TEL-AVIV UNIVERSITY
RAYMOND AND BEVERLY SACKLER
FACULTY OF EXACT SCIENCES
SCHOOL OF COMPUTER SCIENCE

Efficient Construction of Pathways in the Complement of the Union of Balls in $\mathbb{R}^3$

Thesis submitted in partial fulfillment of the requirements for the M.Sc. degree in the School of Computer Science, Tel-Aviv University

by

Eitan Yaffe

The research work for this thesis has been carried out at
Tel-Aviv University
under the supervision of Prof. Dan Halperin

September 2007

Acknowledgments

I deeply thank my supervisor, Prof. Dan Halperin, for his guidance and his help during the work on this thesis. I wish to thank him for his patience, support and encouragement.

I thank Dan Fishelovitch and his lab for fruitful cooperation.

I thank Jean-Daniel Boissonnat, David Cohen-Steiner and all members of INRIA Sophia Antipolis for their help and hospitality.

I thank all the fellow students in the Geometry lab for their assistance.

I thank my brother, Yoav Yaffe, for fruitful discussions and inspiration.

Work in this thesis has been supported in part by the IST Programme of the EU as Shared-cost RTD (FET Open) Project under Contract No IST-006413 (ACS - Algorithms for Complex Shapes), by the Israel Science Foundation (grant no. 236/06), and by the Hermann Minkowski–Minerva Center for Geometry at Tel Aviv University.

Abstract

Given a collection of balls in three-dimensional space we wish to efficiently identify pathways in the complement of their union. The desired pathways should balance between length and clearance. Namely, we prefer short and wide pathways between a given start point and goal point positioned in the complement. In this thesis we provide an algorithm for identifying good pathways of this type and an efficient implementation of the algorithm. A major contribution of the thesis is the notion of the *pathway diagram* which contains an approximation of an idealistic construct related to the medial axis, and which unlike the medial axis is easy to compute in the case of the complement of the union of balls. We provide theoretical analysis of the approximation qualities of the pathway diagram. On top of the algorithm we have developed a software package, MolAxis, to assist the biologist/biochemist to automatically identify good pathways in the complement of molecules. We present experimental results that demonstrate the efficiency of our software in finding pathways in the complement of molecules and attest to the effectiveness of our approximation scheme.

Contents

Part I

Introduction and Algorithm

Chapter 1

Introduction

Let $\mathcal{B}$ be a finite collection of three-dimensional balls and let $\cup\mathcal{B}$ denote their union, namely $\bigcup_{B\in\mathcal{B}} B$. We assume, without loss of generality, that every ball in $\mathcal{B}$ is not smaller than a unit ball. We wish to identify pathways in the complement of $\cup\mathcal{B}$ that balance between length and clearance[1]. The *medial axis* of the complement of $\cup\mathcal{B}$ is the set of points in this complement that have more than one closest points in $\cup\mathcal{B}$. A recent result by Lieutier [20] states that under certain conditions the medial axis of an object and the object itself have the same 'shape' (the technical term is homotopy equivalent), making it suitable for finding the desired pathways.

We follow a standard practice in biology of modeling a molecule by a collection of three-dimensional balls, one ball per atom. The term *channel* is often used in molecular biology to refer to a probable route taken by a small molecule passing through a hole in the molecule. A *pathway* is a curve in the space that lies outside the molecule. If $\mathcal{B}$ is the set of atom balls of a molecule then the boundary surface of $\cup\mathcal{B}$ is called the *van der Waals surface* of the molecule. The *clearance* of a point outside the molecule is the distance between the point and the van der Waals surface of the molecule. Pathways are not unique and more than one pathway may exist between two points. There are several ways to define an optimal pathway between two points. The shortest pathway between two given points typically has the undesirable property that it touches the boundary of the molecule and hence has zero clearance. High clearance pathways, on the other hand, can be extremely long. We are interested in finding pathways that balance between length and clearance. We call such pathways *corridors* [25] and use them to represent channels; a formal definition is given in part III. We construct corridors to represent channels since they are well defined geometric entities that can be approximated in an efficient manner.

The exact medial axis of the complement of the union of balls is a subset of the Voronoi diagram [6, 8] of the balls and can be computed in an exact manner as shown by Boissonnat and Delage [7]. We opt for an approximation approach for two main reasons: simplicity of implementation and speed of computation. There are algorithms that approximate the medial axis of an object from a set of unorganized points sampled on the surface

[1] The *clearance* of a point p on the pathway is distance of p from $\cup\mathcal{B}$.

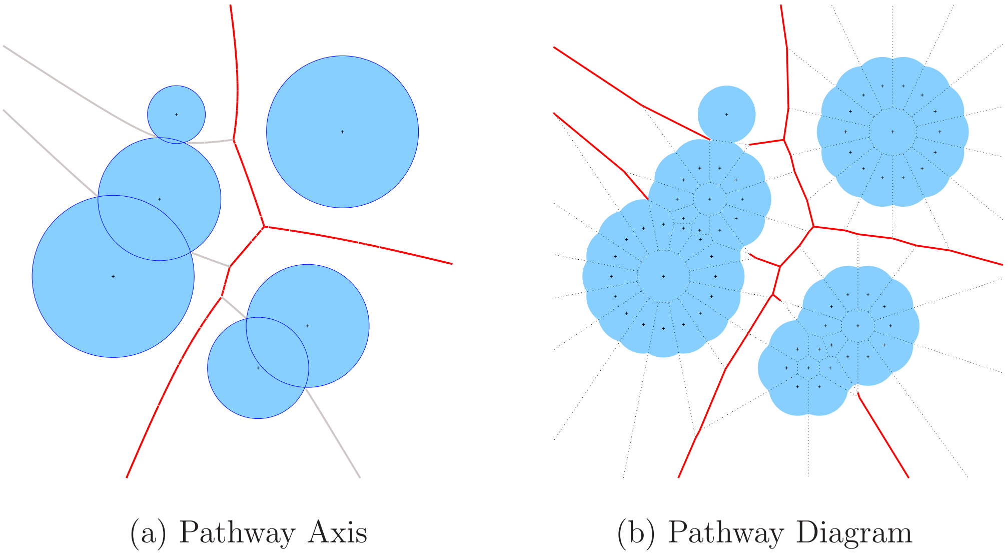
of the object [2, 14]. Oudot and Boissonnat [22] introduce an algorithm for computing the medial axis that has certified results for smooth shapes. In a recent paper Giesen *et al.* [19] approximate a useful subset of the medial axis of a shape with smooth boundary that captures the topology of the shape. However, the complement of the union of a collection of balls is not bounded by a smooth surface, making it difficult to directly apply the techniques (and hence have the topological guarantees) obtained in these papers. In contrast to the aforementioned approaches we sample a volume with balls instead of sampling a surface with points.

The λ-medial axis [9], introduced by Chazal and Lieutier, is a subset of the medial axis, that for some "regular" values of λ remains stable under Hausdorff distance perturbation. This leads to an algorithm [9] that constructs an approximation of the λ-medial axis of an object from a set of noisy unorganized points sampled on or close to the (not necessarily smooth) boundary surface of the object. We apply theoretical ideas introduced there to prove geometric convergence of our approximation.

Edelsbrunner *et al.* [16] define pockets as regions in the complement of a molecule with limited accessibility from the outside. They also describe an efficient algorithm that constructs pockets using the celebrated *weighted alpha shapes* [18]. Pockets are defined using a continuous growth process of the molecule — a pocket is a region in the complement that becomes a void before it completely disappears. In their growth process large atoms grow more slowly, which makes this approach less adequate for locating channels and determining their dimensions. Since in our approach we replace a set of balls of different radii with a set of balls of fixed (unit) radius the growth of the approximating balls is homogeneous, which gives a more intuitive geometric meaning to growth. This property could be useful beyond the usage described in the thesis, in conjunction with any algorithm that makes use of weighted alpha shapes.

In another work, Edelsbrunner *et al.* [17] introduce the notion of topological persistence during a growth process of the union of balls. In that work an efficient algorithm is described that classifies topological changes during the growth process as topological features or topological noise depending on their lifetime during the process. The theoretical notion of persistence was extended independently by Chazal *et al.* [10] and by Cohen-Steiner *et al.* [11]. Cohen-Steiner *et al.* [11] deal with real-valued functions on a topological space. The *persistence diagram* of the distance function from $\cup\mathcal{B}$ encodes topological characteristics of the function, giving a measure on the importance of topological features. To the best of our knowledge our approximation scheme is the first to yield a good approximation of this persistence diagram, which leads us to believe here as well that our approach is applicable more widely.

In this thesis we describe an approximation scheme that approximates $\mathcal{B}$ by a collection of unit balls $\mathcal{K}_\varepsilon$ such that the Hausdorff distance between $\cup\mathcal{B}$ and $\cup\mathcal{K}_\varepsilon$ is not larger than a prescribed ε. We focus on a subset of the Voronoi diagram of the centers of $\mathcal{K}_\varepsilon$ which we call the *pathway diagram* and show how to compute it. The *pathway axis* of $\mathcal{B}$ (defined in Chapter 2) is a core subset of the medial axis of the complement of $\cup\mathcal{B}$, which is composed of all points for which the set of closest balls in $\mathcal{B}$ do not have a common intersection. Informally, it can be seen as a subset of the medial axis that does not include 'dead ends',


(a) Pathway Axis

(b) Pathway Diagram

Figure 1.1: An example of a two-dimensional pathway axis and pathway diagram. (a) The input circles are colored light blue, and their pathway axis is colored red. The discarded parts of the medial axis are colored gray. (b) An example of a collection of same-size circles $\mathcal{K}_\varepsilon$ (light blue) and the pathway diagram of their centers (red). The discarded portions of the Voronoi diagram of their centers is depicted using dotted lines.

which makes it sufficient for identifying pathways. We prove that the pathway diagram contains an approximation of the pathway axis. See Figure 1.1 for a two-dimensional illustration of the pathway axis of a collection of circles and a pathway diagram of these circles. We provide a bound on the number of balls in $\mathcal{K}_\varepsilon$ as a function of ε and the ratio between the largest and the smallest ball in $\mathcal{B}$. We prove that the pathway diagram is either close to the medial axis of the complement of $\cup\mathcal{B}$ or to the boundary surface of $\cup\mathcal{B}$.

We present MolAxis, a new tool designed for the efficient identification of molecular channels. MolAxis makes use of the pathway diagram in order to construct corridors. To the best of our knowledge it is the first attempt to approximate and analyze a subset of the medial axis of the complement of a molecule in order to construct channels. A major advantage of our approach is that since the medial axis is composed of two-dimensional surface patches it reduces the dimension of the problem, i.e., it transforms a three-dimensional problem to a two-dimensional problem. In order to extract the desired corridors we further reduce the dimension of the problem, transforming it to a problem on graphs, as we will see in Part III. This dimension reduction, combined with a novel sampling technique, leads to a highly efficient algorithm.

We implemented MolAxis using the CGAL library. The CGAL [1] open source project is aimed at making the large body of theoretical algorithms and data structures in computational geometry applicable in practice, while focusing on reliability and performance. This allowed us to focus on the application oriented aspects of the tool. We show together with biologists how MolAxis is applied to identify and characterize channels in two different biological settings. In the first setting, which deals with *chamber proteins*, MolAxis finds corridors leading from a source point in an inner chamber (or cavity) of a protein to its surface. MolAxis can automatically compute a source point in the center of the main chamber using topological persistence. We found that the biologically significant chambers were automatically identified with a high success rate in this manner on the datasets that we have examined. We also allow a user-specified source point. In the second setting, which deals with *transmembrane proteins*, MolAxis constructs a single channel crossing a transmembrane protein, using user-defined parameters such as the channel direction vector and a ball fully contained in the channel. MolAxis is highly efficient and therefore can be applied to huge datasets such as multiple snapshots generated in a Molecular Dynamics (MD) simulation of the motion of a protein.

One of the most commonly used tools to compute channel location and dimension is the program HOLE, developed by Smart *et al.* [24]. The HOLE method uses a Monte Carlo simulated-annealing procedure to find the best route for a ball squeezing through the channel, changing its radius as it passes. A recent tool for computing channels in the complement of a molecule is the program CAVER by Petřek *et al.* [23]. It is based on a three-dimensional grid search in the complement space of the molecule. We compare our approach with HOLE and CAVER in terms of accuracy, performance and underlying theoretical guarantees of finding the desired pathways.

MolAxis is available on the web at `http://www.cs.tau.ac.il/~eitanyaf/MOLAXIS`.

Thesis Outline

The rest of the thesis is organized as follows. In Part I we give the needed background and explain in detail our algorithm for the construction of the pathway diagram. Part II is dedicated to the theoretical analysis of our method. We prove properties of our approximation scheme and of the pathway diagram. In Part III we give implementation details and describe how the pathway diagram is used to construct corridors in the complement of molecules, which represent molecular channels.

Chapter 2

Preliminaries

Our work builds on a large body of earlier results concerning Voronoi diagrams and the medial axis. We assume some familiarity with alpha complexes [18] and the λ-medial axis [9]; however, we define them formally below. We borrow notation mainly from the work of Attali *et al.* [4] and the work of Chazal and Lieutier [9]. For any set X we denote by $\bar{X}$, X^o, ∂X, X^c and $|X|$ the closure, the interior, the boundary, the complement and the cardinality of X respectively. $B(x, r)$, $B^o(x, r)$ and $S(x, r)$ denote the closed ball, open ball and sphere of center x and radius r in $\mathbb{R}^d$ respectively. We denote the Euclidean distance between two points $x, y \in \mathbb{R}^d$ by $d(x, y)$. The distance between two subsets A, B of $\mathbb{R}^d$ is defined to be $d(A, B) = \inf_{a \in A, b \in B} d(a, b)$.

2.1 Hausdorff Distance

The *one-sided Hausdorff distance* between two compact subsets A and B of $\mathbb{R}^d$ is:

$$d_H(A|B) = \sup_{x \in A} d(x, B) \ .$$

The *(symmetric) Hausdorff distance* between two compact subsets A and B of $\mathbb{R}^d$ is the maximum of the two one-sided distances, namely $d_H(A, B) = \max(d_H(A|B), d_H(B|A))$. We say that A is a *Hausdorff approximation* of B with an *approximation resolution* of ε if the Hausdorff distance between A and B is not larger than ε. In such a case, we will say for short that A is an ε-approximation of B.

2.2 Medial Axis

Let $\mathcal{O}$ be a bounded open subset of $\mathbb{R}^d$. For any point $x \in \mathcal{O}$, we denote by $\Gamma_{\mathcal{O}}(x)$ the set of closest points to x in the complement $\mathcal{O}^c$, namely $\Gamma_{\mathcal{O}}(x) = \{y \in \mathcal{O}^c : d(x, y) = d(x, \mathcal{O}^c)\}$. The *Medial Axis* $M[\mathcal{O}]$ of the open set $\mathcal{O}$ is the set of points $x \in \mathcal{O}$ that have at least two closest boundary points:

$$M[\mathcal{O}] = \{x \in \mathcal{O} : |\Gamma_{\mathcal{O}}(x)| \geq 2\} \; .$$

We say that a ball B is *empty* in $\mathcal{O}$ if its interior B^o is contained in $\mathcal{O}$. B is *maximal* (or *medial*) in $\mathcal{O}$ if it is empty and not contained in any other empty ball. An alternative definition of the medial axis of $\mathcal{O}$ is the union of the centers of all maximal balls in $\mathcal{O}$.

2.3 Voronoi Diagram and Delaunay Complex

From this point on we restrict ourselves in the thesis to $\mathbb{R}^3$. Let E be a finite point set in $\mathbb{R}^3$. We define the *Voronoi cell* of $e \in E$ to be $V_e = \{x \in R^3 : \forall e' \in E, d(e,x) \leq d(e',x)\}$. In words, V_e is the set of points in $\mathbb{R}^3$ that are at least as close to e as they are to any other point of E. For a subset $T \subseteq E$ we define the *Voronoi face* of T to be $V_T = \bigcap_{e \in T} V_e$. The *Voronoi diagram* of E [6] is the collection of Voronoi cells:

$$\mathcal{V}[E] = \{V_T : \emptyset \neq T \subseteq E\} \; .$$

For $0 \leq k+1 \leq 3$, a *k-simplex* σ in $\mathbb{R}^3$ is the convex hull of $k+1$ affinely independent points. The convex hull of any $0 \leq l+1 \leq k+1$ of these points is an l-simplex and a *face* of σ. Note that $\emptyset$ is the only (-1)-simplex and it is contained in any simplex. A *simplicial complex* is a collection C of simplices that satisfy the following two conditions.

(1) If $\sigma \in C$ and σ' is a face of σ then $\sigma' \in C$.

(2) If σ_1, σ_2 are in C then $\sigma_1 \cap \sigma_2$ is a face of both.

A subset $C' \subset C$ is a *sub-complex* of C if it is a simplicial complex itself, that is, it satisfies the first condition (since the second condition is trivially satisfied). For each Voronoi cell $V_T \neq \emptyset$ we define the dual Delaunay simplex σ_T to be the convex hull of the points of T. The *Delaunay complex* is the collection of the Delaunay simplices:

$$\mathcal{D}[E] = \{\sigma_T : \emptyset \neq V_T \subseteq \mathcal{V}[E]\} \; .$$

Voronoi diagrams and Delaunay complexes are among the most extensively studied tools in computational geometry, used to solve numerous, and surprisingly different, problems; see, e.g. the survey by Aurenhammer and Klein [6].

Chapter 3

Constructing the Pathway Diagram

In this chapter we define and explain what is the pathway diagram of a collection of points in $\mathbb{R}^3$ and give a formal description of our algorithm. The algorithm is fairly simple and it proceeds in two steps. First, we construct a collection $\mathcal{K}_\varepsilon$ of unit balls that constitute an ε-approximation of $\mathcal{B}$ under the Hausdorff metric. In a second step we construct the pathway diagram of the centers of $\mathcal{K}_\varepsilon$, which we denote by $\mathcal{P}_\varepsilon$. We defer technical implementation details to Chapter 7 in Part III. The properties that make $\mathcal{K}_\varepsilon$ and the pathway diagram useful for our purposes are presented and proved in Part II.

3.1 Pathway Diagram

Let E be a finite point set in $\mathbb{R}^3$ and let σ_T be a Delaunay simplex of $\mathcal{D}[E]$. Let R_T denote the radius of the smallest ball that contains all points of T on its boundary surface. We say that the simplex σ_T is *α-exposed* if $\alpha > R_T$ [18]. The collection of α-exposed simplices is a simplicial complex, which is called the *α-complex* of E. We call the collection of the dual Voronoi faces of simplices that are *not* in the α-complex the *α-Voronoi graph* of E (see Figure 3.1 for a two-dimensional illustration). Note that a simplex σ_T is α-exposed if and only if its dual Voronoi face V_T and the set of balls centered at the points of T with radius α, all have a non-empty intersection.

The $(\alpha = 1)$-Voronoi graph of E will play an important role in the thesis, and we shall refer to it as the *pathway diagram* of E. Denoting by $\mathcal{K}(E)$ the collection of unit balls centered at points of E, we can define the pathway diagram of E in a more intuitive manner. It is the set of Voronoi faces in $\mathcal{V}[E]$ that do not intersect $\cup\mathcal{K}(E)$. It is a subset of the medial axis of the complement of $\cup K(E)$ and it contains only flat facets, i.e., patches of planes bounded by simple polygons. Actually, the only difference between the pathway diagram of E and the whole medial axis of the complement of $\cup\mathcal{K}(E)$ is that the medial axis also contains parts of planes bounded by arcs whenever the medial axis reaches the boundary surface of $\cup\mathcal{K}(E)$. The pathway diagram is thus defined such that it is completely piecewise linear and easy to compute, avoiding the need to construct complicated facets that are bounded by arcs.

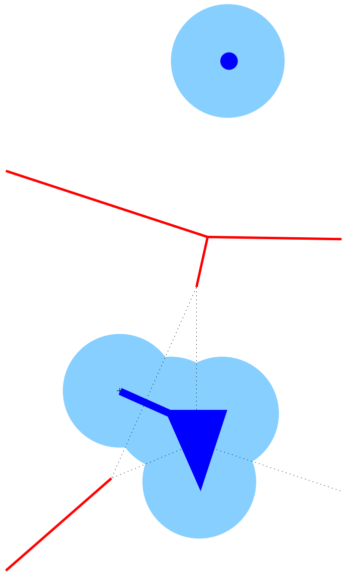


Figure 3.1: The α-complex of a collection E of five points is colored blue. The α-Voronoi graph of E is colored red. For clarity we draw in light blue circles with radius α centered at E. Note that a simplex is part of the α-complex if and only if its dual Voronoi face and the set of circles that are centered on its vertices have a non-empty intersection.

Definition 3.1 (ε-pathway diagram) *Let X be a closed bounded subset of $\mathbb{R}^3$ and let E be a finite point set. If $\cup\mathcal{K}(E)$ is an ε-approximation of X we call the pathway diagram of E an ε-*pathway diagram *of X.*

3.2 Ball ε-Sample and the ε-Flower

Let X be as above, a closed bounded subset of $\mathbb{R}^3$. We call a point on ∂X a *sample point*. We say that a finite point set E is a *point sample* of an object $X \subseteq \mathbb{R}^3$ if E is contained in ∂X. The set E is a *point ε-sample* of X if it is a point sample of X and $d_H(\partial X|E) \leq \varepsilon$. We extend the ε-sample concept from points to balls. We call a ball $B(x,r)$ a *sample ball* of X if it is contained in X and the distance of its center to the boundary of X is equal to its radius, namely $d(x,\partial X) = r$. A set K of balls is a *ball sample* of X if all balls in K are sample balls.

Definition 3.2 (ball ε-sample) *Given a set K of closed balls, a body $X \subset \mathbb{R}^3$ and a real parameter $\varepsilon > 0$ we say that K is a ball ε-sample of X if K is a ball sample and $d_H(\partial X|\bigcup K) \leq \varepsilon$.*

Note that if the balls in K have radius 0, then the definition of the ball sample coincides with the definition of the point set sample. For the next definition, recall that a *spherical shell* is the set difference between two concentric balls of different radii.

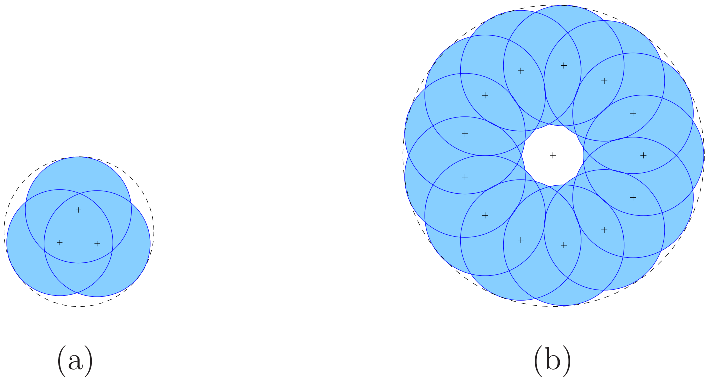


Figure 3.2: (a) An ε-flower (in the plane) without a void inside. (b) An ε-flower with a void inside. In both cases, the dashed line bounds the circle that is being approximated.

Definition 3.3 (ε-flower) *Let $B = B(x, r)$ be a ball and $\varepsilon \geq 0$ be real a parameter. A set K of closed* unit *balls are called an ε-flower of B if they constitute a ball ε-sample of B and $\cup K$ is either homeomorphic to a ball or homeomorphic to a spherical shell.*

3.3 The Algorithm

The purpose of the algorithm is to construct an ε-pathway diagram of $\mathcal{B}$, and an ε-approximation of $\cup\mathcal{B}$ (note that we approximate all of $\cup\mathcal{B}$ and not only its boundary surface). The input is a collection $\mathcal{B}$ of three-dimensional balls such that every ball in $\mathcal{B}$ is not smaller than a unit ball, and a real parameter $\varepsilon < 1/2$. The Algorithm constructs a collection E_ε of points such that the collection $\mathcal{K}_\varepsilon$ of unit balls centered at the points of E_ε constitute an ε-approximation of $\cup\mathcal{B}$. The output of the algorithm is both $\mathcal{K}_\varepsilon$ and the pathway diagram P_ε of E_ε.

In the pseudocode below (Algorithm 1) the procedure DUAL(s) returns the dual Voronoi face of a simplex s, and FLOWER(B, ε) is a procedure that returns the centers of an ε-flower of a ball B. In Chapter 7 we describe our implementation of the procedure FLOWER(B, ε), and explain how we construct, with little effort, the pathway diagram of a point set using the CGAL library.

Remark. Note that for every ball $B \in \mathcal{B}$ the algorithm adds to $\mathcal{K}_\varepsilon$ a set of unit balls that are not necessarily all sample balls of B — some of them are completely inside B. These surplus balls are added to ensure that $\mathcal{K}_\varepsilon$ is a Hausdorff approximation of $\mathcal{B}$ and exist only if the radius of B is larger than two. The number of surplus balls in our applications is small as we shall see in Part III.

Algorithm 1 Pathway diagram construction

```
Input:   A collection $\mathcal{B}$ of balls each not smaller than a unit ball.
Output:  (1) A collection $\mathcal{K}_\varepsilon$ of unit balls such that $\cup\mathcal{K}$ is an $\varepsilon$-approximation of $\cup\mathcal{B}$.
         (2) An $\varepsilon$-pathway diagram of $\mathcal{B}$.

$E \Leftarrow \emptyset$, $\mathcal{P}_\varepsilon \Leftarrow \emptyset$
for all $B = B(x, r) \in \mathcal{B}$ do
  $r' \Leftarrow r$
  while $r' > 0$ do
    $E \Leftarrow E \cup \text{FLOWER}(B(x, r'), \varepsilon)$
    $r' \Leftarrow r' - 1$
  end while
end for
$\mathcal{K}_\varepsilon \Leftarrow \mathcal{K}(E)$
$\mathcal{D}[E] \Leftarrow$ Delaunay triangulation of $E$
for all $s \in \mathcal{D}[E]$ do
  if $s$ is not 1-exposed then
    $\mathcal{P}_\varepsilon \Leftarrow \mathcal{P}_\varepsilon \cup \text{DUAL}(s)$
  end if
end for
return $\mathcal{K}_\varepsilon$, $\mathcal{P}_\varepsilon$
```

Part II

Properties of the Approximation

Chapter 4

Geometric Properties of the Pathway Diagram

The algorithm described in Part I for constructing the pathway diagram is fairly simple. As we will see in Part III, the pathway diagram serves our practical goal of identifying channels in molecules very well. But can we provide any theoretical guarantees on how well does the pathway diagram approximate the pathway axis? This turns out to be a non-trivial question, which we address in this chapter.

In Section 4.2 we formally define the *pathway axis* as a subset of the medial axis of the complement of $\cup\mathcal{B}$, and prove that the pathway diagram contains an approximation of the pathway axis. In order to prove this property we use the λ-medial axis (defined in Section 4.1) as a mediator, i.e., we show that the pathway diagram contains an approximation of the λ-medial axis, and that the λ-medial axis contains the pathway axis.

The *clearance* of a point p in the pathway diagram $\mathcal{P}_\varepsilon$ is the minimal distance between p and $\cup\mathcal{K}_\varepsilon$. The *exact clearance* of a point p in $(\cup\mathcal{B})^c$ is the minimal distance between p and $\cup\mathcal{B}$. In Section 4.3 we prove that the clearance function can serve as a good approximation of the exact clearance function. We rely on this property in Part III, where we are interested in approximating three-dimensional curves in $(\cup\mathcal{B})^c$ that balance between exact clearance and length.

4.1 Preliminaries

λ-Medial Axis

Let $\mathcal{O}$ be a bounded open subset of $\mathbb{R}^d$. The strictly positive, real valued function $\mathcal{R}_\mathcal{O}$ defined on $\mathcal{O}$ is the distance to the boundary:

$$\mathcal{R}_\mathcal{O}(x) = d(x, \mathcal{O}^c) \ .$$

Recall that $\Gamma_\mathcal{O}(x)$ is the set of closest points to x in the complement $\mathcal{O}^c$. There always exists a unique closed ball with minimal radius enclosing $\Gamma_\mathcal{O}(x)$ [9]. The real valued, positive function $\mathcal{F}$ is defined as the radius of this smallest closed ball enclosing $\Gamma_{\mathcal{O}}(x)$, or formally:

$$\mathcal{F}(x) = \inf\{r : \exists y \in \mathbb{R}^d, B(y,r) \supset \Gamma_{\mathcal{O}}(x)\} \ .$$

We denote by $\Theta(x)$ the center of this smallest enclosing ball. Of course, when $x \notin M[\mathcal{O}]$, we have $\Gamma_{\mathcal{O}}(x) = \{\Theta(x)\}$ and $\mathcal{F}(x) = 0$. Given a real $\lambda \geq 0$ the λ-*Medial Axis* is defined to be:

$$\mathrm{M}_\lambda[\mathcal{O}] = \{x \in \mathcal{O} : \mathcal{F}(x) \geq \lambda\} \ .$$

We say that λ is a *regular* value of $\mathcal{O}$ if the function that maps $\nu \in \mathbb{R}$ to $\mathrm{M}_\nu[\mathcal{O}]$ in $\mathbb{R}^d$ is continuous under the Hausdorff metric at $\nu = \lambda$. Formally, λ is a regular value of $\mathcal{O}$ if for every $\delta > 0$ there exists a $\psi > 0$ such that for any $\nu > 0$ that satisfies $|\lambda - \nu| < \psi$ it holds that $d_H(M_\lambda[\mathcal{O}], M_\nu[\mathcal{O}]) < \delta$. It is shown in [9] that if λ is a regular value of a shape $\mathcal{O}$, then the λ-medial axis transform is continuous at $\mathcal{O}$ for the Hausdorff distance:

Theorem 4.1 (Chazal and Lieutier [9]) *Let $\mathcal{O}$ be a bounded open subset of R^d and λ be a regular value of $\mathcal{O}$. For every $\delta > 0$, there exists $\mu > 0$ such that for every open subset $\tilde{\mathcal{O}}$ of $\mathbb{R}^d$,*

$$d_H(\mathcal{O}^c, \tilde{\mathcal{O}}^c) \leq \mu \implies d_H(M_\lambda[\mathcal{O}], M_\lambda[\tilde{\mathcal{O}}]) \leq \delta \ .$$

In other words, if λ is a regular value of $\mathcal{O}$ then the λ-medial axis of $\mathcal{O}$ is stable under the Hausdorff metric. Next we examine the λ-medial axis of the complement of a finite point set.

The λ-medial axis is formally defined on *bounded* open subsets of $\mathbb{R}^3$, yet the complement of a closed bounded set like $\cup\mathcal{B}$ or a finite point set is not bounded. To resolve this technicality we limit ourselves from now on to a large open ball $Q = B(c_q, r_q)$ that contains $\cup\mathcal{B}$. For example, when we refer to the complement of a closed set $C \subset Q$ we mean the intersection of C^c and Q, i.e., $Q \setminus C$.

λ-Voronoi Graph

Let E be a finite point set and let σ_T be a Delaunay simplex of the Delaunay complex $\mathcal{D}[E]$. We say that σ_T is λ-*enclosed* if the points of T can be enclosed in a sphere of radius not greater than λ. The collection of λ-enclosed simplices is a simplicial complex, which is named the λ-*complex*. Note that a simplex σ_T is λ-enclosed if and only if the set of balls centered at T with radius λ have a non-empty intersection. We call the dual of its complement, namely the collection of the dual Voronoi faces of simplices *not* in the λ-complex, the λ-*Voronoi graph* (see Figure 4.1 for a two-dimensional illustration).

By definition, the λ-Voronoi graph of E and the λ-medial axis of E^c are the same. Therefore, we regard the λ-medial axis as an extension of the λ-Voronoi graph from the complement of finite point sets to general open bounded subsets. When dealing with a finite point set we will interchange between both terms.

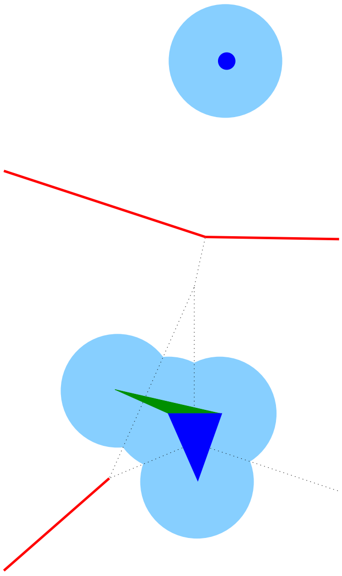


Figure 4.1: The λ-complex of a collection E of five points is colored blue and green. The λ-Voronoi graph of E is colored red. For clarity we draw circles with radius λ centered at E in light blue. Note that a simplex is part of the λ-complex if and only if the circles that are centered at its vertices have a non-empty intersection. Note that the green triangle is λ-enclosed but not α-exposed, for $\alpha = \lambda$, (compare with Figure 3.1) because even though the circles centered at its three vertices share a common intersection, the intersection is disjoint from the dual Voronoi face of the triangle (which is required for a simplex to be α-exposed).

Approximating the λ-Medial Axis

Let $\mathcal{O}$ be a bounded open subset of $\mathbb{R}^d$. We denote the boundary of $\mathcal{O}$ by $S = \partial\mathcal{O} = \bar{\mathcal{O}} \cap \mathcal{O}^c$. The λ-medial axis of S^c is divided into an *inner* and *outer* λ-medial axis. The inner (resp. outer) λ-medial axis is contained in $\mathcal{O}$ (resp. $\mathcal{O}^c$). A finite point set E is called a *μ-noisy sample*[1] of S if the Hausdorff distance between S and E is less than μ. Chazal and Lieutier provide [9] an algorithm for approximating the λ-medial axis of S^c from the λ-Voronoi Graph of a μ-noisy point sample of S, which is based on Theorem 4.1.

λ-Voronoi graph and α-Voronoi graph

Recall that the pathway diagram of E is the $(\alpha = 1)$-Voronoi graph of E. In contrast to the λ-Voronoi graph, the α-Voronoi graph is defined only for finite point sets and cannot be extended to general open subsets of $\mathbb{R}^3$, such as the complement of $\cup\mathcal{B}$. Therefore, to state properties of the pathway diagram we make use of a relation between the λ-Voronoi graph and the α-Voronoi graph for $\alpha = \lambda$, which we discuss next.

For $\alpha = \lambda$, if a simplex $\sigma_T \in \mathcal{D}[E]$ is α-exposed (see definition in Section 3.1) it is

[1] We use 'μ' instead of the natural choice 'ε' since in this thesis 'ε' denotes the Hausdorff approximation quality of $\mathcal{K}_\varepsilon$.

necessarily λ-enclosed but not vice versa. This is because σ_T is λ-enclosed if the set of balls centered at T with radius λ have a non-empty intersection X_T, which is a weaker predicate than the α-exposed predicate which requires that the dual Voronoi face V_T intersects X_T as well ($(X_T \cap V_T) \neq \emptyset$). *Thus the α-Voronoi graph of E contains the λ-Voronoi graph of E for $\alpha = \lambda$.* The lemma below states this and more; denoting by $LVG_\lambda[E]$ the λ-Voronoi graph of E we restate a lemma given in [4], using our terminology and notation:

Lemma 4.2 [4] *For any finite point set E, the pathway diagram of E contains $LVG_1[E]$ and the two are homotopy equivalent.*

4.2 Pathway Axis

We define a function $\mathcal{O}(X)$ that maps a closed bounded subset $X \subseteq Q$ to $Q \setminus X$, which is an open subset of Q. For each point $x \in M[\mathcal{O}(\cup\mathcal{B})]$ we define $I_\mathcal{B}(x)$ to be the set of balls in $\mathcal{B}$ closest to x:

$$I_\mathcal{B}(x) = \{B \in \mathcal{B} : \Gamma_{\mathcal{O}(\cup\mathcal{B})}(x) \cap B \neq \emptyset\} .$$

We define the *pathway axis* of $\mathcal{B}$, denoted by $PA[\mathcal{B}]$, to be the subset of $M[\mathcal{O}(\cup\mathcal{B})]$ for which the balls of $I_\mathcal{B}(x)$ do not share a common point:

$$PA[\mathcal{B}] = \{x \in M[\mathcal{O}(\cup\mathcal{B})] : \bigcap_{B \in I_\mathcal{B}(x)} B = \emptyset\} .$$

We consider the pathway axis sufficient for finding pathways since it is a subset of the medial axis of $\mathcal{O}(\cup\mathcal{B})$ without ‘dead ends’. The following theorem states that the pathway diagram that our algorithm constructs contains an approximation of the pathway axis of $\mathcal{B}$. A collection of balls is said to be in *general position* if no degeneracies occur, namely the common intersection of the boundary spheres of any two, three or four balls in the collection is not a single point.

Theorem 4.3 *Let $\mathcal{B}$ be a finite collection of balls in general position that are each not smaller than a unit ball. For any $\delta > 0$ there exists an $\varepsilon > 0$ such that the pathway diagram P_ε, which Algorithm 1 constructs satisfies $d_H(PA[\mathcal{B}] \mid \mathcal{P}_\varepsilon) < \delta$.*

In order to prove this theorem we first prove several auxiliary claims. Let B_u be the canonical unit ball, i.e., a ball with radius 1 that is centered at the origin. For each ball $B = B(c, r) \in \mathcal{B}$ we define a concentric ball $B_H = B(c, r - 1)$. Let $\mathcal{H}$ denote the collection of balls $\{B_H\}_{B \in \mathcal{B}}$, which we call the *offset balls* of $\mathcal{B}$. Recall that the Minkowski sum of two sets X, Y is $X \oplus Y = \{x + y : x \in X, y \in Y\}$. From the definition of $\mathcal{H}$ and Q we note the following:

Observation 4.4 *$\mathcal{O}(\cup\mathcal{H}) = \mathcal{O}(\cup\mathcal{B}) \oplus B_u$ and $\cup\mathcal{B} = \cup\mathcal{H} \oplus B_u$.*

The first Lemma will allow us to shift our focus from the pathway axis to the λ-medial axis, enabling us to make use of the stability property of the λ-medial axis, as expressed in Theorem 4.1.

Lemma 4.5 *The pathway axis of $\mathcal{B}$, $PA[\mathcal{B}]$, is contained in the $(\lambda = 1)$-medial axis of $\mathcal{O}(\cup\mathcal{H})$, $M_1[\mathcal{O}(\cup\mathcal{H})]$.*

Proof: Let x be a point in $PA[\mathcal{B}]$. It follows that $x \in M[\mathcal{O}(\cup\mathcal{B})]$. The ball $B(x, \mathcal{R}_{\mathcal{O}(\cup\mathcal{B})})$ is a medial ball in $\mathcal{O}(\cup\mathcal{B})$ and thus the ball $B(x, \mathcal{R}_{\mathcal{O}(\cup\mathcal{B})} + 1)$ is a medial ball in $\mathcal{O}(\cup\mathcal{H})$ by Observation 4.4, or in other words $x \in M[\mathcal{O}(\cup\mathcal{H})]$. Let us consider the point set $\Gamma = \Gamma_{\mathcal{O}(\cup\mathcal{H})}(x)$ of closest points to x in $\cup\mathcal{H}$. Assume that $x \notin \mathrm{M}_1[\mathcal{O}(\cup\mathcal{H})]$ and therefore there exists a ball $C = B(c, r)$ with radius not larger than 1 such that $\Gamma \subseteq C$. For each $y \in \Gamma$ it holds that $d(c, y) \leq 1$, and therefore for each $B \in I_{\mathcal{B}}(x)$ it holds that $c \in B$. We get that $c \in \bigcap_{B \in I_{\mathcal{B}}(x)} B$ which means that $x \notin PA[\mathcal{B}]$ in contradiction. It follows that $x \in \mathrm{M}_1[\mathcal{O}(\cup\mathcal{H})]$. □

Recall that E_ε is the finite point set constructed by Algorithm 1. By examining the algorithm we conclude that the point set E_ε satisfies:

(1) Each point p in E_ε lies either on the boundary surface of $\cup\mathcal{H}$ or inside $\cup\mathcal{H}$.

(2) For each point p on the boundary surface of $\cup\mathcal{B}$, it holds that $d(p, \cup\mathcal{K}_\varepsilon) < \varepsilon$.

(3) For each point p in $\cup\mathcal{H}$, it holds that $d(p, E_\varepsilon) < 1$.

We wish to approximate the $(\lambda = 1)$-medial axis of $\mathcal{O}(\cup\mathcal{H})$ using the $(\lambda = 1)$-medial axis of the complement of E_ε. Yet the complement of $\mathcal{O}(\cup\mathcal{H})$ and E_ε are not as close as the conditions of Theorem 4.1 require. We use an intermediate set $\tilde{\mathcal{O}}$ that on the one hand is close to $\mathcal{O}(\cup\mathcal{H})$ and on the other hand has the same λ-medial axis as the complement of E_ε. The lemma below defines $\tilde{\mathcal{O}}$ and proves the latter condition. It is a variant of Lemma 5.2 of Chazal and Lieutier [9]. Given an open subset $\mathcal{O} \subset \mathbb{R}^d$ such that $\mathcal{O}^c$ is bounded, the lemma defines $\tilde{\mathcal{O}}$ and shows that under certain conditions $\tilde{\mathcal{O}}$ and the complement of a finite point set E have the same λ-medial axis. For any bounded open subset X of $\mathbb{R}^3$ we let $X^{+\mu}$ denote the set $\{p \in \mathbb{R}^3 : d(p, X) < \mu\}$. See Figure 4.2 for an illustration.

Lemma 4.6 *Let λ and μ be two real positive numbers such that $\lambda > 2\mu$, and let $\mathcal{O}$ be an open subset of $\mathbb{R}^d$ such that the closed set $C = \mathcal{O}^c$ is bounded. Assume that the finite point set E satisfies:*

(1) $E \subset C$,

(2) $d_H(\partial C \mid E) < \mu$ and

(3) $d_H(C \mid E) < \lambda$.

The λ-medial axis of $\tilde{\mathcal{O}} = \mathcal{O}^{+\mu} \setminus E$ is equal to the λ-Voronoi graph of E.

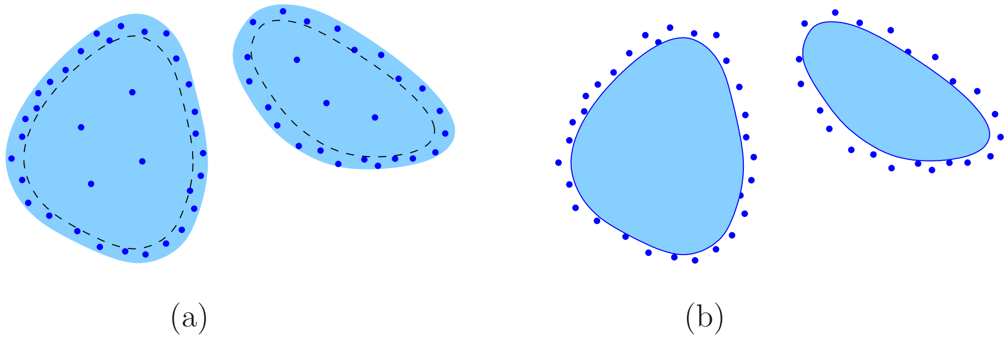


Figure 4.2: Illustration for Lemma 4.6. (a) The closed set C is colored light blue. A finite point set E is depicted using small blue dots (discs). (b) The complement of $\mathcal{O}$. It is composed of a subset of E depicted with dots, and the complement of $(C^c)^{+\mu}$ in light blue. If the conditions of the lemma are satisfied then the λ-medial axis of $\mathcal{O}$ is equal to the λ-Voronoi graph of E.

Proof:

Let x be a point of $M_\lambda[\tilde{\mathcal{O}}]$. Suppose there exists a $z \in (\mathcal{O}^{+\mu})^c$ such that $d(x,z) = R_{\tilde{\mathcal{O}}}(x) > \mu$. Therefore the ball $B(x, R_{\tilde{\mathcal{O}}}(x))$ is contained in $(\mathcal{O}^{+\mu})^c$ and does not contain any point of E. Since x belongs to $\mathcal{O}$ and z belongs to $(\mathcal{O}^{+\mu})^c$, there exists a point y on the segment $[x,z]$ such that $y \in \partial\mathcal{O}$. The ball $B(z,\mu)$ does not intersect $\mathcal{O}$, so $d(z,y) \geq \mu$ and hence $B(y,\mu) \subset B(x, R_{\tilde{\mathcal{O}}}(x))$. But $B(x, R_{\tilde{\mathcal{O}}}(x)) \cap E = \emptyset$ and $d(y,E) < \mu$ because $d_H(\partial\mathcal{O}, E) < \mu$, which is a contradiction. Thus we get that $\Gamma_{\tilde{\mathcal{O}}}(x) \subset E$, which means that x belongs to the λ-Voronoi graph of E. This proves one direction.

Let x be a point of the λ-Voronoi graph of E. We know that $x \in \mathcal{O}^{+\mu}$ since $d_H(C \mid E) < \lambda$. If x is in $\mathcal{O}^{+\mu} \setminus \mathcal{O}$ there exists a point $p \in \partial C$ such that $d(x,p) < \mu$ and a point $e \in E$ such that $d(p,e) < \mu$. Therefore $d(x,e) < 2\mu < \lambda$, in contradiction to the fact that x is a point of the λ-Voronoi graph of E. This means that x is in $\mathcal{O}$. Suppose there exists a point $z \in (\mathcal{O}^{+\mu})^c$ such that $d(x,z) = R_{\tilde{\mathcal{O}}}(x)$. In a manner similar to the proof of the first direction we reach a contradiction and conclude that $\Gamma_{\tilde{\mathcal{O}}}(x) \subset E$. Therefore $x \in M_\lambda[\tilde{\mathcal{O}}]$, which completes the proof of the lemma.

□

The last thing we need before proving Theorem 4.3 is to make sure $\lambda = 1$ is a regular value of $\mathcal{O}(\cup\mathcal{H})$. We state the following observation, regarding the regular values of $\mathcal{O}(\cup\mathcal{H})$, without proof.

Observation 4.7 *If the balls of $\mathcal{B}$ are in general position then $\lambda = 1$ is a regular value of $\mathcal{O}(\cup\mathcal{H})$.*

Proof (Theorem 4.3):

Let $\delta > 0$ be a real number. Since the balls in $\mathcal{B}$ are in general position we can apply Theorem 4.1 to $\mathcal{O}(\cup\mathcal{H})$, thus there exists a $1/2 > \mu > 0$ such that for every open set

$\mathcal{O}$, if $d_H(\mathcal{O}(\cup\mathcal{H})^c, \mathcal{O}^c) < \mu$ then $d_H(M_1[\mathcal{O}(\cup\mathcal{H})], M_1[\mathcal{O}]) < \delta$. We choose $\varepsilon = \frac{2}{5}\mu^2$, and according to Lemma 5.5 of Chapter 5, E_ε contains a μ-noisy sample of the boundary surface of $\cup\mathcal{H}$. Since $d_H(\cup\mathcal{K}_\varepsilon, \cup\mathcal{B}) < \varepsilon < 1$ we also know that $d_H(\cup\mathcal{H} \mid E_\varepsilon) < 1$. We define the open subset $\tilde{\mathcal{O}} = ((\cup\mathcal{H})^c)^{+\mu} \setminus E_\varepsilon$ as in Lemma 4.6. We can apply Lemma 4.6, thus the $(\lambda = 1)$-medial axis of $\tilde{\mathcal{O}}$ (denoted by $M_1[\tilde{\mathcal{O}}]$) is equal to the $(\lambda = 1)$-Voronoi graph of E_ε (denoted by $\mathit{LVG}_1[E_\varepsilon]$). From the definition of $\tilde{\mathcal{O}}$ we know that $d_H(\mathcal{O}(\cup\mathcal{H})^c, \tilde{\mathcal{O}}^c) < \mu$, thus $d_H(M_1[\mathcal{O}(\cup\mathcal{H})] \mid M_1[\tilde{\mathcal{O}}]) < \delta$. To complete the proof of the theorem we need to note two things:

(1) $M_1[\tilde{\mathcal{O}}] = \mathit{LVG}_1[E_\varepsilon] \subseteq \mathcal{P}_\varepsilon$ according to Lemma 4.2

(2) $\mathit{PA}[\mathcal{B}] \subseteq \mathrm{M}_1[\mathcal{O}(\cup\mathcal{H})]$ according to Lemma 4.5.

□

Remark. The algorithm in [9] does not distinguish between the inner or outer part of the λ-medial axis. To distinguish the desired portion (inner or outer) an extra postprocessing stage us required. In our scenario the open set $\mathcal{O}$ is the complement of $\cup\mathcal{H}$. The point sample E_ε that our algorithm constructs contains a μ-sample of the boundary surface of $\cup\mathcal{H}$ for some μ. Yet in addition to these points E_ε contains surplus points that lie completely inside $\cup\mathcal{H}$ that 'cover' $\cup\mathcal{H}$, i.e., the Hausdorff distance between $\cup\mathcal{H}$ and E_ε is bounded by $\lambda = 1$. This ensures us that the outer λ-medial of $\mathcal{O}$ (which is the part that lies within $\cup\mathcal{H}$) is empty, thus we do not need any postprocessing. We proved this property of our algorithm in Lemma 4.6 above. This required a slight modification of Lemma 5.2 in [9].

4.3 Geometric Convergence

The *clearance* $c(p)$ of a point p in the pathway diagram $\mathcal{P}_\varepsilon$ is the minimal distance between p and $\cup\mathcal{K}_\varepsilon$. The *exact clearance* $\bar{c}(p)$ of a point p in $\mathcal{O}(\cup\mathcal{B})$ is the minimal distance between p and $\cup\mathcal{B}$. The lemma below guarantees that the clearance function can serve as a good approximation of the exact clearance function.

Lemma 4.8 *For any point $p \in \mathcal{P}_\varepsilon$ such that $c(p) > \varepsilon$ it holds that $|c(p) - \bar{c}(p)| \leq \varepsilon$.*

Proof:

Let p be a point in $\mathcal{P}_\varepsilon$. The ball $B_0 = B(p, c(p))$ is a medial ball of $\mathcal{O}(\mathcal{K}_\varepsilon)$ by definition. The smaller concentric ball $B_1 = B(p, c(p) - \varepsilon)$ is empty in $\mathcal{O}(\cup\mathcal{B})$, since the Hausdorff distance $d_H(\cup\mathcal{B}, \cup\mathcal{K}_\varepsilon) \leq \varepsilon$. We define the ball B' to be the maximal empty ball in $\mathcal{O}(\cup\mathcal{B})$ that is centered at p, such that $B_1 \subseteq B' \subseteq B_0$. The ball B' exists and its radius $\bar{c}(p)$ satisfies the assertions of the lemma. □

Chapter 5

Complexity of the Approximation

In this chapter we address the complexity of our approximation scheme, focusing on the number of unit balls in $\mathcal{K}_\varepsilon$. In Section 5.1 we prove an upper bound on the ratio $|\mathcal{K}_\varepsilon|/|\mathcal{B}|$. In Section 5.2 we show what is gained from using ball samples over using point samples, the latter being the standard practice (see, e.g., [19, 22]). Finally, in Section 5.3, we prove a lemma which we have already used in Section 4.2.

5.1 Upper Bound on $|\mathcal{K}_\varepsilon|/|\mathcal{B}|$

Based on the definitions of a ball sample and a point sample, presented in Section 3.2, we define the *approximation quality* of a sample. Let X be a closed bounded subset of $\mathbb{R}^3$, let E be a finite point sample of X, and let K be ball sample of X. We call the one-sided Hausdorff distance $d_H(\partial X|E)$ the *approximation quality* of E. In a similar fashion, we call the one-sided Hausdorff distance $d_H(\partial X| \cup K)$ the *approximation quality* of K. Given an integer $\kappa > 0$, and denoting by μ the approximation quality of E, we say that E is *κ-light* if the number of sample points in any ball of radius μ is not greater than κ, namely $\forall x' \in \mathbb{R}^3, |B(x', \mu) \cap E| \leq \kappa$.

The following theorem gives an upper bound on the number of unit balls needed to construct an ε-flower of a single ball $B \in \mathcal{B}$. In this section we prove the following theorem.

Theorem 5.1 *Let $B(x, r)$ be a ball with $r \geq 1$, let E be a finite point sample of $B(x, r-1)$. Let $K = K(E)$ be the collection of unit balls centered at E, which are a ball sample of $B(x, r)$. Denoting by ε the approximation quality of K, if $\varepsilon < 1/2$ and E is κ-light then:*

$$|E| \leq \kappa \frac{16r^2}{3\varepsilon^{\frac{4}{3}}} .$$

The following is a corollary of the theorem which gives a bound on the ratio $|\mathcal{K}_\varepsilon|/|\mathcal{B}|$, assuming the procedure FLOWER(B, ε) produces a κ-light sampling.

Corollary 5.2 *If for every ball $B(x,r) \in \mathcal{B}$ it holds that $r \leq \rho$, and the sampling procedure FLOWER(B,ε) produces a κ-light sampling for an integer κ then:*

$$|\mathcal{K}_\varepsilon|/|\mathcal{B}| \leq \kappa \frac{16\lceil\rho\rceil^3}{3\varepsilon^{\frac{4}{3}}} \ .$$

Proof:

Following Algorithm 1, for every ball $B(x,r) \in \mathcal{B}$ the procedure FLOWER(B,ε) is called at most $\lceil r\rceil \leq \lceil\rho\rceil$ times. Summing over all balls in $\mathcal{B}$ gives the desired result.

□

In order to prove Theorem 5.1 we first prove several auxiliary claims. The following lemma gives an upper and lower bound on the number of points in a μ-sample which is κ-light.

Lemma 5.3 *Let E be a μ-sample of a ball $B = B(x,r)$. If E is κ-light and $\mu < \min(r,1/2)$, then $16r^2/3\mu^2 \leq |E| \leq \kappa(16r^2/3\mu^2)$.*

Proof:

Let $S = S(x,r)$ be the boundary sphere of B. We denote by $A(X)$ the surface area of a bounded surface X. Since E is κ-light it holds that:

$$A(S) \leq \sum_{p\in E} A(S \cap B(p,\mu)) \leq \kappa A(S) \ .$$

It is not difficult to see that for any $p \in E$ it holds that $\frac{3}{4}\pi\mu^2 \leq A(S \cap B(p,\mu)) \leq \pi\mu^2$. Therefore we get:

$$\frac{3}{4}\pi\mu^2|E| \leq \kappa A(S) \ .$$

Since $A(S) = 4\pi r^2$ we get:

$$\frac{16r^2}{3\mu^2} \leq |E| \leq \kappa\frac{16r^2}{3\mu^2} \ .$$

□

In the next lemma we establish a relation between the approximation quality of a point sample E of $B(x,r-1)$ and the approximation quality of the ball sample $K(E)$ of $B(x,r)$:

Lemma 5.4 *Let $B = B(x,r)$ be a ball such that $r \geq 1$, and let E be a point sample of $B(x,r-1)$. Let ε denote $d_H(\partial B(x,r)| \cup K(E))$, and μ denote $d_H(\partial B(x,r-1)|E)$. If $\varepsilon < \min(r-1,1/2)$ then:*

$$\mu^2 \geq 2\varepsilon\frac{r-1}{r} \ .$$

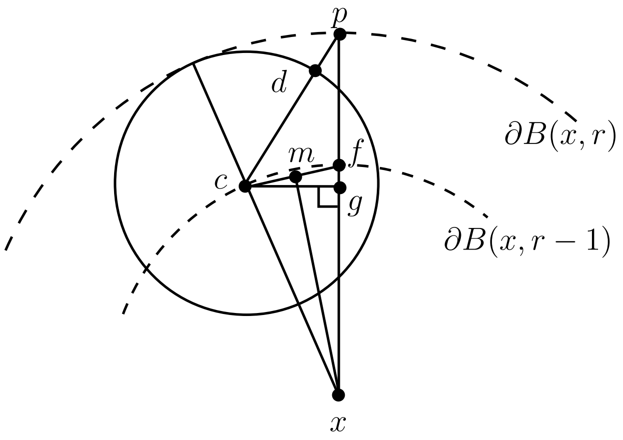


Figure 5.1: Geometric relation between μ and ε. The length of dp is ε and the length of cf bounds μ.

Proof:

Let $p \in \partial B$ be a point at distance $\varepsilon = d_H(\partial B, \cup K)$ from $\cup K$. We know p exists since B and $\cup K$ are compact and ε is the approximation quality. Let $k = B(c,1)$ be a ball in K that is ε distant from p. We denote by f the radial projection of p onto the ball $B(x, r-1)$; see Figure 5.1 for an illustration. We note that $c \in E$ is the closest point of E to f, since k is the closest ball to p. Therefore $\mu \geq d(f,c)$ giving an implicit bound on μ which we will work out below.

We use the following notation. The intersection of the segment $\overline{pc}$ with the boundary of k is marked with d, the closest point to c on the segment $\overline{px}$ is marked with g, and the midpoint of $\overline{cf}$ is marked with m. We denote the length of $\overline{cf}$ by $\bar{\mu}$. Since $\mu \geq d(f,c) = \bar{\mu}$ it suffice to prove the bound for $\bar{\mu}$. First we note that $|\overline{cx}| = r-1$, $|\overline{px}| = r$, $|\overline{pd}| = \varepsilon$, $|\overline{xf}| = r-1$, all from the definition. From triangle similarity we know that $|\overline{fg}|/|\overline{cf}| = |\overline{mf}|/|\overline{xf}|$ or in other words $|\overline{fg}| = \frac{(\bar{\mu})^2}{2(r-1)}$. Considering the triangle $\triangle cgp$ and the equality $|\overline{pg}| = |\overline{fg}| + 1$ we can express ε as a function of $\bar{\mu}$:

$$\begin{aligned}
\varepsilon &= (|\overline{pg}|^2 + |\overline{cg}|^2)^{1/2} - 1 \\
&= ((|\overline{fg}|+1)^2 + |\overline{cg}|^2)^{1/2} - 1 \\
&= (|\overline{cf}|^2 + 2|\overline{fg}| + 1)^{1/2} - 1 \\
&= (\bar{\mu}^2 + 1 + \frac{(\bar{\mu})^2}{r-1})^{1/2} - 1 \ .
\end{aligned}$$

We can now work towards expressing $\bar{\mu}$ as a function of ε:

$$\begin{aligned}
(\varepsilon+1)^2 &= \bar{\mu}^2 + 1 + \frac{(\bar{\mu})^2}{r-1} \\
\varepsilon^2 + 2\varepsilon &= (\bar{\mu})^2 \frac{r}{r-1} \ .
\end{aligned}$$

Reorganizing the terms we obtain

$$\bar{\mu}^2 = (\varepsilon^2 + 2\varepsilon)\frac{r-1}{r} \geq 2\varepsilon\frac{r-1}{r} \ .$$

Recall that $\mu \geq d(f,c)$ or in other words $\mu \geq \bar{\mu}$. The bound asserted in the lemma follows.

□

We can finally prove Theorem 5.1, harnessing the lemmas above.

Proof (Theorem 5.1):

Let E be a sample as defined in the theorem and let μ be its approximation quality. We handle three cases, according to the approximation quality ε of $K(E)$. First, the trivial case of a small ball. If $(r-1) \leq \varepsilon$ then E obviously contains a constant number of points. In the two other cases we can use Lemma 5.4, which states that $\mu^2 \geq 2\varepsilon\frac{r-1}{r}$. We also use Lemma 5.3 which states that $|E| \leq \kappa\frac{16(r-1)^2}{3\mu^2}$. In case $\varepsilon < (r-1) \leq \varepsilon^{\frac{1}{3}}$ we get:

$$\mu^2 \geq 2\varepsilon\frac{r-1}{r} \geq 2\varepsilon - \frac{2\varepsilon}{\varepsilon+1} = \frac{2\varepsilon^2}{\varepsilon+1} \geq \varepsilon^2 \ .$$

We use this bound in conjunction with Lemma 5.3 and get a bound on $|E|$:

$$|E| \leq \kappa\frac{16(r-1)^2}{3\mu^2} \leq \kappa\frac{16\varepsilon^{\frac{2}{3}}}{3\varepsilon^2} \leq \kappa\frac{16}{3\varepsilon^{\frac{4}{3}}} \leq \kappa\frac{16r^2}{3\varepsilon^{\frac{4}{3}}} \ .$$

In the remaining case, where $\varepsilon^{\frac{1}{3}} < (r-1)$ we get:

$$\begin{aligned} \mu^2 &\geq 2\varepsilon\tfrac{r-1}{r} \geq 2\varepsilon - \tfrac{2\varepsilon}{r} \geq 2\varepsilon - \tfrac{2\varepsilon}{\varepsilon^{\frac{1}{3}}+1} = \varepsilon(2 - \tfrac{2}{\varepsilon^{\frac{1}{3}}+1}) \\ \mu^2 &\geq 2\tfrac{\varepsilon^{\frac{4}{3}}}{\varepsilon^{\frac{1}{3}}+1} \geq \varepsilon^{\frac{4}{3}} \ . \end{aligned}$$

By using the above inequality together with the bound of Lemma 5.3 we get:

$$|E| \leq \kappa\frac{16(r-1)^2}{3\mu^2} \leq \kappa\frac{16r^2}{3\varepsilon^{\frac{4}{3}}} \ .$$

□

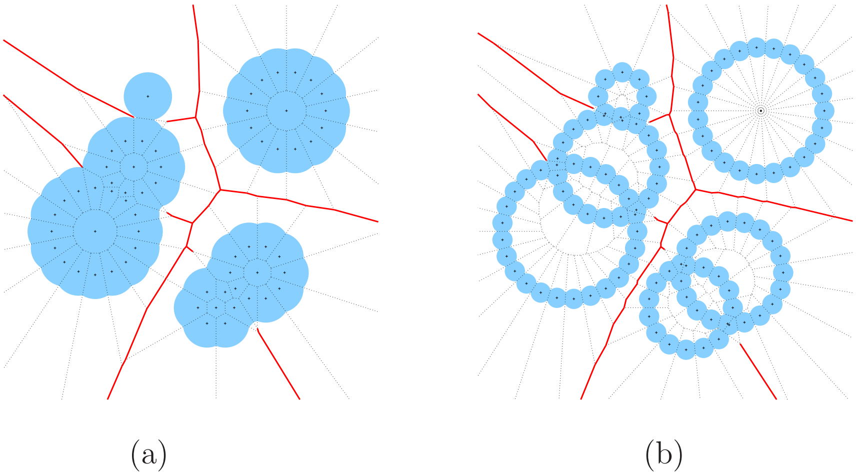


Figure 5.2: Two-dimensional illustration of a ball sample and point sample approach for approximating a subset of the medial axis of the complement of $\cup\mathcal{B}$. (a) The collection of unit balls $\mathcal{K}_\varepsilon$ are colored light blue, and the pathway diagram of their centers is colored red. (b) A point sample S_ε of the boundary surface of the balls in $\mathcal{B}$ is depicted using small crosses, and balls with radius ε that are centered at S_ε are colored light blue. The λ-Voronoi graph of S_ε with $\lambda = \varepsilon$ is colored in red, with the portion that is contained inside $\cup\mathcal{B}$ discarded.

5.2 Comparison with Point Sampling Techniques

We compare here our ball approximation with a standard approach of sampling using points. In this technique we sample the boundary surface of the balls in $\mathcal{B}$ with a collection of points S_ε, such that S_ε contains an ε-sample of the surface of $\cup\mathcal{B}$. We construct an approximation of the λ-medial axis of the complement of $\cup\mathcal{B}$ from the λ-Voronoi graph of S_ε with $\lambda = \varepsilon$, as described in [4] (see Figure 5.2). The λ-Voronoi graph of S_ε contains an irrelevant outer portion which must be discarded, as discussed in the remark at the end of Section 4.2.

Let us focus on a single ball $B \in \mathcal{B}$. The main advantage of our approach is the relatively small number of unit balls needed to construct an ε-flower of B, compared to the number of points needed to construct a point ε-sample of B. From Theorem 5.1, we know that the number N_{ball} of unit balls needed by our algorithm to constitute an ε-flower B is not more than $\kappa(16r^2/3\varepsilon^{\frac{4}{3}})$. On the other hand, the number N_{point} of points needed to constitute a point ε-sample of B is at least $(16r^2/3\varepsilon^2)$, according to Lemma 5.3. Therefore, by regarding κ as a constant, we get that the ratio $N_{\mathrm{ball}}/N_{\mathrm{point}}$ is $\Omega(\varepsilon^{2/3})$.

5.3 Upper Bound on Point Approximation Quality

Let B be a ball in $\mathcal{B}$, let E be a point sample of $B(x, r-1)$, and let $K = K(E)$ be the collection of unit balls centered at E, which are a ball sample of $B(x, r)$. We Denote by ε the approximation quality of K, and by μ the approximation quality of E. To conclude this section we state a lemma that gives an upper bound on μ as a function of ε, which we have already used in Section 4.2.

Lemma 5.5 *Let $B = B(x, r)$ be a ball such that $r \geq 1$, and let E be a point sample of $B(x, r-1)$. Let ε denote $d_H(\partial B(x,r)| \cup K(E))$, and μ denote $d_H(\partial B(x, r-1)|E)$. If $\varepsilon < 1/2$ then:*

$$\mu^2 \leq \frac{5}{2}\varepsilon \ .$$

Proof:

Let E be a point sample as above and let μ be its approximation quality. Let f be a point on $B(x, r-1)$ where the distance from the nearest sample point is μ (see Figure 5.1). Let $c \in E$ be the closest point to f and p be the point on ∂B closest to f. Let $k = B(c, 1)$ denote a ball in $K(E)$. The distance of p to $\cup K$ is achieved on k and is equal to $\bar{\varepsilon}$ for a real $\bar{\varepsilon} \leq \varepsilon$. Applying the same analysis as in Lemma 5.4, we get:

$$\mu^2 = (\bar{\varepsilon}^2 + 2\bar{\varepsilon})\frac{r-1}{r} \leq (\varepsilon^2 + 2\varepsilon)\frac{r-1}{r} \leq \frac{5}{2}\varepsilon \ .$$

□

Chapter 6

Topological Persistence

Our main goal in this thesis is to find pathways in the complement of a molecule. Yet Algorithm 1 has an added value: the collection $\mathcal{K}_\varepsilon$ of unit balls can be used to construct a good approximation of the persistence diagram of the Euclidean distance function from a union of balls. In Part III we make use of the constructed persistence diagram to compute the center of the largest chamber in a molecule, which we use as a root starting point when constructing the desired pathways. In Section 6.1 we repeat verbatim the definitions introduced by Cohen-Steiner *et al.* [11], which we need in order to state our results in Section 6.2. The reader is referred to [21] for an introduction to Homology that is both rigorous and accessible to non-specialists.

6.1 Background

Given a topological space X and an integer k, we denote the k-th singular homology group of X by $H_k(X)$, and the k-th Betti number of X by $\beta_k(X) = \dim H_k(X)$. We work here with modulo 2 coefficients, so that homology groups are vector spaces over $\mathbb{Z}_2 = \mathbb{Z}/2\mathbb{Z}$.

Definition 6.1 *[11] Let X be a topological space and f a real function on X. A homological critical value of f is a real number A for which there exists an integer k such that for all sufficiently small $\varepsilon > 0$ the map $H_k(f^{-1}(-\infty, A-\varepsilon]) \to H_k(f^{-1}(-\infty, A+\varepsilon])$ induced by inclusion is not an isomorphism.*

Definition 6.2 *[11] A function $f : X \to \mathbb{R}$ is tame if it has a finite number of homological critical values and the homology groups $H_k(f^{-1}(-\infty, A])$ are finite-dimensional for all $k \in \mathbb{Z}$ and $A \in \mathbb{R}$.*

In other words, the homological critical values are the levels where the homology of the sub-level sets changes. Assuming a fixed integer k we write $F_x = H_k(f^{-1}(-\infty, x])$, and for $x < y$ we define $f_x^y : F_x \to F_y$ to be the map induced by inclusion of the sub-level set of x in that of y. We write $F_x^y = \text{im } f_x^y$ for the image of F_x in F_y. By convention we set

$F_x^y = \{0\}$ whenever x or y is infinite. Let $\beta_x^y = \dim F_x^y$ denote the *persistent Betti number* for all $-\infty \leq x \leq y \leq +\infty$.

Let $f : X \to \mathbb{R}$ be a tame function, $(a_i)_{i=1\ldots n}$ its homological critical values, and $(b_i)_{i=1\ldots n}$ an interleaved sequence, namely $b_{i-1} < a_i < b_i$ for all i. We set $b_{-1} = a_0 = -\infty$ and $b_{n+1} = a_{n+1} = +\infty$. For two integers $0 \leq i < j \leq n+1$, we define the *multiplicity* of the pair (a_i, a_j) by $\mu_i^j = \beta_{b_{i-1}}^{b_j} - \beta_{b_i}^{b_j} + \beta_{b_i}^{b_{j-1}} - \beta_{b_{i-1}}^{b_{j-1}}$. Denoting by $\bar{\mathbb{R}}$ the union $\mathbb{R} \cup \{-\infty, +\infty\}$ we are ready to define the persistence diagram.

Definition 6.3 *[11] The persistence diagram $D(f) \subset \bar{\mathbb{R}}^2$ of f is the set of points (a_i, a_j), counted with multiplicity μ_i^j for $0 \leq i < j \leq n+1$, union all points on the diagonal, counted with infinite multiplicity.*

For points $p = (p_1, p_2)$ and $q = (q_1, q_2)$ in $\bar{\mathbb{R}}^2$, let $\|p - q\|_\infty$ be the maximum of $|p_1 - q_1|$ and $|p_2 - q_2|$. Similarly for functions f and g, let $\|f - g\|_\infty = \sup_x |f(x) - g(x)|$. Let X and Y be two multisets of points.

Definition 6.4 *[11] The bottleneck distance between X and Y is*

$$d_B(X, Y) = \inf_\gamma \sup_x \|x - \gamma(x)\|_\infty ,$$

where $x \in X$ and $y \in Y$ range over all points and γ ranges over all bijections from X to Y. Cohen-Steiner *et al.* prove [11] that small changes in f imply small changes under the bottleneck metric in the persistence diagram. We use a weakened version of their main theorem that is sufficient for our needs.

Theorem 6.5 *[11] Let A, A' be two subsets of $\mathbb{R}^3$ such that $d_H(A, A') \leq \varepsilon$. Let $f_A, f_{A'}$ denote the distance from A, A' respectively. The persistence diagrams of $f_A, f_{A'}$ satisfy*

$$d_B(D(f_A), D(f_{A'})) \leq \varepsilon .$$

6.2 Our Contribution

The set of unit balls $\mathcal{K}_\varepsilon$, which our algorithm constructs for a given set $\mathcal{B}$ of balls, satisfies $d_H(\cup\mathcal{B}, \cup\mathcal{K}_\varepsilon) \leq \varepsilon$. Let $f_\mathcal{B}, f_{\mathcal{K}_\varepsilon}$ denote the distance functions from $\cup\mathcal{B}, \cup\mathcal{K}_\varepsilon$ respectively. Note that the homology groups of the two unions are not necessarily the same. But the unions are similar in some sense: the bottleneck distance between the persistence diagrams of the two functions is bounded by ε according to Theorem 6.5. This implies that ‘major’ topological features, i.e., large voids or tunnels are the same for both unions, or formally:

Lemma 6.6 *Let $\mathcal{B}$ be a set of balls, each of radius at least 1, and let $\mathcal{K}_\varepsilon$ be a set of unit balls such that $d_H(\cup\mathcal{B}, \cup\mathcal{K}_\varepsilon) \leq \varepsilon$. Let $f_\mathcal{B}(x)$ (resp. $f_{\mathcal{K}_\varepsilon}(x)$) be the distance function defined on the complement of $\cup\mathcal{B}$ (resp. $\cup\mathcal{K}_\varepsilon$) of a point x in $\mathbb{R}^3$ from $\cup\mathcal{B}$ (resp. $\cup\mathcal{K}_\varepsilon$). Then the persistence diagrams of $f_\mathcal{B}(x)$ and $f_{\mathcal{K}_\varepsilon}(x)$ satisfy*

$$d_B(D(f_\mathcal{B}), D(f_{\mathcal{K}_\varepsilon})) \leq \varepsilon .$$

We believe that our approximation scheme can be used whenever the persistent topology bears more meaning than the exact topology, trading topological precision with the Euclidean metrics of the persistence diagram.

Part III

Implementation and Applications

Chapter 7

Pathway Diagram: Implementation Details

In Section 7.1 we describe how we implemented the procedure FLOWER(B, ε) which constructs an ε-flower for a single ball $B \in \mathcal{B}$. In Section 7.2 we explain how we construct the pathway diagram $\mathcal{P}_\varepsilon$ using ready-made tools. (See Chapter 3 for the definitions of these entities.)

7.1 ε-Flower Construction

For each ball $B = B(c, r) \in \mathcal{B}$ we construct a set of unit balls K_B that are an ε-flower of B. Constructing an ε-flower with a minimal number of unit balls is an optimization problem, closely related to the following problem: "how can n points be distributed on a unit sphere such that they maximize the minimum distance between any pair of points?". Such a configuration of points is called a *spherical code* and its construction has been intensively studied [26]. We employ two heuristic sampling techniques for producing an ε-flower.

The first heuristic is **Icosahedron refinement.** An *icosahedron* is the Platonic solid P_3 having 12 vertices, 30 edges, and 20 congruent equilateral triangular faces. We denote by $I_0(c, r)$ an icosahedron that has its vertices on the sphere $S = S(c, r-1)$. $I_0(c, r)$ can be refined by adding a vertex in the midpoint of each polyhedron edge and centrally projecting it onto the sphere S. Three new polyhedron edges are added within each triangle, that connect the three new vertices on its boundary edges. Each triangle of $I_0(c, r)$ is split into four new smaller triangles. In this way we get a *refined icosahedron* of degree one, denoted by $I_1(c, r)$, which compromises 42 vertices, 80 triangles and 120 edges. We continue the refinement recursively and define a sequence of refined icosahedra: $I_0(c, r), I_1(c, r), I_2(c, r)$, etcetera. For any natural number η we define $\varepsilon_{\text{ico}}(\eta)$ to be the one-sided Hausdorff distance between the boundary surface of B and the union of the collection of unit balls centered at the vertices of $I_\eta(c, r)$. We compute $\varepsilon_{\text{ico}}(\eta)$ by scanning all triangles of $I_\eta(c, r)$. Given the user-specified parameter $\varepsilon > 0$ we find the smallest η such that $\varepsilon_{\text{ico}}(\eta) \leq \varepsilon$, and return the vertices of $I_\eta(c, r)$.

The second heuristic that we employ is **Random points.** The icosahedron refinement technique has a major drawback: The number of vertices in each icosahedron in the sequence jump in large steps, i.e., $12, 42, \ldots$ Since the first icosahedron has 12 vertices, we have a 'gap' between 2 and 11 that we wish to bridge. We use a naive random sampling technique to generate preprocessed samples as described next.

We repeat the following procedure for each $i = 2, \ldots 11$. Using a typically large integer constant N_{rnd}, we generate $j = 1, \ldots, N_{\text{rnd}}$ random point sets E_{ij}, each containing exactly i points, such that the points of E_{ij} are located on the sphere $S(c, r-1)$. Recall that for any finite point collection $E \subset \mathbb{R}^3$ we denote by $K(E)$ the collection of unit balls centered at E. We choose the set $E_{i\tilde{j}}$, $1 < \tilde{j} \leq N_{\text{rnd}}$, such that:

- The union of the unit balls in $K(E_{i\tilde{j}})$ is homeomorphic to a ball or a spherical shell.
- For any $1 < j \leq N_{\text{rnd}}$ it holds that $d_H(\partial B| \cup K(E_{i\tilde{j}})) \leq d_H(\partial B| \cup K(E_{i\tilde{j}}))$.

We denote $E_{i\tilde{j}}$ by E_i and denote the one-sided Hausdorff distance $d_H(\partial B| \cup K(E_{i\tilde{j}}))$ by $\varepsilon_{\text{rnd}}(i)$. After completing the procedure for all $2 \leq i \leq 11$ we have 10 computed point sets $\{E_i\}_{i=2,\ldots 11}$ with their respective one-sided Hausdorff distances $\{\varepsilon_{\text{rnd}}(i)\}_{i=2,\ldots 11}$. Now, given the user-specified parameter ε, we choose the minimal $2 \leq i \leq 11$ such that $\varepsilon_{\text{rnd}}(i) \leq \varepsilon$. If i exists we use E_i as the centers of the ε-flower. If no i satisfies $\varepsilon_{\text{rnd}}(i) \leq \varepsilon$ we use icosahedron refinement.

7.2 Computing the Pathway Diagram

We compute the $(\alpha = 1)$-Voronoi graph of E_ε using the 3D Alpha Shapes package [12] of the CGAL library. We use exact arithmetic to ensure that the algorithm is robust. The Alpha Shapes data structure allows to retrieve the α-complex for any α value. We set the pathway diagram $\mathcal{P}_\varepsilon$ to be the collection of dual Voronoi faces of the simplices that are *not* in the $(\alpha = 1)$-complex of E_ε.

Chapter 8

MolAxis

We call a possible route in a protein where smaller molecules can pass a *corridor* (we formally define corridors in Section 8.1). Based on the ideas presented so far in the thesis we devised a program, MolAxis, to assist the biologist/biochemist in finding corridors in molecules. Given a molecule represented as a union of balls, on top of the pathway diagram of the molecule, we construct a tree which we call the *corridor tree*, which captures possible relevant pathways in the pathway diagram. In this and the next chapter we assume some familiarity with basic terminology in molecular biology.

First we wish to provide some intuition as to what is a corridor. Imagine a volcano erupting at a given point which lies outside the molecule volume (for example, located inside a molecular chamber). The lava is flowing out of the volcano mouth in a set of streams that flow faster where the passage is wide and slower where the passage is narrow. Whenever a stream reaches an obstacle (like the molecule or another stream) that cannot be bypassed it stops flowing. Streams tend to balance between length and clearance and they represent corridors in this analogy.

As already noted, we model a molecule using a collection of three-dimensional balls, one ball per atom and assign each atom ball its corresponding van der Waals (VDW) radius. For a given molecule, we set our input ball collection $\mathcal{B}$ to be the collection of atom balls. We scale the atoms such that the smallest atom ball is a unit ball. Given a user-specified parameter $\varepsilon > 0$ we employ Algorithm 1 of Part I, which constructs a set of sample points E_ε, the collection of unit balls $\mathcal{K}_\varepsilon$ centered at E_ε, and the pathway diagram $\mathcal{P}_\varepsilon$ of E_ε. In this chapter we elaborate on how we define and extract corridors from the pathway diagram, as implemented in the MolAxis program.

8.1 Definitions

During the execution of Algorithm 1 each atom ball $B \in \mathcal{B}$ is replaced by a collection K_B of unit balls. We call each ball $K \in K_B$ an *approximate ball* of B, and refer to the collection $\mathcal{K}_\varepsilon = \cup_{B \in \mathcal{B}} K_B$ of unit balls as the *approximate balls* of the molecule. We keep a two-directional mapping between each atom ball $B \in \mathcal{B}$ and the collection K_B of unit balls in $\mathcal{K}_\varepsilon$, which approximate B. We call the union of the approximate balls the *approximate molecule.* A *pathway* π is a curve in space that lies outside the approximate molecule and is contained in the pathway diagram.

The *approximate VDW surface* is the boundary surface of the approximate molecule. Let π be a pathway and p be a point on it. The *clearance* $c(p)$ of p is the minimal distance between p and the approximate VDW surface. The *lining balls* of p are the collection of (one or more) approximate balls with a minimal distance to p. We call an atom $B \in \mathcal{B}$ a *lining atom* of p if at least one of the lining ball(s) of p approximates B. A *lining residue* of p is a residue that contain one or more lining atoms of p.

The *profile* of π is the clearance of the points on π as a function of the distance along the pathway. The *pathway ball* of $p \in \pi$ is the ball with radius $c(p)$ that is centered at p. The *pathway surface* of π is the boundary (envelope) surface of the union of all pathway balls of π. The *bottleneck radius*[1] of π is the minimal clearance along the pathway, and the *bottleneck point* of π is the point in π where the bottleneck radius is achieved. The *bottleneck atoms* (resp. *bottleneck residues*) are the lining atoms (resp. lining residues) of the bottleneck point.

The *exact clearance* $\bar{c}(p)$ of a point p in the complement of $\cup\mathcal{B}$ is the distance between p and the (exact) VDW surface. We restate the following observation which is proved in Lemma 4.8 of Part II:

Observation 8.1 *For any $p \in \mathcal{P}_\varepsilon$ such that $c(p) > \varepsilon$ it holds that $|c(p) - \bar{c}(p)| \leq \varepsilon$.*

This observation justifies our use of the clearance function as an approximation of the exact clearance function.

8.2 Pathway Graph Construction

The pathway diagram is composed of two-dimensional patches. In order to reduce the problem to a one-dimensional problem we create a graph which contains only vertices and edges. First, we discard all facets of the pathway diagram. Due to geometric properties of the Voronoi diagram of points, the maximal clearance of a bounded facet is achieved on one of its boundary vertices. Thus discarding facets favors pathway clearance (at the possible expense of increasing the pathway length). A more robust approach could have been to sample vertices within the facet, yet we did not find this improvement necessary. Second, some edges of the pathway diagram are rays, going to infinity. We replace the rays with segments by intersecting the pathway diagram with a bounding sphere, as we describe next.

In Chapter 4 we used a large bounding sphere $Q = B(c_q, r_q)$ that contains $\cup\mathcal{B}$ for the sake of the analysis. Here we set Q to be user-specified sphere. Q represents the locus of interest within $\mathbb{R}^3$, i.e., all computation will be limited to the inside of Q. For each Voronoi edge $e = (v_i, v_o)$ that intersects Q, with v_i inside Q and v_o outside Q, we construct

[1] Do not confuse this term with the unrelated *bottleneck distance* which is defined and used in Chapter 6.

a new vertex v_e that is the intersection of e and Q. We call v_e a *boundary vertex*. We also construct a *boundary edge* $e' = (v_i, v_e)$, which is the part of e that lies within Q. We define V to be the collection of Voronoi vertices of the pathway diagram that lie within Q along with all the boundary vertices. In a similar fashion, we define E to be the collection of Voronoi edges of the pathway diagram that lie completely within Q along with all the boundary edges. We construct a graph $G = G(V, E)$ which we call the *pathway graph*. From this point on we restrict ourselves to pathways that are contained in the pathway graph.

8.3 Corridor Tree Construction

Pathways are not unique and more than one pathway can exist between two points. There are several ways to define an optimal pathway between two points. One way is the shortest pathway between the two points. Another way is to focus only on the clearance of the pathway. The shortest pathway between two given points typically has the undesirable property that it can get arbitrarily close to the boundary of the molecule and hence has close to zero bottleneck radius. High clearance pathways, on the other hand, can be extremely long. We are therefore interested in finding pathways that balance between length and clearance.

For any vertex $v \in V$ located at $p \in \mathbb{R}^3$ we define its clearance to be $c(v) = c(p)$. For each edge $e = (v, v') \in E$, the *edge clearance* $c(e)$ is defined as follows:

$$c(e) = \min\left(C_{\max}, \frac{c(v)}{2} + \frac{c(v')}{2}\right),$$

where $C_{\max}$ is a user-defined constant, which serves as an upper bound on the clearance. We define a weight function $w(e)$ over the edges that accounts for the length of the edge and the clearance of points along the edge as follows:

$$w(e) = \frac{d(v, v')}{(c(e))^2}.$$

We employ a minimum weight optimization algorithm on the graph G and compute a tree. The weight function $w(e)$ favors pathways that are both short and wide and can be seen as a flux optimization. The weight function can be easily modified and adapted to optimize other criteria. We select a root vertex s in a manner described below and compute the tree rooted at s using Dijkstra's algorithm [15] on G with the weights defined by $w(e)$. We call this tree the *corridor tree* of the molecule.

During the computation of the tree each vertex v is assigned a *flux weight* $W(v)$, which is the sum of the weights of the edges on the path between s and v. We say that $u \in V$ is an *ancestor* of $v \in V$ if it is contained in the (single) path from the root vertex s to v in the corridor tree. A vertex $v \in V$ is called a *leaf vertex* if it is a leaf in the corridor tree, i.e., v is not an ancestor of any vertex in V. A *corridor* π is a path in the corridor tree that reaches a leaf vertex v_π. We define the flux weight of a corridor π to be the flux weight $W(v_\pi)$ of its leaf vertex v_π.

8.4 Querying the Corridor Tree

The corridor tree construction is done as a preprocess, and it is saved to a file. We support various queries to allow the user to identify, display and analyze a single corridor in the corridor tree. The user can query the corridor tree in order to identify the corridor that has the smallest flux weight and that passes through a user-specified sphere. MolAxis gives as output the corridor profile, lining atoms, lining residues, bottleneck radius, bottleneck atoms and bottleneck residues. For visualization purposes, MolAxis constructs the corridor surface of channels either as a collection of balls (see Figure 9.4) or as a meshed surface (see Figure 9.7).

MolAxis supports special queries designed for two scenarios, namely *chamber channels* (see Figure 9.4) and *cross channels* (see Figure 9.2). Chamber channels are channels that connect an inner chamber to the outside. MolAxis identifies chamber channels by reporting on corridors that connect the root vertex to boundary vertices as described below. A cross channel is a channel that crosses the protein from side to side, like a transmembrane channel. In this case MolAxis gives as a result a concatenation of two corridors that together represent the channel as described below.

8.5 Chamber Channels

In the first scenario, dealing with chamber channels, we select the root vertex s to be one of the vertices within the chamber. This is done either by selecting a vertex closest to a user-specified point or by automatically computing the center of the largest chamber in the protein. The latter option is called *auto mode*. The largest chamber is deduced using persistent topology techniques similar to the one used by Edelsbrunner *et al.* [17] (see also Chapter 6). It is the vertex in the center of the last remaining void if the approximated balls are inflated in a uniform manner. We distinguish between two types of chamber corridors. Corridors that reach a boundary vertex are called *exit corridors* since they exit the chamber, while all other corridors are called *dead-end corridors*. We use exit corridors to represent the molecular channels.

We define the *forking vertex*[2] $v(\pi_1, \pi_2)$ of two corridors π_1, π_2 to be the last identical vertex in the path from s to the leaf vertices of π_1 and π_2. The vertex $v(\pi_1, \pi_2)$ might be located far outside the chamber or even outside the convex hull of the molecule, which means that the two corridors actually represent the same channel. In this case one of the corridors should be discarded. We introduce a user-specified parameter $F_{\max}$ called the *forking threshold* to control when corridors are discarded as described below.

[2] The forking vertex is also known as the *least common ancestor* of the leaf vertices of π_1 and π_2.

First we color all vertices in V *blue*. We traverse all exit corridors in a sequence $\pi_1, \pi_2 \ldots$, sorted in ascending order of their flux weight, i.e., starting from the exit corridor that has the best (lowest) flux weight. Let π_i be an exit corridor in the sequence. The *forking weight* of π_i for $i > 1$ is the maximal flux weight of all forking vertices with respect to all previous corridors in the sequence:

$$f(\pi_i) = \max_{0<j<i}(W(v(\pi_i, \pi_j))) \ .$$

We set $f(\pi_i)$ to be the flux weight of the path between the root vertex s and the last *red* vertex in π_i (in the first iteration no vertex is colored *red*, so we trivially set $f(\pi_1) = 0$). We report to the user the corridor π_i only if its forking weight is smaller than the forking threshold F_{max}. If the forking weight of π_i is not smaller than the forking threshold, we regard π_i as similar to a previously reported exit corridor and therefore discard it. We then color the vertices of π_i *red*, and continue to the next exit corridor in the sequence.

8.6 Cross Channels

In this scenario our primary purpose is to identify transmembrane (TM) channels. We add an imaginary vertex v_∞ at infinity, connect it with edges to all the boundary vertices and set the weight of these new edges to zero. We set the root s to be v_∞ and compute the corridor tree. The user must specify a *cross plane*, which is a plane that splits the boundary vertices into two groups: *above vertices* and *below vertices*. Given an edge $e = e(v_1, v_2) \in E$ that is *not* in the corridor tree, we call e a *crossing edge* if the ancestor boundary vertex u_1 of v_1 is an above vertex and the ancestor boundary vertex u_2 of v_2 is a below vertex. In this case, we define the *crossing corridor* π_e to be the concatenation of the corridor leading from u_1 down to v_1 in the corridor tree, the edge e, and the corridor leading from v_2 up to u_2 in the corridor tree. There might be crossing corridors that bypass the transmembrane channel. Therefore we discard crossing corridors that do not pass through a user-specified sphere. If no crossing corridors are left we report that the channel is closed. If there is more than one candidate for a crossing corridor we report the crossing corridor with the smallest flux weight.

Chapter 9

Experimental Results and Discussion

All tests were carried out on a Pentium IV 3.0 GHz machine with 1GB of RAM running a LINUX native operating system. Recall that the user-defined constant $C_{\max}$ is an upper bound on the clearance. In all runs $C_{\max}$ was set to 1.4Å, and the forking threshold $F_{\max}$ was set to 6.

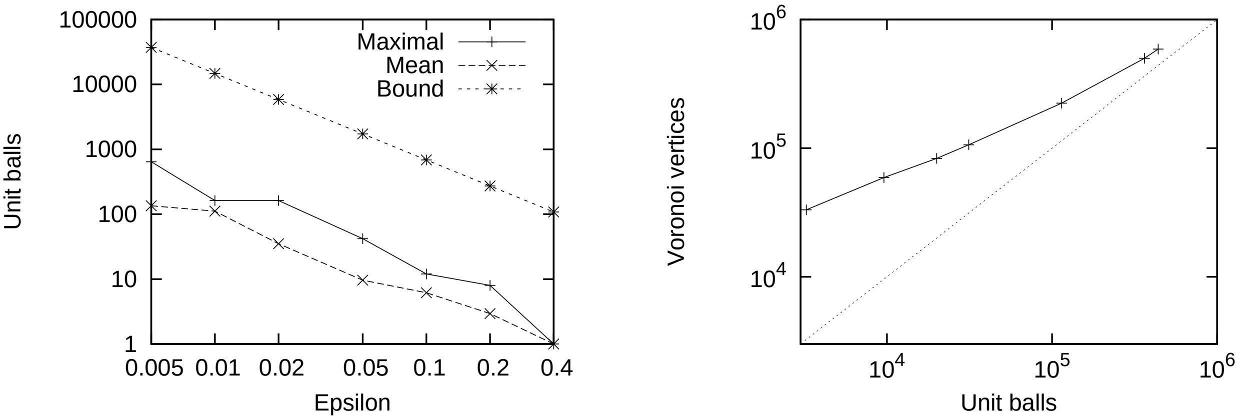
9.1 Experimental Results

First, we report on a set of basic tests that were performed on a collection of 3251 balls, which model a P450 Enzyme isozyme (H atoms discarded). The ratio between the largest and smallest input balls is $\rho = 1.21$. In Table 9.1 we report on a set of tests performed on these input balls with varying resolution. In Figure 9.1 we provide (a) a graph of the size of $\mathcal{K}_\varepsilon$ as a function of ε, comparing the experimental results to the theoretical bound proved in Theorem 5.1 and (b) a graph of the number of Voronoi vertices in the pathway diagram as a function of the number of balls in $\mathcal{K}_\varepsilon$. Note that as ε becomes smaller and the number of unit balls increases, the ratio between the number of Voronoi vertices and number of unit balls tends to one. We believe that the explanation of this phenomenon is that our sampled points behave like points sampled on a smooth surface [5] or points sampled on a polyhedral surface [3]. In these two cases the number of faces in the Voronoi diagrams is less than the worst case $\Theta(n^2)$ faces, and is $O(n \log n)$. Table 9.2 we provide the runtime breakdown of three runs (for $\varepsilon = 0.4$, $\varepsilon = 0.1$, $\varepsilon = 0.01$).

In the following sections we describe how MolAxis was employed in a biological context. MolAxis was found to be a very efficient and sensitive tool in identifying transmembrane (TM) channels within proteins and pathways leading from buried cavities within enzymes to their surfaces. In two different TM channels, MolAxis detected the main TM channel and its axis taking into account the global geometry of the protein rendering a fairly smooth and representative route for a ligand passing through the channel. MolAxis also detected in few seconds all previously characterized substrate and water channels connecting the P450 enzymes main cavity, which is the active site, to the outside of the protein. MolAxis identified channels even if they are nearly closed and pointed to new channels previously

Resolution	Max BPS	Unit Balls	Vertices	Runtime	File size
0.4	1	3251	33171	2.456	2.6
0.2	8	9585	59064	5.120	4.5
0.1	12	19996	83272	8.737	6.3
0.05	42	31287	106107	12.229	7.9
0.02	162	114222	223864	35.522	16.9
0.01	162	362262	499761	101.830	37.7
0.005	642	439182	589304	123.444	44.4

Table 9.1: Complexity analysis for a set of runs with different resolutions on a single instance of the P450 enzyme. *Resolution* is the Hausdorff approximation quality ε. *Max BPS* is the maximal number of unit balls used to approximate a single ball. *Unit Balls* is the total number of unit balls used in the approximation (the number of balls in $\mathcal{K}_\varepsilon$). *Vertices* is the number of vertices in the pathway graph. *Runtime* is the total runtime in seconds, including the construction of the pathway diagram, pathway graph and corridor tree. *File size* is the size in MB of the file that contains the corridor tree (called the *.pma* file).



(a)

(b)

Figure 9.1: Graphs generated for the set of runs that are analyzed in Table 9.1. All scales are logarithmic. (a) Experimental and theoretical values of the ratio $|\mathcal{B}|/|\mathcal{K}_\varepsilon|$. *Maximal* is the maximal BPS, i.e. the maximal number of unit balls used to approximate a single ball. *Mean* is the mean BPS over all input balls. *Bound* is the theoretical upper bound as expressed in Theorem 5.1, with $\kappa = 3$. Note that for $\varepsilon > 0.2$ we need exactly one unit ball to approximate each input ball. (b) The number of Voronoi vertices in the pathway graph as a function of the number of unit balls in $\mathcal{K}_\varepsilon$. We draw as a reference the function $y = x$ using a dotted line. Note that as the number of unit balls increases the ratio between the number of Voronoi vertices and the number of unit balls becomes closer to one.

Phase	$\varepsilon = 0.4$	$\varepsilon = 0.1$	$\varepsilon = 0.01$
Construct point sample	0.016 (0.6%)	0.088 (1%)	1.720 (1.7%)
Compute triangulation	0.088 (3.5%)	0.568 (6.5%)	10.645 (10.5%)
Compute alpha shape family	0.456 (18.5%)	2.848 (34%)	51.471 (51.5%)
Construct corridor tree	0.112 (4.5%)	0.328 (3.9%)	2.692 (2.7%)
Stream to file	1.000 (40%)	2.412 (28%)	14.25 (14.2%)

Table 9.2: Runtime breakdown of a MolAxis run on a single instance of the P450 enzyme (see the relevant rows in Table 9.1 for more details about these runs). The time is in seconds, the percentage of the total runtime is in parenthesis. We report major phases only (that is why the percentages do not sum to 100%). Note that the most time-consuming phase is the Alpha shape computation, and as ε decreases it becomes more dominant.

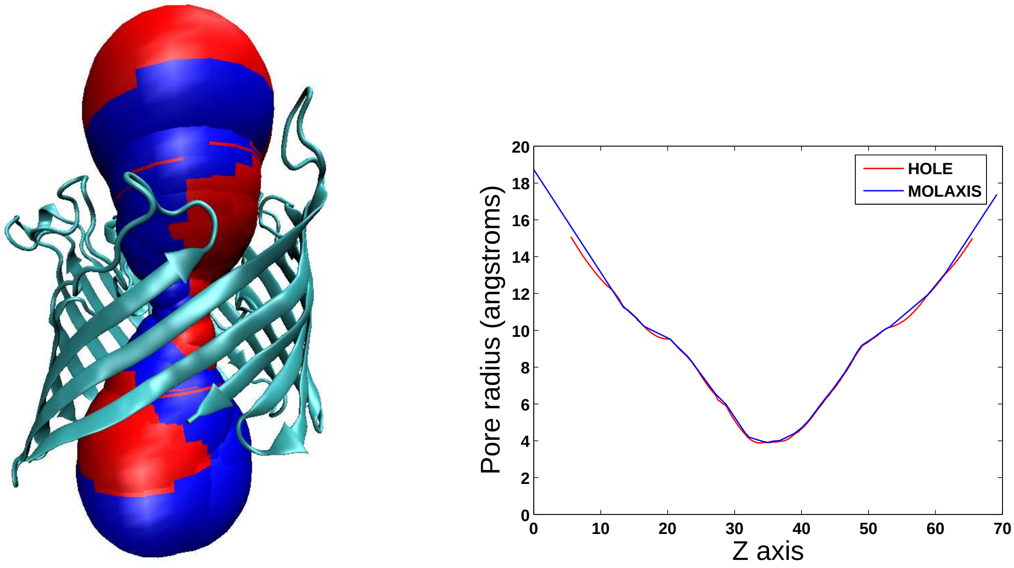
unidentified. As being accurate and highly efficient, MolAxis was used to analyze large data driven from Molecular Dynamics (MD) simulation of the human CYP3A4 enzyme in order to understand channels dynamics and gating mechanisms along time.

9.2 Large Pore Channels

Large pore channels (LPC) are membrane proteins located in the outer membrane of the bacteria and allow the supply of macromolecules (usually sugars) to the cytoplasm. The specific LPC structure that we experiment with, PDB code 1PRN, consists of a β-barrel that forms a straight channel with a narrowing in its middle and a wide open conformation at both its ends. As can be seen in Figure 9.2, the HOLE and MolAxis programs agree on the results and almost identically find the same pathway and profile with a similar running time of about 5 seconds. The bottleneck radius of the channel, as computed by MolAxis, is 3.9Å.

9.3 ABC Transporter

Adenosine triphosphate (ATP) binding cassette (ABC) transporters catalyze the translocation of substrate against a transmembrane concentration gradient by coupling this unfavorable passage to hydrolysis of ATP. In general both HOLE and MolAxis have similar outputs. Although HOLE and MolAxis calculate similar pore radius profiles, the axes of the channel are quite different as seen in Figure 9.3. The path found by HOLE has a 'zigzagging' pattern while MolAxis detects a smoother path of the channel which better represents the channel geometry. We further compare MolAxis and HOLE in Section 9.9.



(a) (b)

Figure 9.2: (a) LPC channel surface as calculated by MolAxis (blue) and HOLE (red). The LPC is shown in cartoons. (b) LPC pore radius along the Z-axis (in Å) as calculated by MolAxis (blue) and HOLE (red). The Z-axis vector is roughly aligned with the channel direction.

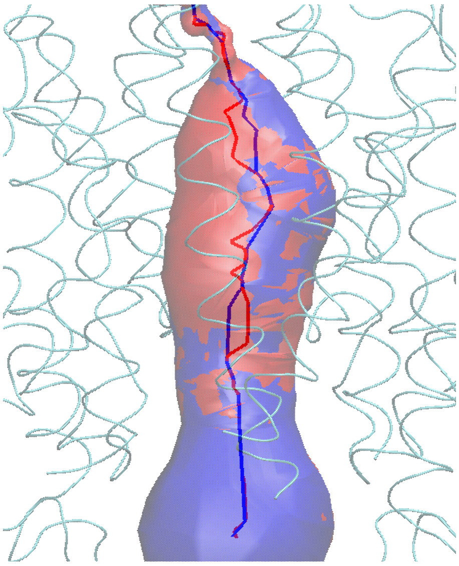


Figure 9.3: Pathway and pathway surface of the ABC transporter as computed by MolAxis (blue) and HOLE (red). Note that the pathway found by HOLE has a 'zigzagging' pattern while the corridor constructed by MolAxis is smoother.

9.4 Static Analysis of P450

Cytochrome P450 proteins constitute a large family of mono-oxygenases heme containing enzymes that oxidize a variety of chemical compounds in microorganisms. The oxidation of a substrate occurs at the hydrophobic core of the protein. It is of great mechanistic and biochemical interest to identify and characterize all channels that link the active site to bulk solvent both statically and dynamically by means of MD simulations. We focus here on the human CYP3A4, a P450 isozyme.

We match a corridor to a channel if it exits through the relevant secondary structure elements of the channel. We found that the correspondence between corridors and channels is high but it is not one-to-one. First, some channels had no corresponding corridors. This can happen when a channel is closed or when it is nearly closed and its exit mouth is close to a wider channel. We call the latter phenomenon *overshadowing*, since another channel is hiding the relevant channel; see more details in Section 9.6 below. Second, multiple corridors can match the same channel. We allow the user to address this problem by adjusting the forking threshold (see Section 8.5). A third possibility is corridors with no corresponding channels. This either signifies a possibly newly discovered channel or a random exit route that opens for a short time during an MD simulation.

In Figure 9.4 we show the channels of the human CYP3A4 as computed using MolAxis. In Table 9.3 we show the correspondence between the corridors found by MolAxis and the channels of CYP3A4.

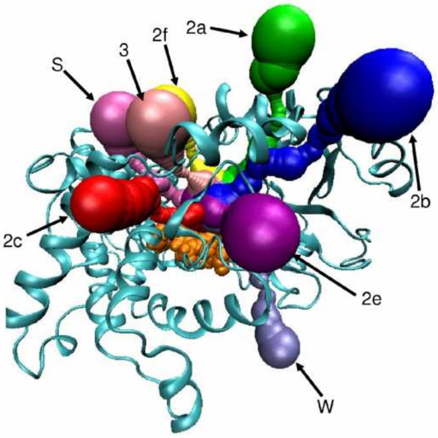


Figure 9.4: CYP3A4 channels as detected by MolAxis. CYP3A4 is represented by cartoons and the heme prosthetic group is represented by its VDW surface and colored orange. Each channel surface is colored in a different color for the sake of clarity.

ε	**2a**	**2b**	**2c**	**2d**	**2f**	**3**	**S**	**W**	**Other**
0.2	3	2	5	1	6	7	4	8	9,10
0.3	3	2	4	1	7	6,9	5	-	8,10
0.4	4	2	3	1	7	5	6	9	8,10

Table 9.3: Ranking of the first 10 identified exit corridors according to the flux weight, in the human CYP3A4. The biological channels are called '2a', '2b' etc. Exit corridors that do not correspond to any known channel are placed in the 'Other' column.

9.5 MD Simulation Analysis of P450

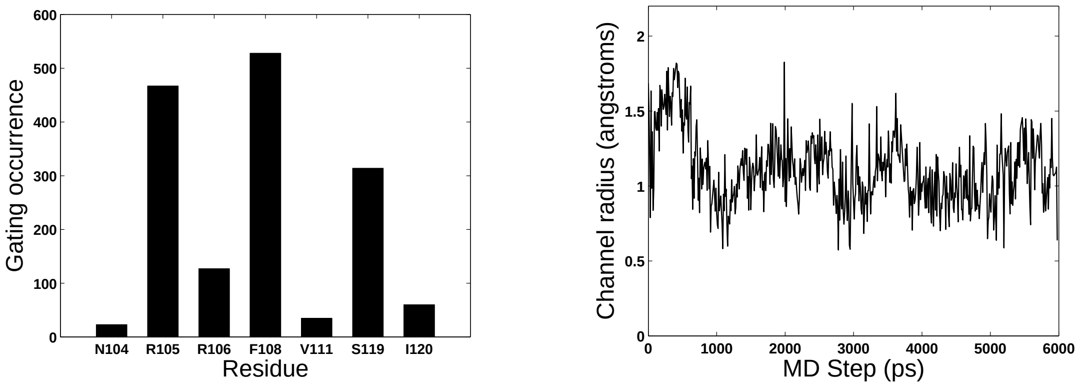
Enzyme dynamics and motions which may control the opening and closing of channels are not apparent in static crystal structures and may be missed in structural analysis. In addition, changes in the dimensions of viewable channels may be overlooked. MD simulation is a good tool for assessing channel movements along time and for comprehension of the channel's gating mechanism and the (cooperative) behavior of residues involved in its opening and closing.

Like in the previous section we focused here on the human CYP3A4. When employing MolAxis we used *auto mode* to find the largest cavity (which is the active site) and to detect all channels emanating from it to the surface of the enzyme. In most MD snapshots (roughly 95%) MolAxis detected the center of the active site of the human CYP3A4 as the largest void. The rest of the snapshots were overlooked as their largest cavities found were not placed in the middle of the active site and thus not biologically significant. As there are many substrate channels we simplified this example by focusing on one substrate channel, denoted '2e'. The major goal of this analysis is to obtain better insights of the gating mechanism and residues of the '2e' substrate access channel. We calculated the corridor surface, bottleneck radius and lining residues of channel '2e' along time. In Figure 9.5 we show the frequent bottleneck residues and the bottleneck radius of the channel over time. The analysis of a large number of snapshots was made possible by the high efficiency of MolAxis.

9.6 Channel and Corridor Correspondence

Recall that a corridor can exist without a corresponding channel, i.e., with no known biological related data, consequently pointing to a potential newly discovered channel. Interestingly, the opposite is also possible. In cases where a channel is closed or almost closed, the corridors that pass through it might actually be dead-end corridors. We call this phenomenon *overshadowing* and it occurs when the mouth of a narrow channel is close to the mouth of a wide channel. In this case we say that the wide channel overshadows the narrow channel.

For example, let us consider a conformation in which there are two channels, one wide and one narrow, which leave the chamber in about the same direction. Let us focus on



(a) (b)

Figure 9.5: Analysis of an MD simulation of channel 2e of the human CYP3A4 enzyme. (a) Portion of time each residue is a bottleneck residue of the corridor. (b) Bottleneck radius of the corridor over time. The x-axis units are picoseconds.

a vertex in the pathway graph that lies in the mouth of the narrow channel, such that any pathway passing through the narrow channel must pass through it; see Figure 9.6 for an illustration. Assume that there are two pathways reaching the vertex in the pathway graph. The first pathway passes through the narrow channel and reaches the vertex. The other pathway passes through the wide channel, then passes close to the outer surface of the molecule and finally reaches the vertex. This detour can happen within the convex hull of the molecule, i.e., along a groove. The narrow channel will be reported as open if the optimal pathway (in flux weight terms) is the one going through the narrow channel, otherwise it will be reported as closed. The user can control this phenomenon by lowering the user-defined constant $C_{\max}$, giving the clearance a lesser weight.

9.7 Geometric Convergence

Corridors found by MolAxis lie close to the medial axis of the complement of the molecule. We found that the computed corridors are not always identical at different resolutions. This happens since the medial axis is composed of surface patches, which leaves some freedom in choosing the one-dimensional pathways. Even if moving on the medial axis there might be numerous pathways that cross a channel, with a similar flux weight. We observed that at high resolution of less than 0.05Å, the corridor tree seems to converge (data not shown). Even so, we note that the seemingly different corridors which are obtained at low resolutions have similar features, e.g., they pass near the same amino acids and have similar profiles. Therefore we conclude that MolAxis can be run at low resolutions (such as 0.1-0.5Å).

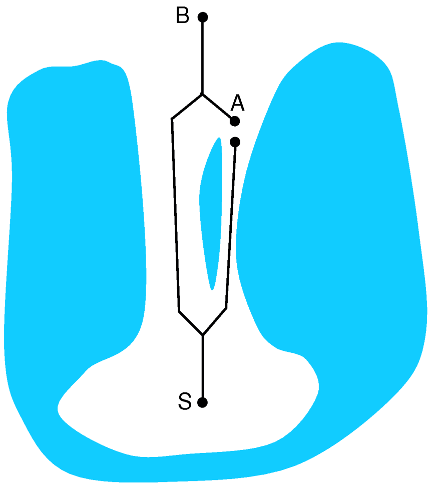


Figure 9.6: An overshadowed channel. The corridor tree is depicted using black lines, S is the root vertex, A is a vertex close to the mouth of the narrow right channel and B is a boundary vertex. Since the left channel is wide the shortest corridor reaching vertex A passes through the left wide channel. Therefore the corridor that passes through the right channel is a dead-end corridor and thus the channel is reported as closed.

9.8 Comparison with a Grid Based Approach

We compare our approach to a grid-based approach as implemented in the CAVER tool [23]. CAVER defines a three-dimensional grid covering the convex hull of the given molecule (represented by a union of balls). Each grid cell is then marked as *inner* or *outer* to the molecule, according to the center of the grid cell. All inner grid cells are discarded. The centers of the outer grid cells are set as vertices of a graph. CAVER connects neighboring grid vertices with an edge, and gives weights to the vertices according to their distance to the surface of the molecule. The corridors are computed using a version of Dijkstra's algorithm, similar to the one we use. The main difference from our approach is the number of vertices needed for the approximation. While limiting ourselves to the medial axis which is a two-dimensional entity, we construct much fewer vertices, which explains the extremely large difference in the running time between the two programs (from a couple of seconds of our program up to hours of CAVER, on the same input). This huge difference allows application of MolAxis along molecular dynamic trajectories, enabling to follow the channel dynamics (see Section 9.5). The pathways found by CAVER are close to the medial axis, which means that the grid points sampled far from the medial axis were actually not needed (see Figure 9.7a). Even so, note that this is not a property of all weight functions. For example, if the goal is to minimize a user-specified energy function, the desired pathway might not necessarily be close to the medial axis. A grid based approach can handle such a case whereas our approach would probably miss the desired pathway.

9.9 Comparison with a Monte Carlo Approach

The HOLE method [24] finds a possible route for a ball squeezing through the channel (changing its radius as it passes). Given a starting point in the channel cavity and a *channel vector*, which is a vector in the direction of the channel, HOLE moves a plane $P(t)$ that is orthogonal to the channel vector in steps along the vector using a parameter t (see Figure 9.7b). We denote by $S_{\text{opt}}(t)$ the largest sphere centered at the plane which can be accommodated in the channel without overlap with the VDW surface of the molecule. For each plane $P(t)$ HOLE uses a Monte Carlo simulated-annealing procedure to construct a sphere $S(t)$ on the plane $P(t)$ that is close to $S_{\text{opt}}(t)$. This procedure is iterated in the direction of the channel vector until a series of sphere positions is generated that represents the channel.

Note that the center of $S_{\text{opt}}(t)$ is actually the point of the medial axis centered at the plane $P(t)$ with the highest clearance. We compare with HOLE in two theoretical aspects. First, due to the non deterministic nature of the Monte Carlo procedure, $S(t)$ might in fact be far away from $S_{\text{opt}}(t)$. Second, since the optimization is done separately for each plane the computed pathway can be erratic or 'zigzagging', as seen in Figure 9.3. Furthermore, even if we assume HOLE managed to find $S_{\text{opt}}(t)$ the result pathway might exhibit an unwanted behavior. For the sake of the discussion let us extend the function $S_{\text{opt}}(t)$ from a discrete set of t values to all real values. Due to properties of the medial axis the function $S_{\text{opt}}(t)$ is not necessarily continuous in t, which results in 'jumps' in the pathway. We conclude that the optimization done by HOLE can be seen as a clearance-only optimization. In contrast, the pathways that MolAxis constructs are continuous, and due to the global optimization approach the pathways balance between length and clearance.

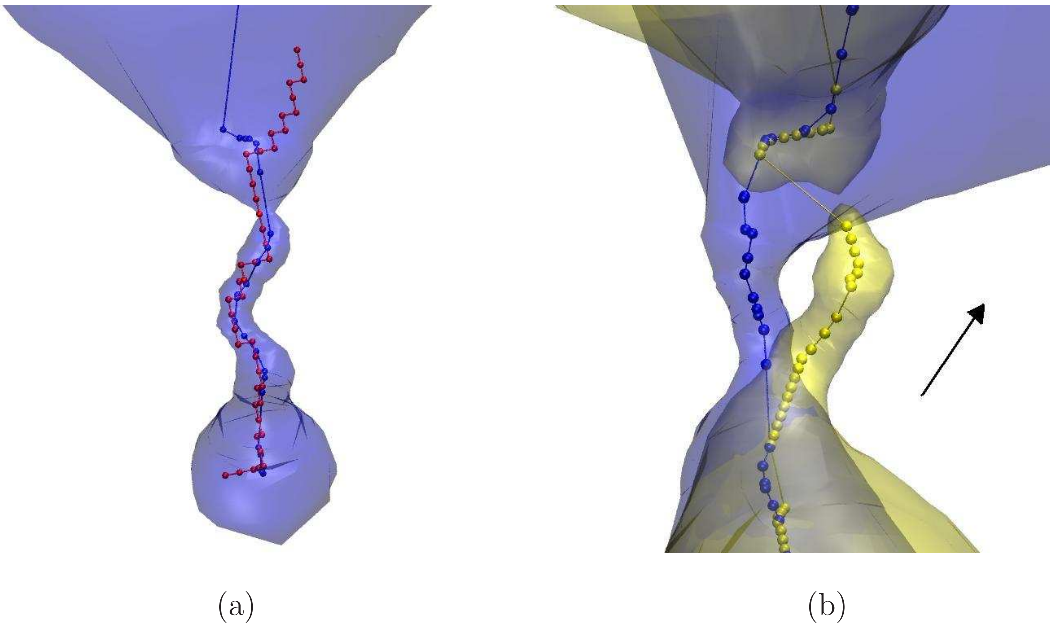


Figure 9.7: Comparison of pathways found by MolAxis with pathways found by CAVER and HOLE. (a) Pathways found by MolAxis (blue) and CAVER (red) of a P450 channel. Note the high correspondence between MolAxis and CAVER, suggesting CAVER pathways are near the medial axis. (b) Zoom in on pathways found by MolAxis (blue) and HOLE (yellow) of the ABC transporter. The channel vector used (black arrow) was not exactly aligned with the channel, i.e., there was roughly an angle of 45 degrees between the two. It is visible that MolAxis is not as sensitive to the channel vector as the HOLE program.

Chapter 10

Conclusions and Future Work

In this thesis we have introduced the pathway diagram of a collection $\mathcal{B}$ of balls in $\mathbb{R}^3$ each not smaller than a unit ball. We have shown how we employ it to construct pathways in the complement of their union. We proved several properties of the pathway diagram and reported on experimental results where the pathway diagram was used to identify pathways in the complement of molecules.

We have shown how our approximation scheme can be used to approximate the persistence diagram J of the Euclidean distance function from $\cup\mathcal{B}$. An alternative approach for approximating J is to compute the persistence diagram of the *power* distance from $\cup\mathcal{B}$, by using the power diagram of $\mathcal{B}$. The main drawback of the latter approach is that the quality of the approximation deteriorates as the radius of the largest ball in $\mathcal{B}$ increases. We suggest combining both approaches into a new hybrid approach, in which balls larger than a real parameter $r_{\mathrm{c}} > 1$ are replaced by balls with a radius of r_{c}, producing a collection Y of balls that have a radius between 1 and r_{c}, such that the Hausdorff distance between $\cup\mathcal{B}$ and $\cup Y$ is not greater than ε. The persistence diagram of the *power* distance function from $\cup Y$ can serve as an approximation of the desired persistence diagram J. The main theoretical question here is what is the optimal value of r_{c} as a function of $\mathcal{B}$ and ε, i.e., what value of r_{c} minimizes the number of balls in Y. This analysis can produce a new algorithm which can be useful for finding pathways in the complement of molecules in an even more efficient manner.

The pathway diagram is a two-dimensional entity. Instead of the naive approach taken in Part III for reducing the two-dimensional medial axis to a one-dimensional skeleton it is possible to employ techniques like the one introduced by Dey *et al.* [13] to simplify the pathway diagram. This will produce a *skeleton* of the complement of a molecule, which is a collection of one-dimensional curves. Viewing the skeleton of the complement of complex molecules with multiple channels might be insightful for the biologist/biochemist, and using the skeleton for finding pathways might lead to an even more efficient algorithm.

Bibliography

[1] The CGAL Manaul, Release 3.2.1, 2006, `http://www.cgal.org`.

[2] N. Amenta, S. Choi, and R. Kolluri. The power crust, unions of balls, and the medial axis transform. *Computational Geometry: Theory and Applications*, 19(2–3):127–153, 2001.

[3] D. Attali and J.-D. Boissonnat. A linear bound on the complexity of the Delaunay triangulation of points on polyhedral surfaces. *Discrete & Computational Geometry*, 31(3):369–384, 2004.

[4] D. Attali, J.-D. Boissonnat, and H. Edelsbrunner. Stability and computation of medial axes: A state of the art report. In B. H. T. Möller and B. Russell, editors, *Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration*. Springer-Verlag, Mathematics and Visualization, 2007.

[5] D. Attali, J.-D. Boissonnat, and A. Lieutier. Complexity of the Delaunay triangulation of points on surfaces: The smooth case. In *Proceedings of the Symposium on Computational Geometry*, pages 201–210, 2003.

[6] F. Aurenhammer and R. Klein. Voronoi diagrams. In J.-R. Sack and J. Urrutia, editors, *Handbook of Computational Geometry*, pages 201–290. Elsevier Science Publishers B.V. North-Holland, Amsterdam, 2000.

[7] J.-D. Boissonnat and C. Delage. Convex hull and Voronoi diagram of additively weighted points. In *Proceedings of the European Symposium on Algorithms*, pages 367–378, 2005.

[8] J.-D. Boissonnat and M. Yvinec. *Algorithmic Geometry*. Cambridge University Press, UK, 1998. Translated from the French version by H. Brönnimann.

[9] F. Chazal and A. Lieutier. The "Lambda-medial axis". *Graphical Models*, 67(4):304–331, 2005.

[10] F. Chazal and A. Lieutier. Weak feature size and persistent homology: Computing homology of solids in $\mathbb{R}^n$ from noisy data samples. In *Proceedings of the Symposium on Computational Geometry*, pages 255–262, 2005.

[11] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *Proceedings of the Symposium on Computational Geometry*, pages 263–271, 2005.

[12] T. K. F. Da and M. Yvinec. 3D Alpha Shapes. In CGAL Editorial Board, editor, CGAL- *3.2 User and Reference Manual*. 2006. `http://www.cgal.org/Manual/3.2/doc_html/cgal_manual/Alpha_shapes_3/Chapter_main.html`.

[13] T. K. Dey and J. Sun. Defining and computing curve-skeletons with medial geodesic function. In *Proceedings of the Eurographics Symposium on Geometry Processing*, pages 143–152, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association.

[14] T. K. Dey and W. Zhao. Approximate medial axis as a Voronoi subcomplex. In *Proceedings of the Symposium on Solid Modeling and Applications*, pages 356 – 366. ACM Press, 2002.

[15] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[16] H. Edelsbrunner, M. A. Facello, and J. Liang. On the definition and the construction of pockets in macromolecules. *Discrete Applied Mathematics*, 88(1-3):83–102, 1998.

[17] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, 2002.

[18] H. Edelsbrunner and E. P. Mücke. Three-dimensional alpha shapes. *ACM Trans. Graph.*, 13(1):43–72, 1994.

[19] J. Giesen, E. A. Ramos, and B. Sadri. Medial axis approximation and unstable flow complex. In *Proceedings of the Symposium on Computational Geometry*, pages 327–336, 2006.

[20] A. Lieutier. Any open bounded subset of $\mathbb{R}^n$ has the same homotopy type than its medial axis. In *Proceedings of the Symposium on Solid Modeling and Applications*, pages 65–75, 2003.

[21] J. Munkres. *Elements of Algebraic Topology*. Addison-Wesley, 1984.

[22] S. Oudot and J.-D. Boissonnat. Provably good surface sampling and approximation. In *Proceedings of the Symposium on Geometry Processing*, pages 9–19, 2003.

[23] M. Petřek, M. Otyepka, P. Banáš, P. Košinová, J. Koča, and J. Damborský. CAVER: A new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics*, 7(316), 2006.

[24] O. Smart, J. Neduvelil, X. Wang, B. Wallace, and M. Sansom. HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *J. Mol Graphics*, 14(6):354–60,376, 1996.

[25] R. Wein, J. van den Berg, and D. Halperin. The visibility-Voronoi complex and its applications. *Computational Geometry: Theory and Applications*, 36(1):66–78, 2007.

[26] E. W. Weisstein. Spherical code, from mathworld — a wolfram web resource, http://mathworld.wolfram.com/sphericalcode.html, 2000.