63 BIOLOGICAL APPLICATIONS OF COMPUTATIONAL TOPOLOGY

Herbert Edelsbrunner

INTRODUCTION

Structural molecular biology is a relatively recent application area for computational geometry and topology, but one with enormous potential. We currently observe a bi-partition of computational research in this field: the *bio-informatics* branch focuses on strings, which are abstractions of the hereditary information stored in the DNA of living organisms, while the *molecular simulation* branch studies organic molecules in their natural three-dimensional habitat. Perhaps it is not surprising that the application of numerical algorithms is significantly more developed than that of geometric algorithms. One of the goals of this article is to raise the general consciousness for the importance of geometric methods to elucidate the mysterious foundations of our very existence. Another goal is the broadening of what we consider a geometric algorithm. There is plenty of valuable no-mans-land between combinatorial and numerical algorithms, and it seems opportune to explore this land with a computational geometry frame of mind.

63.1 BIO-MOLECULES

GLOSSARY

Central dogma: The proven claim that proteins are created in two steps by transcribing genes to RNA and translating RNA to protein.

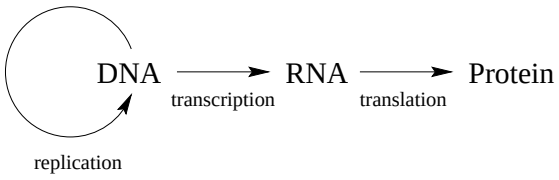

FIGURE 63.1.1
The DNA gets replicated as a whole. Pieces of DNA referred to as genes are transcribed into pieces of RNA, which are then translated into proteins.

DNA: Deoxyribonucleic Acid. The material that carries all hereditary information. A double-stranded helix that encodes information into two anti-parallel sequences of nucleotides.

Replication: Process in which the two strands of DNA are separated and both

0-8493-8524-5/97/$0.00+$.50

strands are complemented to form new double-strands.

Genome: Complete set of genetic material of a living organism. For humans, it is divided into twenty-three *chromosomes*, each a long double-strand of DNA.

Gene: Subsequence of DNA capable of being transcribed to produce a functional RNA molecule.

Transcription: Process in which the two strands of DNA are locally separated and one strand is copied to a piece of RNA.

RNA: Ribonucleic Acid. A single-stranded structure that is chemically almost identical to DNA.

Translation: Process in which a strand of RNA is read by the ribosome and translated into a protein.

Protein: A linear sequence of amino acids connected to each other by peptide bonds.

Amino Acid: Consists of a central carbon atom (C_α) linked to an amino group, a carboxyl group, one hydrogen atom, and a side chain. A *residue* is an amino acid whose $NC_\alpha C$ sequence is linked into the polypeptide chain of a protein.

Protein backbone: Polypeptide chain consisting of repeated $CC_\alpha N$ units. The bond between N and C is rigid, but the bonds connecting C_α to C and C_α to N can be rotated around the connecting edges.

Protein folding: Process in which a polypeptide chain folds up to a usually globular shape that is characteristic for each type of protein.

FROM DNA TO PROTEIN

Organic life is based on a surprising small number of molecule types. Most prominently, we have *DNA*, *RNA* and *protein*. All of them have the simple structure of a linear sequence consisting of a chain or backbone with attached side-chains. DNA and RNA each use an alphabet of only four nucleotides, while proteins use an alphabet of twenty amino acids. As discovered by Watson and Crick [WC53], the natural form of DNA consists of two sequences or strands that are held together by complementary nucleotide pairs. DNA has the ability to *replicate* itself, which is done by separating the two strands and complementing both with the matching strand made from free nucleotides in the surrounding solution. DNA is the memory of evolution that gives coherence to all living species; it forms the material basis of heredity as studied by Mendel in the nineteenth century [Men66]. Apparently, only a small fraction of the DNA in any organism represents used information. The used pieces are the *genes*, which are *transcribed* into RNA in a process similar to replication. RNA remains single-stranded and most of it gets *translated* into protein. This happens in the *ribosome*, which is a large molecular machine made out of proteins and RNA molecules. A single strand of RNA is fed into the ribosome, and each triplet of nucleotides is translated into an amino acid, which is appended to the growing peptide chain. Upon completion, this chain leaves the ribosome as the final protein. This scenario is reminiscent of the Touring machine model of computing, in which information is read from an input tape and the results of the computations are printed on an output tape.

FROM SEQUENCE TO FUNCTION

When the protein leaves the ribosome, it folds up to form a shape that is characteristic for its sequence of amino acids. The proteins constitute the work-force that maintains organic life. Specific proteins fulfill specific functions within the organism, and the particular shape it assumes is crucial:

$$\text{Sequence} \implies \text{Form} \implies \text{Function.}$$

This is why geometry is important in molecular biology. It is essential to learn the shapes of all proteins and to understand what is important about them. Most functions are tied up in the interaction of proteins with each other and with other molecules. The replication of DNA, the transcriptions to RNA, and the translation to protein are but three examples, and each is served by a complicated machine made up of different proteins and RNA molecules. In other words, proteins are the pieces of a huge three-dimensional dynamic puzzle whose solution requires, upon others things, a good understanding of the involved shapes. A major difficulty in the field of molecular biology is the miniscule scale of space and time at which the processes happen. The actors and their scripts are complicated and observations are indirect. The experimental work is generally complemented by computational simulations, which are referred to as theoretical work in this area.

63.2 GEOMETRIC MODELS

Proteins are complicated objects, which have been abstracted into a number of different models emphasizing different aspects of their behavior. We may think of them as curves in space modeling the backbone, or as a collection of balls or spheres representing it at the level of individual atoms.

GLOSSARY

Space-filling diagram: Model that represents a protein by the space it occupies. Most commonly, each atom is represented by a ball (a solid sphere) and the protein is the union of these balls.

Van der Waals surface: Boundary of space-filling diagram defined as the union of balls with van der Waals radii. The sizes of these balls are chosen to reflect the transition from an attractive to a repulsive van der Waals force.

Solvent accessible surface: Boundary of space-filling diagram in which each van der Waals ball is enlarged by the radius of the solvent sphere. Alternatively, it is the set of centers of solvent spheres that touch but do not otherwise intersect the van der Waals surface.

Molecular surface: Boundary of the portion of space inaccessible to the solvent. It is obtained by rolling the solvent sphere about the van der Waals surface.

Power distance: Square length of tangent line segment from a point x to a sphere with center z and radius r. It is also referred to as the *weighted square distance* and formally defined as $\|x - z\|^2 - r^2$.

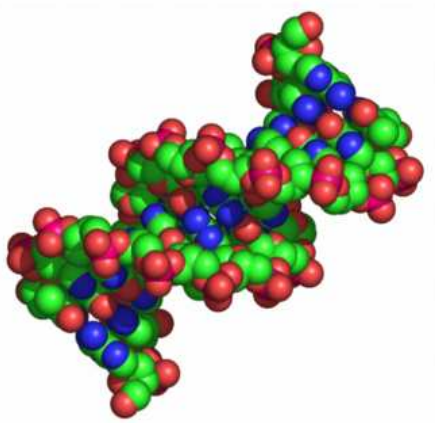

FIGURE 63.2.1
A short segment of a DNA double-helix in space-filling representation. DNA uses an alphabet of four nucleotides: adenine (A), guanine (G), cytosine (C) and thymine (T). In the picture they are barely visible since the nucleotides are packed in the middle, using hydrogen bonds to hold the strands together.

Voronoi diagram: Decomposition of space into convex polyhedra. Each polyhedron belongs to a sphere in a given collection and consists of all points for which this sphere minimizes the power distance. This decomposition is also known as the *power diagram* and the *weighted Voronoi diagram*.

Delaunay triangulation: Dual to the Voronoi diagram. For generic collections of spheres, it is a simplicial complex consisting of tetrahedra, triangles, edges and vertices. This complex is also known as the *regular triangulation*, the *coherent triangulation* and the *weighted Delaunay triangulation*.

Dual complex: Dual to the Voronoi decomposition of a union of balls. It is a subcomplex of the Delaunay triangulation.

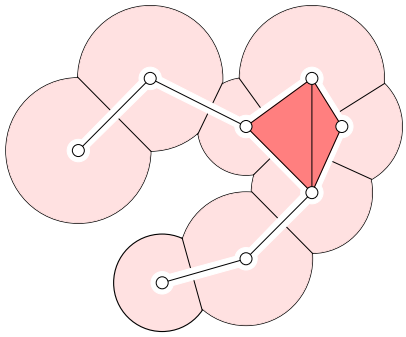

FIGURE 63.2.2
Each Voronoi polygon intersects the union of disks in a convex set, which is the intersection with its defining disk. The drawing shows the Voronoi decomposition of the union and the dual complex superimposed.

Growth model: Rule for growing all spheres in a collection continuously and simultaneously. The rule that increases the square radius r^2 to $r^2 + t$ at time t keeps the Voronoi diagram invariant at all times.

Alpha complex: The dual complex at time $t = \alpha$ for a collection of spheres that grow while keeping the Voronoi diagram invariant. The *alpha shape* is the underlying space of the alpha complex.

Filtration: Nested sequence of complexes. The prime example in this paper is the sequence of alpha complexes.

SPACE-FILLING DIAGRAMS

Our starting point is the *van der Waals force*, which is based on quantum mechanical effects. At short range up to a few Angstrom, the force is attractive but significantly weaker than covalent or ionic bonds. At very short range, the force is strongly repulsive. We may assign *van der Waals radii* to the atoms such that the force changes from attractive to repulsive when the corresponding spheres touch [GR01]. The *van der Waals surface* is the boundary of the space-filling diagram made up of the balls with van der Waals radii. In the 1970s, Richards and collaborators extended this idea to capture the interaction of a protein with the surrounding solvent [LR71, Ric77]. The *solvent accessible surface* is the boundary of the space-filling diagram in which the balls are grown by the radius of the sphere that models a single solvent molecule. Usually the solvent is water represented by a sphere of radius 1.4 Angstrom. The *molecular surface* is obtained by rolling the solvent sphere over the van der Waals surface and filling in the inaccessible crevices and cusps. This surface is sometimes referred to as the *Connolly surface*, named after the creator of the first software representing this surface by a collection of dots [Con83].

DUAL STRUCTURES

We complement the space-filling representations of proteins with geometrically dual structures. A major advantage of these dual structures is their computational convenience. We begin by introducing the *Voronoi diagram* of a collection of balls or spheres, which decomposes the space into convex polyhedra [Vor07]. Next we intersect the union of balls with the Voronoi diagram and obtain a decomposition of the space-filling diagram into convex *cells*. Indeed, these cells are the intersections of the balls with their corresponding Voronoi polyhedra. The *dual complex* is the collection of simplices that express the intersection pattern between the cells: we have a vertex for every cell, an edge for every pair of cells that share a common facet, a triangle for every triplet of cells that share a common edge, and a tetrahedron for every quadruplet of cells that share a common point [EKS83, EM94]. This exhausts all possible intersection patterns in the assumed generic case. We get a natural embedding if we use the sphere centers as the vertices of the dual complex.

GROWTH MODEL

One and the same Voronoi diagram corresponds to more than just one collection of spheres. For example, if we grow the square radius r_i^2 of the i-th sphere to $r_i^2 + t$, for every i, we get the same Voronoi diagram. Think of t as time parametrizing this particular growth model of the spheres. While the Voronoi diagram remains fixed, the dual complex changes. The cells in which the balls intersect the Voronoi polyhedra grow monotonically with time, which implies that the dual complex can acquire but not lose simplices. We thus get a nested sequence of dual complexes,

$$\emptyset \;=\; K_0 \;\subseteq\; K_1 \;\subseteq\; \ldots \;\subseteq\; K_m \;=\; D,$$

which begins with the empty complex at time $t = -\infty$ and ends with the Delaunay triangulation [Del34], at time $t = \infty$. We refer to this sequence as a *filtration* of

the Delaunay triangulation and think of it as the dual representation of the protein at all scale levels.

63.3 MESHING

We introduce yet another surface bounding a space-filling diagram of sorts. The *molecular skin* is the boundary of the union of infinitely many balls. Besides the balls with van der Waals radii representing the atoms, we have balls interpolating between them that give rise to blending patches and altogether to a tangent continuous surface. The molecular skin is rather similar in appearance to the molecular surface but uses hyperboloids instead of tori to blend between the spheres [Ede99]. The smoothness of the surface permits a mesh whose triangles are all approximately equiangular [CDES01]. Applications of this mesh include the representation of proteins for visualization purposes and the solution of differential equations defined over the surface by finite element and other numerical methods.

GLOSSARY

Molecular skin: Surface of a molecule that is geometrically similar to the molecular surface but uses hyperboloid instead of torus patches for blending.

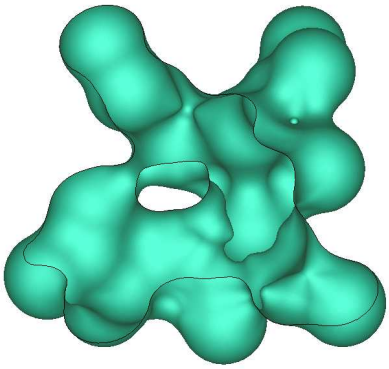

FIGURE 63.3.1
Cut-away view of the skin of a small molecule. We see a blend of sphere and hyperboloid patches. The surface is inside-outside symmetric: it can be defined by a collection of spheres on either of its two sides.

Mixed complex: Decomposition of space into shrunken Voronoi polyhedra, shrunken Delaunay tetrahedra and shrunken products of corresponding Voronoi polygons and Delaunay edges as well as Voronoi edges and Delaunay triangles. It decomposes the skin surface into sphere and hyperboloid patches.

Maximum normal curvature: The larger absolute value $\kappa(x)$ of the two principal curvatures at a point x of the surface.

ε-sampling: Collection of points S on the molecular skin M such that every point $x \in M$ has a point $u \in S$ at distance $\|x - u\| \leq \varepsilon/\kappa(x)$.

Restricted Delaunay triangulation: Dual to the restriction of the (three-dimensional) Voronoi diagram of S to the molecular skin M.

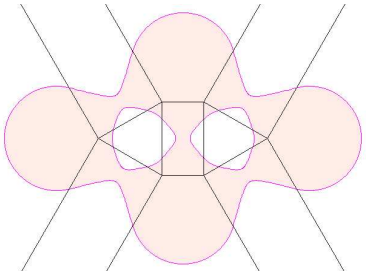

FIGURE 63.3.2
The skin curve defined by four circles in the plane. The mixed complex decomposes the curve into pieces of circles and hyperbolas.

Closed ball property: The condition that every cell in a complex is a topologically simple closed ball of the appropriate dimension.

Morphing: Process of deforming one given shape into another.

Shape space: Locally parametrized space of shapes. The prime example in this paper is the $(k-1)$-dimensional space generated by k shapes each specified by a collection of spheres in $\mathbb{R}^3$.

TRIANGULATION

The molecular skin has geometric properties that can be exploited to construct a numerically high-quality mesh and to maintain that mesh during deformation. The most important of these is the continuity of the *maximum normal curvature* function $\kappa : M \to \mathbb{R}$. To define it, consider the 1-parameter family of geodesics passing through x and let $\kappa(x)$ be the maximum of their curvatures at x. We use this function to guide the local density of the points distributed over M that are used as vertices of the mesh. Given such a collection of points S, we construct a mesh using their Voronoi diagram restricted to M. The polyhedra decompose the surface into patches, and the mesh is constructed as the dual of that decomposition [Che93]. As proved in [ES97], the mesh is homeomorphic to the surface if the pieces of the restricted Voronoi diagram are topologically simple sets of the appropriate dimensions. In other words, the intersection of each Voronoi polyhedron, polygon, or edge with M is either empty or a topological disk, and interval, or a single point. Because of the smoothness of M, this topological property is implied if the points form an ε-sampling, with $\varepsilon = 0.279$ or smaller [CDES01].

DEFORMATION AND SHAPE SPACE

The variation of the maximum normal curvature function can be bounded by the one-sided Lipschitz condition $|1/\kappa(x) - 1/\kappa(y)| \leq \|x - y\|$, where the distance is measured in $\mathbb{R}^3$. The continuity over $\mathbb{R}^3$ and not just over M is crucial when it comes to maintaining the mesh while changing the surface. This leads us to the topic of deformations and shape space. The latter is constructed as a parametrization of the deformation process. The deformation from a shape A_0 to another shape A_1 can be written as $\lambda_0 A_0 + \lambda_1 A_1$, with $\lambda_1 = 1 - \lambda_0$. Accordingly, we may think of the unit interval as a one-dimensional shape space. We can generalize this to a k-dimensional shape space as long as the different ways of arriving at $(\lambda_0, \lambda_1, \ldots, \lambda_k)$, with $\sum \lambda_i = 1$ and $\lambda_i \geq 0$ for all i, all give the same shape $A = \sum \lambda_i A_i$. How to

define deformations such that this is indeed the case is explained in [CEF01].

63.4 CONNECTIVITY AND SHAPE FEATURES

Protein connectivity is often understood in terms of its covalent bonds, in particular along the backbone. In this section, we discuss a different notion, namely the topological connectivity of the space assigned to a protein by its space-filling diagram. We mention *homeomorphisms*, *homotopies*, *homology groups* and *Euler characteristics*, which are common topological concepts used to define and talk about connectivity. Of particular importance are the homology groups and their ranks, the *Betti numbers*, as they lend themselves to efficient algorithms. In addition to computing the connectivity of a single space-filling diagram, we study how the connectivity changes when the balls grow. The sequence of space-filling diagrams obtained this way corresponds to the filtration of dual complexes introduced earlier. We use this filtration to define basic shape features, such as pockets in proteins and interfaces between complexed proteins and molecules.

GLOSSARY

Topological equivalence: Equivalence relation between topological spaces defined by *homeomorphisms*, which are continuous bijections with continuous inverses.

Homotopy equivalence: Weaker equivalence relation between topological spaces $\mathbb{X}$ and $\mathbb{Y}$ defined by maps $f : \mathbb{X} \to \mathbb{Y}$ and $g : \mathbb{Y} \to \mathbb{X}$ whose compositions $g \circ f$ and $f \circ g$ are homotopic to the identities on $\mathbb{X}$ and on $\mathbb{Y}$.

Deformation retraction: A homotopy between the identity on $\mathbb{X}$ and a retraction of $\mathbb{X}$ to $\mathbb{Y} \subseteq \mathbb{X}$ that leaves $\mathbb{Y}$ fixed. The existence of the deformation implies that $\mathbb{X}$ and $\mathbb{Y}$ are homotopy equivalent.

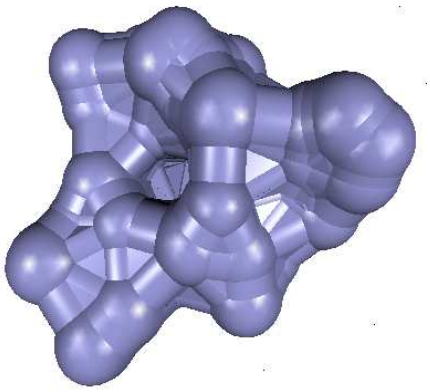

FIGURE 63.4.1
Snap-shot during the deformation retraction of the space-filling representation of gramicidin to its dual complex. The spheres shrink to vertices while the intersection circles become cylinders that eventually turn into edges.

Homology groups: Quotients of cycle groups and their boundary subgroups. There is one group per dimension. The *k-th Betti number*, β_k, is the rank of the

k-th homology group.

Euler characteristic: The alternating sum of Betti numbers: $\chi = \sum_{k \geq 0} (-1)^k \beta_k$.

Voids: Bounded connected components of the complement. In this paper, we are primarily interested in voids of space-filling diagrams embedded in $\mathbb{R}^3$.

Pockets: Maximal regions in the complement of a space-filling diagram that become voids before they disappear. Here we assume the growth model that preserves the Voronoi diagram of the spheres.

Persistent homology groups: Quotients of the cycle groups at some time t and their boundary subgroups a later time $t + p$. The ranks of these groups are the *persistent Betti numbers*.

Protein complex: Two or more docked proteins. A complex can be represented by a single space-filling diagram of colored balls.

Molecular interface: Surface consisting of bi-chromatic Voronoi polygons that separate the proteins in the complex. The surface is usually retracted to the region in which the proteins are in close contact.

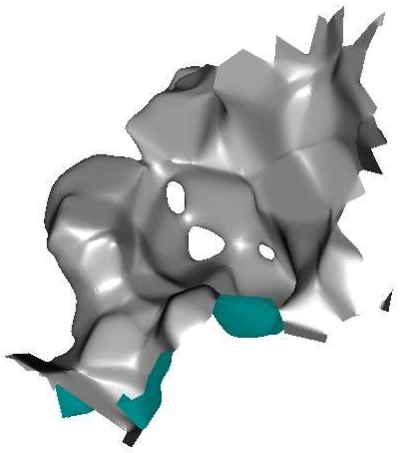

FIGURE 63.4.2
Molecular interface of the neurotoxic vipoxin complex. The surface has non-zero genus, which is unusual. In this case, we have genus equal to three, which implies the existence of three loops from each protein that are linked with each other. The linking might explain the unusually high stability of the complex, which remains for years in solution. The piecewise linear surface has been smoothed to improve visibility.

CLASSIFICATION

The connectivity of topological spaces is commonly discussed by forming equivalence classes of spaces that are connected the same way. Sameness may be defined as being homeomorphic, being homotopy equivalent, having isomorphic homology groups, or having the same Euler characteristic. In this sequence, the classification gets progressively coarser but also easier to compute. Homology groups seem to be a good compromise as they capture a great deal of connectivity information and have fast algorithms. The classic approach to computing homology groups is algebraic and considers the incidence matrices of adjacent dimensions. Each matrix is reduced to *Smith normal form* using a Gaussian elimination like reduction algorithm. The ranks and torsion coefficients of the homology groups can be read off directly from the reduced matrices [Mun84]. Depending on which coefficients we

use and how exactly we reduce, the running time can be anywhere between cubic in the number of simplices and exponential or worse.

INCREMENTAL ALGORITHM

Space-filling diagrams are embedded in $\mathbb{R}^3$ and enjoy properties that permit much faster algorithms. To get started, we use the existence of a deformation retraction from the space-filling diagram to the dual complex, which implies that the two have isomorphic homology groups [Ede95]. The embedding in $\mathbb{R}^3$ prohibits non-zero torsion coefficients [AH35]. We therefore limit ourselves to Betti numbers, which we compute incrementally, by adding one simplex at a time in an order that agrees with the filtration of the dual complexes. Upon adding a k-dimensional simplex σ, the k-th Betti number goes up by one if σ belongs to a k-cycle, and the $(k-1)$-st Betti number goes down by one if σ does not belong to a k-cycle. The two cases can be distinguished in a time that, for all practical purposes, is constant per operation, leading to an essentially linear time algorithm for computing the Betti numbers of all complexes in the filtration [DE95].

PERSISTENCE

To get a handle on the stability of a homology class, we observe that the simplices that create cycles can be paired with the simplices that destroy cycles. The *persistence* is the time lag between the creation and the destruction [ELZ02]. The idea of pairing lies also at the heart of two types of shape features relevant in the study of protein interactions. A *pocket* in a space-filling diagram is a portion of the outside space that becomes a void before it disappears [EFL98, Kun92]. It is represented by a triangle-tetrahedron pair: the triangle creates a void and the tetrahedron is the last piece that eventually fills that same void. The *molecular interface* consists of all bi-chromatic Voronoi polygons of a protein complex. To identify the essential portions of this surface, we again observe how voids are formed and retain the bi-chromatic polygons inside pockets while removing all others [BER03]. A different geometric formalization of the same biochemical concept can be found in [VBR+95]. Preliminary experiments suggest that the combination of molecular interfaces and the idea of persistence can be used to predict the hot-spot residues in protein-protein interactions [Wel96].

63.5 DENSITY MAPS

Continuous maps over manifolds arise in a variety of settings within structural molecular biology. One is *x-ray crystallography*, which is the most common method for determining the three-dimensional structure of proteins [BJ76, Rho00]. While casting x-rays on a crystal of purified protein, we observe defraction patterns from which the electron density of the protein can be obtained via an inverse Fourier transform. Another setting is *molecular mechanics*, whose central object is the force field that drives atomic motions. We may, for example, be interested in the electrostatic potential induced by a protein and visualize it as a density map over

three-dimensional space or over a surface embedded in that space. As a third setting, we mention the *protein docking* problem. Given two proteins, or a protein and a ligand, we try to fit protrusions of one into the cavities of the other [Con86]. We make up continuous functions related to the shapes of the surfaces and identify protrusions and cavities as local extremes of these functions. Morse theory is the natural mathematical framework for studying these maps [Mil63, Mat02].

GLOSSARY

Morse function: Generic smooth map over a manifold, $f : \mathbb{M} \to \mathbb{R}$. In particular, the genericity assumption includes that all critical points are non-degenerate and have different function values.

Gradient, Hessian: The vector of first derivatives and the matrix of second derivatives.

Critical point: Point at which the gradient of f vanishes. It is *non-degenerate* if the Hessian is invertible. The *index* of a non-degenerate critical point is the number of negative eigenvalues of the Hessian.

Integral line: Maximal curve whose velocity vectors agree with the gradient of the Morse function. Two integral lines are either disjoint or the same.

Stable manifold: Union of integral lines that converge to the same critical point. We get *unstable manifolds* if we negate f and thus effectively reverse the gradient.

Morse-Smale complex: Collection of cells obtained by intersecting stable with unstable manifolds. We require that f is a *Morse-Smale function* that satisfies the additional genericity assumption that these intersections are transversal.

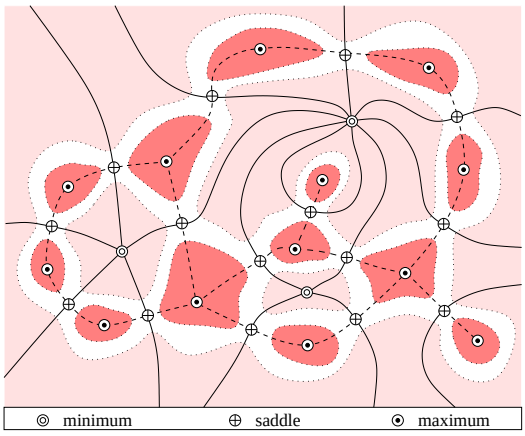

FIGURE 63.5.1
Portion of the Morse-Smale complex of a Morse-Smale function over a 2-manifold. The solid stable 1-manifolds and the dashed unstable 1-manifolds are shows together with two dotted level sets. Observe that all two-dimensional regions of the complex are quadrangular.

Cancellation: Local change of the Morse function that removes a pair of critical points. Their indices are necessarily contiguous.

CRITICAL POINTS

Classic Morse theory applies only to generic smooth maps over manifolds, $f : M \to \mathbb{R}$. Maps that arise in practice are rarely smooth and generic, or more precisely, the information we are able to collect about maps is rarely enough to go beyond a piecewise linear representation. To illustrate this point, we discuss critical points, which for smooth functions are characterized by a vanishing gradient: $\nabla f = 0$. If we draw a small circle around a non-critical point u on a 2-manifold, we get one arc along which the function takes on values less than $f(u)$ and a complementary arc along which the function is greater than or equal to $f(u)$. Call the former arc the *lower link* of u. We get different lower links for critical points: the entire circle for a *minimum*, two arcs for a *saddle*, and the empty set for a *maximum*. A typical representation of a piecewise linear map is a triangulation with function values specified at the vertices and linearly interpolated over the edges and triangles. The lower link of a vertex can still be defined and the criticality of the vertex can be determined from the topology of the lower link [Ban67].

MORSE-SMALE COMPLEXES

In the smooth case, each critical point defines a *stable manifold* of points that converge to it by following the gradient flow. Symmetrically, if defines an *unstable manifold* of points that converge to it by following the reversed gradient flow. These manifolds define decompositions of the manifold into simple cells [Tho49]. Extensions of these ideas to construct similar cell decompositions of manifolds with piecewise linear continuous functions can be found in [EHZ01]. In practice, it is essential to be able to simplify these decompositions, which can be done by cancelling critical points in pairs in the order of increasing persistence [ELZ02].

63.6 MATCH AND FIT

Proteins can be similar in a variety of ways: they can have similar residue sequences, they can have backbones that are layed out similarly in space, and they can have similar shapes after folding. The first two notions are important to gain insight into the evolutionary development of proteins. The corresponding computational problems are sequence alignment and structure alignment. The question of shape similarity, and in particular of partial shape similarity, is relevant to understand the interaction between proteins and their substrates, which can be proteins or other molecules. Indeed, many interactions seem to require a high degree of partial shape complementarity, which we interpret as a high degree of partial shape similarity between the protein and the complement of its substrate.

GLOSSARY

Rigid motion: Orientation and distance preserving motion. The primary example in this paper are rigid motions of three-dimensional space, $\mu : \mathbb{R}^3 \to \mathbb{R}^3$. Each rigid motion can be decomposed into a rotation followed by a translation.

RMSD: Root mean square distance. Root of the average square distance between two sets of points with a given bijection.

Dynamic programming: Algorithmic paradigm which computes the optimum from pre-computed optimal solutions to sub-problems.

Sequence alignment: Collection of monotonically increasing maps to the integers, one per sequence. Each letter gets either matched or skipped.

Structure alignment: Collection of monotonically increasing maps to the integers, one per chain of points modeling a protein backbone.

Protein docking: Process in which a protein forms a complex with another molecule. The complex usually exists only temporarily and facilitates an interaction between the molecules.

STRUCTURE ALIGNMENT

There are two approaches to structure alignment. The first compares the matrices of internal sequences between the points [HS93]. We only discuss the second approach, which is a direct extension of the work on sequence alignments in bioinformatics [Gus97]. Instead of letters representing residues, we align points in space, which are the centers of the alpha carbon atoms along the two backbones. For decomposable score functions, we can find the optimal alignment with dynamic programming in time that is quadratic in the the number of points. One such function suggested in [SLL93] penalizes unmatched points and for every matched pair (u_i, v_j) adds

$$\delta(u_i, v_j) \quad = \quad \frac{100}{5 + \|u_i - v_j\|^2}$$

to the score. The dynamic programming approach works only for two fixed sequences, and the six degrees of freedom we gain by allowing rigid motions complicate matters considerably. Nevertheless, it is possible to compute an approximation to the optimal alignment in time that is polynomial in the number of points and the tolerated error [KL03].

RIGID MOTIONS

Let $u_1, u_2, \ldots, u_n$ and $v_1, v_2, \ldots, v_n$ be two sequences of points in $\mathbb{R}^3$. For a given rigid motion μ, the *root mean square distance* between the sequences is

$$f(\mu) \quad = \quad \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|u_i - \mu(v_i)\|^2}.$$

It is perhaps surprising that the dependence of f on μ can be expressed by a quadratic function which, in the generic case, has a unique local minimum. To describe the minimizing rigid motion, we decompose it into a translation followed by a rotation. Assuming the centroid of the u_i lies at the origin, the optimum translation moves the centroid of the v_i to the origin, and the optimum rotation can be computed by solving a straightforward eigenvalue problem. One of the earliest references to this result is Kabsch [Kab78]. A lucid description of the proof using quaternions to represent rotations can be found in Horn [Hor87].

PROTEIN DOCKING

A good local geometric fit is a necessary condition for a complex between two or more proteins to be formed. There are however additional factors, such as electrostatic and hydrophobic forces. To further complicate the issue, proteins are somewhat flexible and can sometimes avoid otherwise prohibitive steric clashes [ESM01]. Taking all these factors into account seems prohibitive and most computational approaches to protein docking explore the space of rigid motions using relatively simple geometric score functions [HMWN02]. An example is the number of atoms in close but not too close distance from each other. The space of rigid motions is six-dimensional and exploring it is time-consuming, even with simple score functions. The idea of Connolly to use critical points of Morse functions to identify motions [Con86] seems promising but not yet fully explored. It is usually combined with geometric hashing to enumerate the motions suggested by the critical point patterns [NLWN94].

63.7 MEASURES AND DERIVATIVES

Computing the volume and the surface area of a space-filling diagram are two of the most fundamental means to characterize the geometry of a protein. To mention a specific application, we consider the computation of the *solvation energy*, which is central in the simulation of folding and docking processes. Many simulations use *implicit solvent models* and describe the hydrophobic part of the solvation energy as a weighted sum of the accessible surface area or, alternatively, as a weighted sum of volumes. The weights are experimentally determined *solvation parameters* that assess the contributions of different atom types to the hydrophobic term [EM86]. A molecular dynamics simulation requires the weighted area or volume and their derivatives in order to estimate the contribution of the hydrophobic term to the energy that drives the process.

GLOSSARY

Indicator function: Maps a point x to 1 if $x \in P$ and to 0 if $x \notin P$, where P is some fixed set. In this paper, we are interested in convex polyhedra P and use the alternating sum of faces visible from x as indicator.

Inclusion-exclusion: Principle used to compute the volume of a union of bodies as the alternating sum of volumes of k-fold intersections, for $k \geq 1$.

Stereographic projection: Mapping of the 3-sphere minus a point to the three-dimensional Euclidean space. The map preserves spheres and angles.

Angle of revolution: The normalized k-dimensional volume of a subset of the k-dimensional sphere.

Atomic solvation parameters: Experimentally determined numbers that assess the hydrophobicity of different atoms.

Weighted volume: Volume of a space-filling diagram in which the contribution of each individual ball is weighted by its atomic solvation parameter. Also a function $V : \mathbb{R}^{3n} \to \mathbb{R}$ obtained by parametrizing a space-filling diagram by the coordinates of its n ball centers.

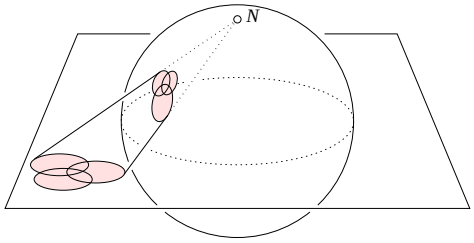

FIGURE 63.7.1
Stereographic projection from the north-pole. The preimage of a circle in the plane is a circle on the sphere, which is the intersection of the sphere with a plane. By extension, the preimage of a union of disks is the intersection of the sphere with the complement of a convex polyhedron.

Weighted volume derivative: The linear map $\mathrm{D}V_{\mathbf{z}} : \mathbb{R}^{3n} \to \mathbb{R}$ defined by $\mathrm{D}V_{\mathbf{z}}(\mathbf{t}) = \langle \mathbf{v}, \mathbf{t} \rangle$, where $\mathbf{z} \in \mathbb{R}^{3n}$ specifies the space-filling diagram, $\mathbf{t} \in \mathbb{R}^{3n}$ lists the coordinates of the motion vectors and $\mathbf{v} = \nabla V(\mathbf{z})$ is the gradient of V at $\mathbf{z}$. It is also the map $\mathrm{D}V : \mathbb{R}^{3n} \to \mathbb{R}^{3n}$ that maps $\mathbf{z}$ to $\mathbf{v}$.

GEOMETRIC INCLUSION-EXCLUSION

Work on computing the volume and the area of a space-filling diagram $F = \bigcup_i B_i$ can be divided into approximate [Row63] and exact methods [Ric74]. According to the principle of inclusion-exclusion, the volume of F can be expressed by an alternating sum of volumes of intersections:

$$\operatorname{vol} F \quad = \quad \sum_{\Lambda} (-1)^{\operatorname{card}\Lambda + 1} \operatorname{vol} \bigcap_{i \in \Lambda} B_i,$$

where Λ ranges over all non-empty subsets of the index set. The size of this formula is exponential in the number of balls, and the individual terms can be quite complicated. Most of the terms are however redundant and a much smaller formula based on the dual complex K of the space-filling diagram F has be given [Ede95]:

$$\operatorname{vol} F \quad = \quad \sum_{\sigma \in K} (-1)^{\dim \sigma} \operatorname{vol} \bigcap \sigma,$$

where $\bigcap \sigma$ denotes the intersection of the $\dim \sigma + 1$ balls whose centers are the vertices of σ. The proof is based on the Euler formula for convex polyhedra and uses stereographic projection to relate the space-filling diagram in $\mathbb{R}^3$ with a convex polyhedron in $\mathbb{R}^4$. Precursors of this result include the existence proof of a polynomial size inclusion-exclusion formula [Kra78] and the presentation of such a formula using the simplices in the Delaunay triangulation [NW92]. We note that it is straightforward to modify the formula to get the weighted volume: decompose the terms $\operatorname{vol} \bigcap \sigma$ into the portions within the Voronoi cells of the participating balls and weigh each portion accordingly.

DERIVATIVES

The relationship between the weighted and the unweighted volume derivatives is less direct than that between the weighted and the unweighted volumes. Just to state the formula for the weighted volume derivative requires more notation that we are willing to introduce here. Instead, we describe the two geometric ingredients, both of which can be computed by geometric inclusion-exclusion [EK02]. The first

ingredient is the area of the portion of the disk spanned by the circle of two intersecting spheres that belongs to the Voronoi diagram. This facet is the intersection of the disk with the corresponding Voronoi polygon. The second ingredient is the weighted average vector from the center of the disk to the boundary of said facet. The weight is the infinitesimal contribution to the area as we rotated the vector to sweep out the facet.

63.8 SOURCES AND RELATED MATERIAL

FURTHER READING

For background reading in **algorithms** we recommend: [CLR90], which is a comprehensive introduction to combinatorial algorithms; [Gus97], which is an algorithms text specializing in bioinformatics; [Str93], which is an introduction to linear algebra; and [Sch02], which is a numerical algorithms text in molecular modeling.

For background reading in **geometry** we recommend: [Ped88], which is a geometry text focusing on spheres; [Nee97], which is a lucid introduction to geometric transformations; [FT72], which studies packing and covering in two and three dimensions; and [Ede01], which is an introduction to computational geometry and topology, focusing on Delaunay triangulations and mesh generation.

For background reading in **topology** we recommend: [Ale98], which is a compilation of three classical texts in combinatorial topology; [Gib77], which is a very readable introduction to homology groups; [Mun84], which is a comprehensive text in algebraic topology; and [Mat02], which is a recent introduction to Morse theory.

For background reading in **biology** we recommend: [ABL+94], which is a basic introduction to molecular biology on the cell level; [Str88], which is a fundamental text in biochemistry; and [Cre93], which is an introduction to protein sequences, structures and shapes.

RELATED CHAPTERS

Chapter 2: Packing and covering
Chapter 20: Voronoi diagrams and Delaunay triangulations
Chapter 22: Triangulations
Chapter 28: Computational topology

REFERENCES

[ABL+94] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts and J. D. Watson. *Molecular Biology of the Cell.* Garland, New York, 1994

[Ale98] P. S. Alexandrov. *Elementary Concepts of Topology.* Dover, New York, 1961.

[AH35] P. S. Alexandrov and H. Hopf. *Topologie I.* Julius Springer, Berlin, Germany, 1935.

[BER03] Y.-E. Ban, H. Edelsbrunner and J. Rudolph. A definition of interfaces for protein oligomers. Manuscript, Duke Univ., Durham, North Carolina, 2002.

[Ban67] T. F. Banchoff. Critical points and curvature for embedded polyhedra. *J. Differential Geometry* **1** (1967), 245–256.

[BJ76] T. Blundell and L. Johnson. *Protein Crystallography.* Academic Press, New York, 1976.

[CDES01] H.-L. Cheng, T. K. Dey, H. Edelsbrunner and J. Sullivan. Dynamic skin triangulation. *Discrete Comput. Geom.* **25** (2001), 525–568.

[CEF01] H.-L. Cheng, H. Edelsbrunner and P. Fu. Shape space from deformation. *Comput. Geom. Theory Appl.* **19** (2001), 191–204.

[Che93] L. P. Chew. Guaranteed-quality mesh generation for curved surfaces. *In* "Proc. 9th Ann. Sympos. Comput. Geom., 1993", 274–280.

[Con83] M. L. Connolly. Analytic molecular surface calculation. *J. Appl. Crystallogr.* **6** (1983), 548–558.

[Con86] M. L. Connolly. Measurement of protein surface shape by solid angles. *J. Mol. Graphics* **4** (1986), 3–6.

[CLR90] T. H. Cormen, C. E. Leiserson and R. L. Rivest. *Introduction to Algorithms.* MIT Press, Cambridge, Massachusetts, 1990.

[Cre93] T. E. Creighton. *Proteins.* Freeman, New York, 1993.

[Del34] B. Delaunay. Sur la sphère vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk* **7** (1934), 793–800.

[DE95] C. J. A. Delfinado and H. Edelsbrunner. An incremental algorithm for Betti numbers of simplicial complexes on the 3-sphere. *Comput. Aided Geom. Design* **12** (1995), 771–784.

[Ede95] H. Edelsbrunner. The union of balls and its dual shape. *Discrete Comput. Geom.* **13** (1995), 415–167.

[Ede99] H. Edelsbrunner. Deformable smooth surface design. *Discrete Comput. Geom.* **21** (1999), 87–115.

[Ede01] H. Edelsbrunner. *Geometry and Topology for Mesh Generation.* Cambridge Univ. Press, England, 2001.

[EFL98] H. Edelsbrunner, M. A. Facello and J. Liang. On the definition and the construction of pockets in macromolecules. *Discrete Appl. Math.* **88** (1998), 83–102.

[EHZ01] H. Edelsbrunner, J. Harer and A. Zomorodian. Hierarchy of Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete Comput. Geom.*, to appear.

[EKS83] H. Edelsbrunner, D. G. Kirkpatrick and R. Seidel. On the shape of a set of points in the plane. *IEEE Trans. Inform. Theory* **IT-29** (1983), 551–559.

[EK02] H. Edelsbrunner and P. Koehl. The weighted volume derivative of a space-filling diagram. Manuscript, Duke Univ. Durham, North Carolina, 2002.

[ELZ02] H. Edelsbrunner, D. Letscher and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.* **28** (2002), 511–533.

[EM94] H. Edelsbrunner and E. P. Mücke. Three-dimensional alpha shapes. *ACM Trans. Graphics* **13** (1994), 43–72.

[ES97] H. Edelsbrunner and N. R. Shah. Triangulating topological spaces. *Internat. J. Comput. Geom. Appl.* **7** (1997), 365–378.

[EM86] D. Eisenberg and A. McLachlan. Solvation energy in protein folding and binding. *Nature (London)* **319** (1986), 199–203.

[ESM01] A. H. Elcock, D. Sept and J. A. McCammon. Computer simulation of protein-protein interaction. *J. Phys. Chem.* **105** (2001), 1504–1518.

[FT72] L. Fejes Tóth. *Lagerungen in der Ebene, auf der Kugel und im Raum.* Second edition, Springer-Verlag, New York, 1972.

[GR01] M. Gerstein and F. M. Richards. Protein geometry: distances, areas, and volumes. Chapter 22 in *The International Tables for Crystallography, Vol. F*, M. G. Rossmann and E. Arnold (eds.), Kluwer, Dordrecht, the Netherlands, 2001, 531–539.

[Gib77] P. J. Giblin. *Graphs, Surfaces, and Homology. An Introduction to Algebraic Topology.* Chapman and Hall, London, 1977.

[Gus97] D. Gusfield. *Algorithms on Strings, Trees, and Sequences.* Cambridge Univ. Press, England, 1997.

[HMWN02] I. Halperin, B. Mao, H. Wolfson and R. Nussinov. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins* **47** (2002), 409–443.

[HS93] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233** (1993), 123–138.

[Hor87] B. K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Amer. A* **4** (1987), 629–642.

[Kab78] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A* **34** (1978), 827–828.

[KL03] R. Kolodny and N. Linial. Approximate protein structural alignment in polynomial time. Manuscript, Stanford Univ., Stanford, California, 2002.

[Kra78] K. W. Kratky. The area of intersection of n equal circular disks. *J. Phys. A: Math. Gen.* **11** (1978), 1017–1024.

[Kun92] I. D. Kuntz. Structure-based strategies for drug design and discovery. *Science* **257** (1992), 1078–1082.

[LR71] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55** (1971), 379–400.

[Mat02] Y. Matsumoto. *An Introduction to Morse Theory.* American Mathematical Society, Providence, Rhodes Island, 2002.

[Men66] G. Mendel. Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines*, Abhandlungen, Brünn **4** (1866), 3–47.

[Mil63] J. Milnor. *Morse Theory.* Princeton Univ. Press, New Jersey, 1963.

[Mun84] J. R. Munkres. *Elements of Algebraic Topology.* Addison-Wesley, Redwood City, California, 1984.

[NW92] D. Q. Naiman and H. P. Wynn. Inclusion-exclusion Bonferroni identities and inequalities for discrete tube-like problems via Euler characteristics. *Ann. Statist.* **20** (1992), 43–76.

[Nee97] T. Needham. *Visual Complex Analysis.* Clarendon Press, Oxford, England, 1997.

[NLWN94] R. Norel, S. L. Lin, H. Wolfson and R. Nussinov. Shape complementarity at protein-protein interfaces. *Biopolymers* **34** (1994), 933–940.

[Ped88] D. Pedoe. *Geometry. A Comprehensive Course.* Dover, New York, 1988.

[Rho00] G. Rhodes. *Crystallography Made Crystal Clear.* Second edition, Academic Press, Sand Diego, California, 2000.

[Ric74] F. M. Richards. The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82** (1974), 1–14.

[Ric77] F. M. Richards. Areas, volumes, packing and protein structures. *Ann. Rev. Biophys. Bioeng.* **6** (1977), 151–176.

[Row63] J. S. Rowlinson. The triplet distribution function in a fluid of hard spheres. *Molec. Physics* **6** (1963), 517–524.

[Sch02] T. Schlick. *Molecular Modeling and Simulation.* Springer-Verlag, New York, 2002.

[Str93] G. Strang. *Introduction to Linear Algebra.* Wellesley-Cambridge Press, Wellesley, Massachusetts, 1993.

[Str88] L. Stryer. *Biochemistry.* Freeman, New York, 1988.

[SLL93] S. Subbiah, D. V. Laurents and M. Levitt. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Current Biol.* **3** (1993), 141–148.

[Tho49] R. Thom. Sur une partition en cellules associée à une fonction sur une variété. *Comptes Rendus l'Acad. Sci.* **228** (1949), 973–975.

[VBR+95] A. Varshney, F. P. Brooks, Jr., D. C. Richardson, W. V. Wright and D. Minocha. Defining, computing and visualizing molecular interfaces. *In* "Proc. IEEE Visualization, 1995", 36–43.

[Vor07] G. Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *J. Reine Angew. Math.* **133** (1907), 97–178, and **134** (1908), 198–287.

[WC53] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acid: a structure for deoxyribose nucleic acid. Genetic implications of the structure of deoxyribonucleic acid. *Nature* **171** (1953), 737–738 and 964–967.

[Wel96] J. A. Wells. Binding in the growth hormone receptor complex. *Proc. Natl. Acad. Sci.* **93** (1996), 1–6.