic Algorithms*, pages 574–581, 1995.

[45] A. Piccolboni and G. Mauri. Application of evolutionary algorithms to protein prediction. In N. E. A. Kasabov, editor, *Proceedings of I-CONIP97*, pages 574–581. Springer, 1998.

[46] S. Plotkin, S. Rao, and W.D. Smith. Shallow excluded minors and improved graph decomposition. In *SIAM 5th Symp. on Discrete Algorithms*, pages 462–470, 1990.

[47] A.A. Rabow and Scheraga H.A. Improved genetic algorithm for the protein folding problem by use of a cartesian combination operator. *Protein Science*, 5:1800–1815, 1996.

[48] 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*, 106:2418, 1997.

[49] S.S. Ravi and H.B. III Hunt. Application of the planar separator theorem to computing problems. *Information processing letter*, 25(5):317–322, 1987.

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

[51] W.D. Smith and N.C. Wormald. Application of geometric separator theorems. In *FOCS 1998*, pages 232–243, 1998.

[52] D.A. Spielman and S.H. Teng. Disk packings and planar separators. In *12th Annual ACM Symposium on Computational Geometry*, pages 349–358, 1996.

[53] U. Unger and J. Moult. A genetic algorithm for three dimensional protein folding simulations. In *5th Internatinal Confernece on Genetic Algorithms*, pages 581–588, 1993.

[54] U. Unger and J. Moult. Genetic algorithms for protein folding simulations. *Journal of Molecular Biology*, 231(1):75–81, 1993.

[55] J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 4356, April 25, 1953.

[56] R. Weinstock. *Calculus of variations*. McGraw-Hill, 1952.

[57] K. Yue and K.A. Dill. Sequence-structure relationships in proteins and copolymers. *Review E*, 48:2267–2278, 1993.

Vita

Sorinel Adrian Oprisan was born in Salceni, Vaslui, Romania. He graduated from "Alexandru Ioan Cuza" University of Iasi, Romania, in 1987. His first appointment as a middle school physics teacher was in Tecuci, Romania. Between 1990 and 1997 he functioned as Assistant Professor of Theoretical Physics at "Alexandru Ioan Cuza" University of Iasi, Romania. Between 1997 and 1999 he functioned as a Lecture and in 1999 was promoted to Associate Professor of Theoretical Physics at "Alexandru Ioan Cuza" University. In 1998 he obtained his Doctor of Philosophy degree with a thesis on statistical foundations of self-organizing systems under the supervision of Professor Margareta Ignat.

In 1999 he moved to the United States of America to work in computational neuroscience with Dr. Carmen C. Canavier at the University of New Orleans. At the University of New Orleans he also served as adjunct faculty for the Department of Physics. Dr. Oprisan taught a wide range of physics courses from general physics to statistical mechanics, thermodynamics, electrodynamics, classical mechanics, computational physics, nonlinear dynamics, statistical mechanics of nonequilibrium phenomena, and self-organizing phenomena. His research interests cover statistical mechanics, thermodynamics, nonlinear dynamics, chaos and fractals, biophysics, biocomplexity, and computational neuroscience. He published more than 40 peer-reviewed papers in prestigious journals such as Physical Review, Journal of Physics A, Physics Letters A, Bioinformatics, Chaos Solitons and Fractals, Chaos, Neural Computation, Complex Systems, Journal of Differential Equations and Dynamical Systems, Romanian Journal of Chemistry, Neurocomputing, and Biophysical Journal.

In 2001 he was admitted to the Master's program in Computer Science at the University of New Orleans.