65 COMPUTATIONAL TOPOLOGY FOR STRUCTURAL MOLECULAR BIOLOGY

Herbert Edelsbrunner and Patrice Koehl

INTRODUCTION

The advent of high-throughput technologies and the concurrent advances in information sciences have led to a data revolution in biology. This revolution is most significant in molecular biology, with an increase in the number and scale of the "omics" projects over the last decade. *Genomics* projects, for example, have produced impressive advances in our knowledge of the information concealed into genomes, from the many genes that encode for the proteins that are responsible for most if not all cellular functions, to the noncoding regions that are now known to provide regulatory functions. *Proteomics* initiatives help to decipher the role of post-translation modifications on the protein structures and provide maps of protein-protein interactions, while *functional genomics* is the field that attempts to make use of the data produced by these projects to understand protein functions. The biggest challenge today is to assimilate the wealth of information provided by these initiatives into a conceptual framework that will help us decipher life. For example, the current views of the relationship between protein structure and function remain fragmented. We know of their sequences, more and more about their structures, we have information on their biological activities, but we have difficulties connecting this dotted line into an informed whole. We lack the experimental and computational tools for directly studying protein structure, function, and dynamics at the molecular and supra-molecular levels. In this chapter, we review some of the current developments in building the computational tools that are needed, focusing on the role that geometry and topology play in these efforts. One of our goals is to raise the general awareness about the importance of geometric methods in elucidating 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-man's-land between combinatorial and numerical algorithms, and it seems opportune to explore this land with a computational-geometric frame of mind.

65.1 BIOMOLECULES

GLOSSARY

DNA: Deoxyribo Nucleic Acid. A double-stranded molecule found in all cells that is the support of genetic information. Each strand is a long polymer built from four different building blocks, the nucleotides. The sequence in which these nucleotides are arranged contains the entire information required to describe cells

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

and their functions. The two strands are complementary to each other, allowing for repair should one strand be damaged.

RNA: Ribo Nucleic Acid. A long polymer much akin to DNA, being also formed as sequences of four types of nucleotides. RNAs can serve as either carrier of information (in their linear sequences), or as active functional molecules whose activities are related to their 3-dimensional shapes.

Protein: A long polymer, also called a *polypeptide chain*, built from twenty different building blocks, the amino acids. Proteins are active molecules that perform most activities required for cells to function.

Genome: Genetic material of a living organism. It consists of DNA, and in some cases, of RNA (RNA viruses). For humans, it is physically divided into 23 *chromosomes*, each forming a long double-strand of DNA.

Gene: A gene is a segment of the genome that encodes a functional RNA or a protein product. The transmission of genes from an organism to its offsprings is the basis of the heredity.

Central dogma: The Central Dogma is a framework for understanding the transfer of information between the genes in the genome and the proteins they encode for. Schematically, it states that "DNA makes RNA and RNA makes protein."

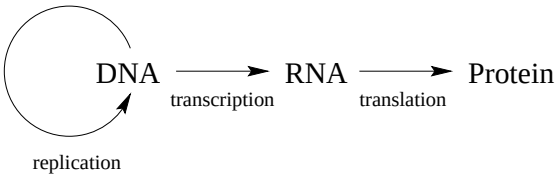


FIGURE 65.1.1
The DNA gets replicated as a whole. Pieces of DNA referred to as genes are transcribed into pieces of RNA, which are then translated into proteins.

Replication: Process of producing two identical replicas of DNA from one original DNA molecule.

Transcription: First step of gene expression, in which a particular segment of DNA (gene) is copied into an RNA molecule.

Translation: Process in which the messenger RNA produced by transcription from DNA is decoded by a ribosome to produce a specific amino acid chain, or protein.

Protein folding: Process in which a polypeptide chain of amino acid folds into a usually unique globular shape. This 3D shape encodes for the function of the protein.

Intrinsically disordered protein (IDP): A protein that lacks a fixed or ordered three-dimensional structure or shape. Despite their lack of stable structure, they form a very large and functionally important class of proteins.

INFORMATION TRANSFER: FROM DNA TO PROTEIN

One of the key features of biological life is its ability to self-replicate. Self-replication is the behavior of a system that yields manufacturing of an identical copy of itself. Biological cells, given suitable environments, reproduce by cell division. During cell division, the information defining the cell, namely its genome, is replicated and

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

then transmitted to the daughter cells: this is the essence of heredity. Interestingly, the entire machinery that performs the replication as well as the compendium that defines the process of replication are both encoded into the genome itself. Understanding the latter has been at the core of molecular biology. Research in this domain has led to a fundamental hypothesis in biology: the Central Dogma. We briefly describe it in the context of information transfer.

The genome is the genetic material of an organism. It consists of DNA, and in a few rare cases (mostly some viruses), of RNA. The DNA is a long polymer whose building blocks are nucleotides. Each nucleotide contains two parts, a backbone consisting of a deoxyribose and a phosphate, and an aromatic base, of which there are four types: adenine (A), thymine (T), guanine (G) and cytosine (C). The nucleotides are linked together to form a long chain, called a *strand*. Cells contain strands of DNA in pairs that are mirrors of each other. When correctly aligned, A pairs with T, G pairs with C, and the two strands form a double helix [WC53]. The geometry of this helix is surprisingly uniform, with only small, albeit important structural differences between regions of different sequences. The order in which the nucleotides appear in one DNA strand defines its sequence. Some stretches of the sequence contain information that can be transcribed first into an RNA molecule and then translated into a protein (Central Dogma). These stretches are called *genes*. It is estimated, for example, that the human genome contains around 20,000 genes [PS10], which represent 1-3% of the whole genome. For a long time, the remainder was considered to be nonfunctional, and therefore dubbed to be "junk" DNA. This view has changed, however, with the advent of the genomic projects. For example, the international Encyclopedia of DNA Elements (ENCODE) project has used biochemical approaches to uncover that at least 80% of human genomic DNA has biochemical activity [ENC12]. While this number has been recently questioned as being too high [Doo13, PG14], as biochemical activities may not imply function, it remains that a large fraction of the noncoding DNA plays a role in regulation of gene expression.

DNA replication is the biological process of generating two identical copies of DNA from one original DNA molecule. This process occurs in all living organisms; it is the basis for heredity. As DNA is made up of two complementary strands wound into a double helix, each strand serves as a template for the production of the complementary strand. This mechanism was first suggested by Watson and Crick based on their model of the structure for DNA [WC53]. As replication is the mechanism that ensures transfer of information from one generation to the other, most species have developed control systems to ensure its fidelity. Replication is performed by DNA polymerases. The function of these molecular machines is not quite perfect, making about one mistake for every ten million base pairs copied [MK08]. Error correction is a property of most of the DNA polymerases. When an incorrect base pair is recognized, DNA polymerase moves backwards by one base pair of DNA, excises the incorrect nucleotide and replaces it with the correct one. This process is known as *proofreading*. It is noteworthy that geometry plays an important role here. Incorporation of the wrong nucleotide leads to changes in the shape of the DNA, and it is this change in geometry that the polymerase detects. In addition to the proofreading process, most cells rely on post-replication mismatch repair mechanisms to monitor the DNA for errors and correct them. The combination of the intrinsic error rates of polymerases, proofreading, and post-replication mismatch repair usually enables replication fidelity of less than one mistake for every billion nucleotides added [MK08]. We do note that this level of fidelity may

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

vary between species. Unicellular organisms that rely on fast adaptation to survive, such as bacteria, usually have polymerases with much lower levels of fidelity [Kun04].

Transcription is the first step in the transfer of information from DNA to its end product, the protein. During this step, a particular segment of DNA is copied into RNA by the enzyme RNA polymerase. RNA molecules are very similar to DNA, being formed as sequences of four types of nucleotides, namely A, G, C, and uracil (U), which is a derivative of thymine. In contrast to the double-stranded DNA, RNA is mostly found to be singled-stranded. This way, it can adopt a large variety of conformations, which remain difficult to predict based on the RNA sequence [SM12]. Interestingly, RNA is considered an essential molecule in the early steps of the origin of life [Gil86, Cec93].

Translation is the last step in gene expression. In translation, the messenger RNA produced by transcription from DNA is decoded by a ribosome to produce a specific amino acid chain, or polypeptide. There are 20 types of amino acids, which share a common *backbone* and are distinguished by their chemically diverse *side-chains*, which range in size from a single hydrogen atom to large aromatic rings and can be charged or include only nonpolar saturated hydrocarbons. The order in which amino acids appear defines the *primary sequence*, also referred to as the *primary structure*, of the polypeptide. In its native environment, the polypeptide chain adopts a unique 3-dimensional shape, in which case it is referred to as a *protein*. The shape defines the *tertiary* or *native structure* of the protein. In this structure, nonpolar amino acids have a tendency to re-group and form the core, while polar amino acids remain accessible to the solvent.

We note that the scenario "DNA makes RNA and RNA makes protein" captured by the Central Dogma is reminiscent of the Turing 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

Proteins, the end products of the information encoded in the genome of any organism, play a central role in defining the life of this organism as they catalyze most biochemical reactions within cells and are responsible, among other functions, for the transport of nutrients and for signal transmission within and between cells. Proteins become functional only when they adopt a 3-dimensional shape, the so-called tertiary, or native structure of the protein. This is by no means different from the macroscopic world: most proteins serve as tools in the cell and as such either have a defined or adaptive shape to function, much like the shapes of the tools we use are defined according to the functions they need to perform. Understanding the shape (the geometry) of a protein is therefore at the core of understanding how cells function. From the seminal work of Anfinsen [Anf73], we know that the sequence fully determines the 3-dimensional structure of the protein, which itself defines its function. While the key to the decoding of the information contained in genes was found more than fifty years ago (the genetic code), we have not yet found the rules that relate a protein sequence to its structure [KL99, BS01]. Our knowledge of protein structure therefore comes from years of experimental studies, either using X-ray crystallography or NMR spectroscopy. The first protein structures to be solved were those of hemoglobin and myoglobin [KDS$^+$60, PRC$^+$60].

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

As of June 2016, there are more than 110,000 protein structures in the database of biomolecular structures [BWF+00]; see `http://www.rcsb.org`. This number remains small compared to the number of existing proteins. There is therefore a lot of effort put into predicting the structure of a protein from the knowledge of its sequence: one of the "holy grails" in molecular biology, namely the protein structure prediction problem [Dil07, EH07, DB07, Zha08, Zha09]. Efforts to solve this problem currently focus on protein sequence analysis, as a consequence of the wealth of sequence data resulting from various genome-sequencing projects, either completed or ongoing. As of May 2016, there were more than 550,000 protein sequences deposited in SwissProt-Uniprot version 2016-05, the fully annotated repository of protein sequences. Data produced by these projects have already led to significant improvements in predictions of both protein 3D structures and functions; see for example [MHS11]. However, we still stand at the dawn of understanding the information encoded in the sequence of a protein.

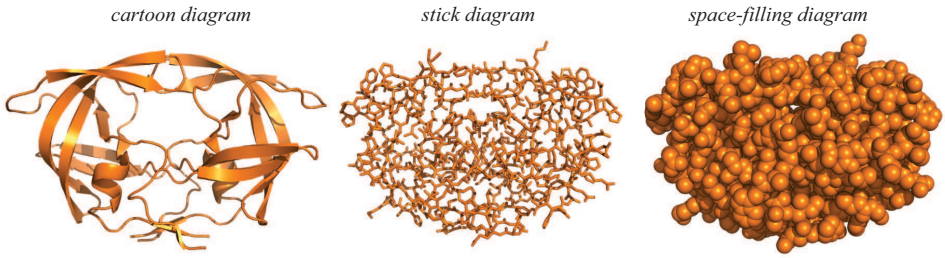
It is worth noting that if the paradigm shape-defines-function is the rule in biology, intrinsically disordered proteins form a significant class of exceptions, as they lack stable structures [DW05, DSUS08]. Shape, however, remains important for those proteins, although it is its flexibility and plasticity that is of essence, as shown for example in the case of P53 [OMY+09].

65.2 GEOMETRIC MODELS

The shape of a protein and its chemical reactivity are highly correlated as the latter depends on the positions of the nuclei and electrons within the protein: this correlation is the rationale for high-resolution experimental and computational studies of the structures and shapes of proteins. Early crystallographers who studied proteins could not rely (as it is common nowadays) on computers and computer graphics programs for representation and analysis of their structures. They had developed a large array of finely crafted physical models that allowed them to have a feeling for the shapes of these molecules. These models, usually made out of painted wood, plastic, rubber, and/or metal, were designed to highlight different properties of the protein under study. The current models in computer graphics programs mimic those early models. The *cartoon diagrams*, also called *ribbon diagrams* or *Richardson diagrams* [Ric85] show the overall path and organization of the protein backbone in 3D. Cartoon diagrams are generated by interpolating a smooth curve through the polypeptide backbone. In the *stick models*, atoms are represented as points (sometimes as small balls) attached together by sticks that represent the chemical bonds. These models capture the stereochemistry of the protein. In the *space-filling models*, such as those of Corey-Pauling-Koltun (CPK) [CP53, Kol65], atoms are represented as balls, whose sizes are set to capture the volumes occupied by the atoms. The radii of those balls are set to the van der Waals radii of the atoms. The *CPK model* has now become standard in the field of macromolecular modeling: a protein is represented as the union of a set of balls, whose centers match with the atomic centers and radii defined by van der Waals radii. The structure of a protein is then fully defined by the coordinates of these centers, and the radii values. The *macromolecular surface* is the geometric surface or boundary of these unions of balls. Note that other definitions are possible; this will be discussed in more detail below.

GLOSSARY

FIGURE 65.2.1
Three representations (diagrams) of the same protein, the HIV-1 protease (Protein Data Bank [BWF+00], identifier: 3MXE). The cartoon diagram on the left characterizes the geometry of the backbone of the protein, the stick diagram in the middle shows the chemical bonds, and the space-filling diagram highlights the space occupied by the protein. The three diagrams complement each other in their representation of relevant information.

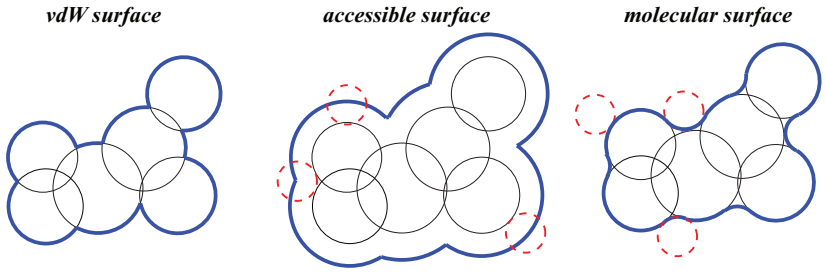


Cartoon diagram: Model that represents the overall path and organization of the protein backbone in 3D. The cartoon diagram is generated by interpolating a smooth curve through the protein backbone.

Stick diagram: Model that represents the chemical connectivity in a protein by displaying chemical bonds as sticks (edges). Atoms are usually just vertices where the edges meet. Some stick diagrams use balls to represent those vertices, the ball-and-stick models.

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.

FIGURE 65.2.2
Three most common molecular surface models for representing proteins (2D examples). Dashed, red circles represent the probe solvent spheres.



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,

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

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 $\|x - z\|^2 - r^2$.

Voronoi diagram: Decomposition of space into convex polyhedra. Each polyhedron corresponds 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.

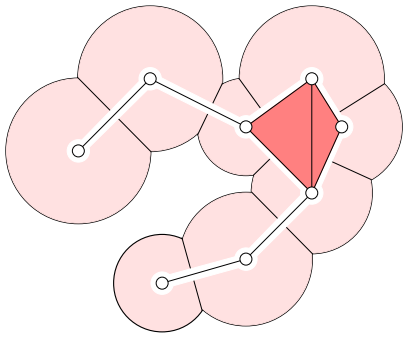


FIGURE 65.2.3
Each Voronoi polygon intersects the union of disks in a convex set, which is the intersection with its defining disk. The drawing shows the Voronoi decomposition of the union and the dual complex superimposed.

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 here is the sequence of alpha complexes.

ALTERNATIVE SURFACE REPRESENTATIONS

While geometric models for the molecular surface provide a deterministic description of the boundary for the shape of a biomolecule, models using implicit or parametric surfaces may be favorable for certain applications [Bli82, ZBX11]. The *implicit molecular surface models* use a level set of a scalar function $f\colon \mathbb{R}^3 \to \mathbb{R}$ that maps each point from the 3-dimensional space to a real value [OF03, CCW06, CP13]. The most common scalar function used for macromolecular surfaces is a

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

summation of Gaussian functions [GP95]. Other scalar functions, such as polynomial and Fermi-Dirac switching functions, have been used as well [LFSB03]. Bates *et al.* [BWZ08] proposed the *minimal molecular surface* as a level set of a scalar function that is the output from a numerical minimization procedure. *Parametric surface models* specify each point on the macromolecular surface by a pair of real value variables. Piecewise polynomials such as Non-Uniform Rational B-spline (NURBS) and Bernstein-Bézier have been proposed to generate parametric representations for molecular surfaces [BLMP97, ZBX11]. Spherical harmonics and their extensions parametrize the macromolecular surface using spherical coordinates and provide a compact analytical representation of macromolecular shapes [MG88, DO93a, DO93b].

We note that neither implicit nor parametric macromolecular surface models are independent from the geometric models based on the union of balls, as they usually have a set of parameters that are tuned such that they provide a reasonable approximation of the surface of the latter.

SPACE-FILLING DIAGRAMS

Our starting point is the *van der Waals force*. These forces capture interactions between atoms and molecules and mostly include attraction and repulsion. At short range up to a few Angstrom, this force is attractive but significantly weaker than covalent or ionic bonds. At very short range, the force is strongly repulsive. We can assign *van der Waals radii* to the atoms so 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*, after the creator of the first software representing this surface by a collection of dots [Con83]. We mention that this surface may have sharp edges, namely when the solvent sphere cannot quite squeeze through an opening of the protein and thus forms a circular or similar curve feature on the surface.

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,

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

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 ith 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.

ALPHA SHAPE THEORY

The dual structures and the growth model introduced above form the basis of the alpha shape theory and its applications to molecular shapes. Alpha shapes have a technical definition that was originally introduced to formalize the notion of 'shape' for a set of points [EKS83]. It can be seen as a generalization of the convex hull of this set. One *alpha complex* is a subcomplex of the Delaunay triangulation, and the corresponding *alpha shape* is the union of the simplices in the alpha complex. Such an alpha shape is characterized by a parameter, α, that corresponds to the parameter t defined above. This parameter controls the level of detail that is desired: the set of all alpha values leads to a family of shapes capturing the intuitive notion of 'crude' versus 'fine' shape of the set.

In its applications to structural biology, the set of points corresponds to the collection of atoms of the molecule of interest, with each atom assigned a weight corresponding to its van der Waals radius. The Delaunay triangulation of this set of weighted points is computed. Most applications require the alpha complex corresponding to $\alpha = 0$, as the corresponding alpha shape best represents the space-filling diagram (either delimited by the vdW surface or by the solvent accessible surface). The alpha complex, K_0, can then be used to measure the molecular shape. The complete filtration can also be used to characterize the topology of the biomolecule, as captured by the simplices of the dual complexes and of the Delaunay triangulation. This will be discussed below.

65.3 MOLECULAR SKIN OF A PROTEIN

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

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

the balls with van der Waals radii representing the atoms, we have balls interpolating between them that give rise to blending patches and, all together, 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. In mathematical terms, it is the boundary of the union of interpolated spheres, which we construct from the set of given spheres as follows. Supposing we have two spheres with centers a_1, a_2 and radii r_1, r_2 such that the distance between the two centers is smaller than $\sqrt{2}(r_1 + r_2)$. For each real number $0 \leq \lambda \leq 1$, the corresponding *interpolated sphere* is obtained by first increasing the radii to $\sqrt{2}r_1$ and $\sqrt{2}r_2$, second fixing the new center to $a_3 = (1 - \lambda)a_1 + \lambda a_2$, third choosing the new radius such that the sphere passes through the circle in which the given two spheres intersect, and fourth shrinking to radius $r_3/\sqrt{2}$. If the distance between the centers is larger than $\sqrt{2}(r_1 + r_2)$, then we extend the construction to include spheres with imaginary radii, which correspond to empty balls and therefore do not contribute to the surface we construct.

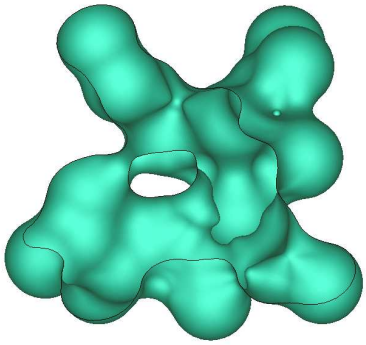


FIGURE 65.3.1
Cutaway view of the skin of a small molecule. We see a blend of sphere and hyperboloid patches. The surface is inside-outside symmetric: it can be defined by a collection of spheres on either of its two sides.

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-sample: A collection S of points on the molecular skin $\mathbb{M}$ such that every point $x \in \mathbb{M}$ has a point $u \in S$ at distance $\|x - u\| \leq \varepsilon/\kappa(x)$.

Restricted Delaunay triangulation: Dual to the restriction of the (3-dimensional) Voronoi diagram of S to the molecular skin $\mathbb{M}$.

Shape space: Locally parametrized space of shapes. The prime example here is

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

the $(k-1)$-dimensional space generated by k shapes, each specified by a collection of spheres in $\mathbb{R}^3$.

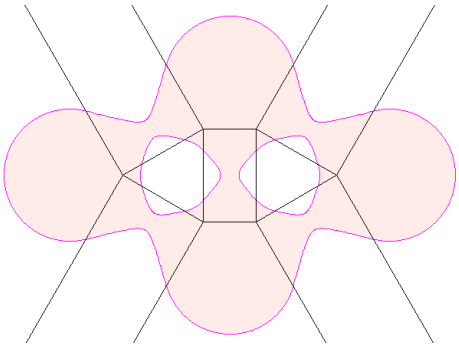


FIGURE 65.3.2
The skin curve defined by four circles in the plane. The mixed complex decomposes the curve into pieces of circles and hyperbolas.

TRIANGULATED MOLECULAR SKIN

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 $\kappa\colon \mathbb{M} \to \mathbb{R}$. 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 $\mathbb{M}$ that are used as vertices of the mesh. Given such a collection S of points, we construct a mesh using its Voronoi diagram restricted to $\mathbb{M}$. The polyhedra decompose the surface into patches, and the mesh is constructed as the dual of this 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 $\mathbb{M}$ is either empty or a topological disk, interval, or single point. Because of the smoothness of $\mathbb{M}$, this topological property is implied if the points form an ε-sampling, with $\varepsilon = 0.279$ or smaller [CDES01]. An alternative approach to triangulating the molecular skin can be found in [KV07].

DEFORMATION AND SHAPE SPACE

The variation of the maximum normal curvature function can be bounded by the one-sided Lipschitz condition $|1/\kappa(x) - 1/\kappa(y)| \leq \|x - y\|$, in which the distance is measured in $\mathbb{R}^3$. The continuity over $\mathbb{R}^3$ and not just over $\mathbb{M}$ 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 1-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 so that this is indeed the case is explained in [CEF01].

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

65.4 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 $f\colon \mathbb{X} \to \mathbb{Y}$ and $g\colon \mathbb{Y} \to \mathbb{X}$ 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.

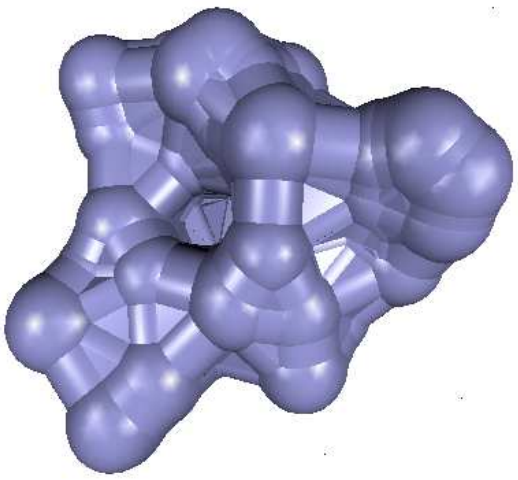


FIGURE 65.4.1
Snapshot during the deformation retraction of the space-filling representation of gramicidin to its dual complex. The spheres shrink to vertices while the intersection circles become cylinders that eventually turn into edges.

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

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

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

Voids: Bounded connected components of the complement. Here, 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 bichromatic Voronoi polygons that separate the proteins in the complex. The surface is retracted to the region in which the proteins are in close contact.

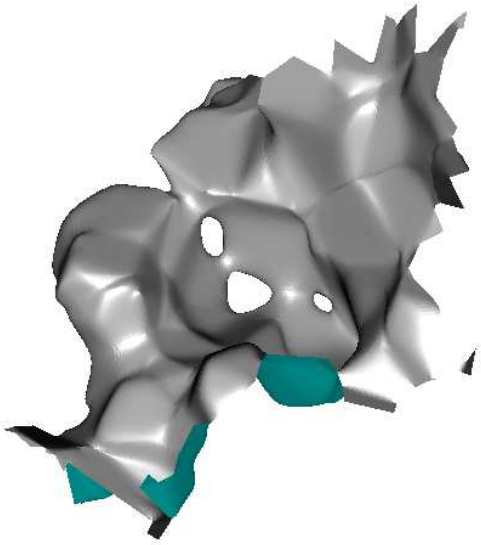


FIGURE 65.4.2
Molecular interface of the neurotoxic vipoxin complex. The surface has nonzero genus, which is unusual. In this case, we have genus equal to three, which implies the existence of three loops from each protein that are linked with each other. The linking might explain the unusually high 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 exactly how we reduce, the running time can be anywhere between cubic in the number of simplices and exponential or worse.

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

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 nonzero 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. When we add a k-dimensional simplex, σ, the kth 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 bichromatic Voronoi polygons of a protein complex. To identify the essential portions of this surface, we again observe how voids are formed and retain the bichromatic polygons inside pockets while removing all others [BER06]. A different geometric formalization of the same biochemical concept can be found in [VBR+95].

Preliminary experiments in the 1990s suggested 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]. In the 2000s, persistent homology was used to characterize structural changes in membrane fusion over the course of a simulation [KZP+07]. More recently, persistent homology has been used for extracting molecular topological fingerprints (MTFs) of proteins, based on the persistence of molecular topological invariants. These fingerprints have been used for protein characterization, identification, and classification [XW14, XW15a], as well as for cryo-EM data analysis [XW15b].

65.5 DENSITY MAPS IN STRUCTURAL BIOLOGY

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 3-dimensional structure of proteins [BJ76, Rho00]. The key to X-ray crystallography is to obtain first pure crystals of the protein of interest. The crystalline atoms cause a beam of incident X-rays to diffract into many specific directions. By measuring the angles and intensities of these diffracted beams, a

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

crystallographer can produce a 3-dimensional picture of the density of electrons within the crystal. From this electron density, the mean positions of the atoms can be determined, the chemical bonds that connect them, as well as their disorder. The first two protein structures to be solved using this technique were those of hemoglobin and myoglobin [KDS+60, PRC+60]. Of the 110,000 protein structures present in the database of biomolecular structures (PDB) as of June 2016, more than 99,000 were determined using X-ray crystallography.

A second setting is *molecular mechanics*, whose central focus is to develop insight on the forces that stabilize biomolecular structures. Describing the state of a biomolecule in terms of its energy landscape, the native state corresponds to a large basin, and it is mostly the structure of this basin that is of interest. Theoretically, the laws of quantum mechanics completely determine the energy landscape of any given molecule by solving Schrödinger's equation. In practice, however, only the simplest systems such as the hydrogen atom have an exact, explicit solution to this equation and modelers of large molecular systems must rely on approximations. Simulations are based on a space-filling representation of the molecule, in which the atoms interact through empirical forces. There is increased interest in the field of structure biology to map the results of those simulations onto the structures of the molecules under study, to help visualize their properties. We may, for example, be interested in the electrostatic potential induced by a protein and visualize it as a density map over 3-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, LLY+15].

GLOSSARY

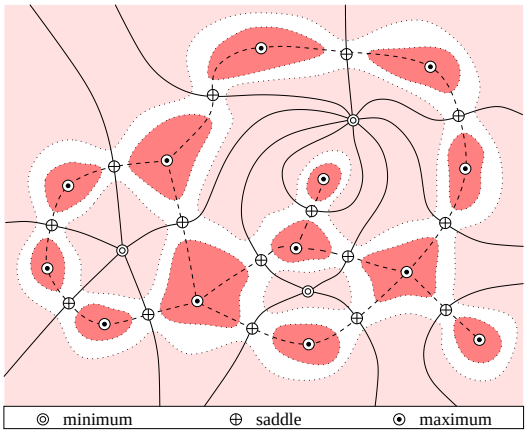


FIGURE 65.5.1
Portion of the Morse-Smale complex of a Morse-Smale function over a 2*-manifold. The solid stable* 1*-manifolds and the dashed unstable* 1*-manifolds are shown together with two dotted level sets. Observe that all* 2*-dimensional regions of the complex are quadrangular.*

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

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

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 ***nondegenerate*** if the Hessian is invertible. The ***index*** of a nondegenerate 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 f to be a *Morse-Smale function* satisfying the additional genericity assumption that these intersections are transversal.

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 on manifolds, $f : \mathbb{M} \to \mathbb{R}$. 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. This is however no reason to give up on applying the underlying ideas of Morse theory. 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 noncritical 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, it 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 [EHZ03]. In practice, it is essential to be able to simplify these decompositions, which can be done by canceling critical points in pairs in the order of increasing persistence [ELZ02].

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

65.6 MEASURING BIOMOLECULES

Protein dynamics is multi-scale: from the jiggling of atoms (pico-seconds), to the domain reorganizations in proteins (micro to milliseconds), to protein folding and diffusion (milli-second to seconds), and finally to binding and translocation (seconds to minutes). Connecting these different scales is a central problem in polymer physics that remains unsolved despite numerous theoretical and computational developments; see [Gue07, DP11]. Computer simulations play an essential role in all corresponding multi-scale methods, as they provide information at the different scales. Usually, computer simulations of protein dynamics start with a large system containing the protein and many water molecules to mimic physiological conditions. Given a model for the physical interactions between these molecules, their space-time trajectories are computed by numerically solving the equations of motion. These trajectories however are limited in scope. Current computing technologies limit the range of time-scales that can be simulated to the microsecond level, for systems that contain up to hundred thousands of atoms [VD11]. There are many directions that are currently explored to extend these limits, from hardware solutions including the development of specialized computers [SDD$^+$07], or by harnessing the power of graphics processor units [SPF$^+$07], to the development of simplified models that are computationally tractable and remain physically accurate. Among such models are those that treat the solvent implicitly, reducing the solute-solvent interactions to their mean-field characteristics. These so-called *implicit solvent models* are often applied to estimate free energy of solute-solvent interactions in structural and chemical processes, folding or conformational transitions of proteins and nucleic acids, association of biological macromolecules with ligands, or transport of drugs across biological membranes. The main advantage of these models is that they express solute-solvent interactions as a function of the solute degrees of freedom alone, more specifically of its volume and surface area. Methods that compute the surface area and volume of a molecule, as well as their derivatives with respect to the position of its atoms are therefore of great interest for computational structural biology.

GLOSSARY

Indicator function: It maps a point x to 1 if $x \in P$ and to 0 if $x \notin P$, in which P is some fixed set. Here, we are interested in convex polyhedra P and can therefore use the alternating sum of the number of faces of various dimensions visible from x as an indicator; see [Ede95] for details.

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 a sphere minus a point to Euclidean space. The map preserves spheres and angles. We are primarily interested in the case in which both the sphere and the Euclidean space are 3-dimensional.

Atomic solvation parameters: Experimentally determined numbers that assess the hydrophobicity of different atoms [EM86].

Weighted area: Area of the boundary of a space-filling diagram in which the contribution of each individual ball is weighted by its atomic solvation parameter.

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

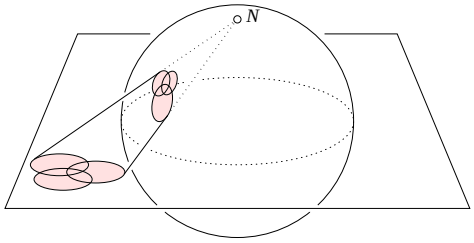


FIGURE 65.6.1
Stereographic projection from the north pole. The preimage of a circle in the plane is a circle on the sphere, which is the intersection of the sphere with a plane. By extension, the preimage of a union of disks is the intersection of the sphere with the complement of a convex polyhedron.

Also a function $A\colon \mathbb{R}^{3n} \to \mathbb{R}$ obtained by parametrizing a space-filling diagram by the coordinates of its n ball centers.

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

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 $V\colon \mathbb{R}^{3n} \to \mathbb{R}$ obtained by parametrizing a space-filling diagram by the coordinates of its n ball centers.

Weighted-volume derivative: The linear map $\mathrm{D}V_{\mathbf{z}}\colon \mathbb{R}^{3n} \to \mathbb{R}$ defined by $\mathrm{D}V_{\mathbf{z}}(\mathbf{t}) = \langle \mathbf{v}, \mathbf{t} \rangle$, in which $\mathbf{z} \in \mathbb{R}^{3n}$ 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 $\mathrm{D}V\colon \mathbb{R}^{3n} \to \mathbb{R}^{3n}$ 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 as an alternating sum of volumes of intersections:

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

in which Λ ranges over all nonempty 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 redundant, however, and a much smaller formula based on the dual complex K of the space-filling diagram F has been given [Ede95]:

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

in which $\bigcap \sigma$ 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

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

is not difficult to modify the formula to get the weighted volume: decompose the terms $\mathrm{vol} \bigcap \sigma$ into the portions within the Voronoi cells of the participating balls and weight each portion accordingly.

DERIVATIVES

The relationship between the weighted- and unweighted-volume derivatives is less direct than that between the weighted and unweighted volumes. Just to state the formula for the weighted-volume derivative requires more notation than we are willing to introduce here. Instead, we describe the two geometric ingredients, both of which can be computed by geometric inclusion-exclusion [EK03]. 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 rotate the vector to sweep out the facet. A similar approach allows for the computation of the weighted area derivative [BEKL04].

VOIDS AND POCKETS

A *void* V is a maximal connected subset of space that is disjoint from and completely surrounded by the union of balls. Its surface area is easily computed by identifying the sphere patches on the boundary of the union that also bound the void. It helps to know that there is a deformation retraction from the union of balls $\bigcup_i B_i$ to the dual complex K [Ede95]. Similarly, there is a corresponding void in K represented by a connected set of simplices in the Delaunay triangulation, that do not belong to K. This set U is open and its boundary (the simplices added by closure) forms what one may call the dual complex of the boundary of V. The volume and surface area of the void V are then computed based on U [Ede95].

It would be interesting to generalize these ideas to pockets as defined in [EFL98]. In contrast to a void, a *pocket* is not completely surrounded by the balls but connected to the outside through narrow channels. Again we have a corresponding set of simplices in the Delaunay triangulation that do not belong to the dual complex, but this set is partially closed at the places the pocket connects to the outside. The inclusion-exclusion formulas still apply, but there are cases in which the cancellation of terms near the connecting channel is not complete and leads to slightly incorrect measurements.

65.7 SHAPE IN STRUCTURAL BIOLOGY

As bio-molecules are usually represented as unions of balls, it is not surprising to see geometric algorithms being adapted to characterize the shapes of molecules. We have discussed above the *alpha shape theory* [EM94], whose first applications in biology focused on computing the volume and surface area of molecular shapes [LEF+98] as well as on characterizing the cavities and pockets formed by a molecule

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

[LEW98, EFL98]. While these applications of the alpha shape theory remain popular in structural biology — with new and improved software implementations being released regularly, such as CASTp [DOT+06], Volume [CKL11], and UnionBall [MK11] — many applications in new domains have been proposed. Here we review a few of these applications.

GLOSSARY

Atom packing: A measure of how tight atoms are packed within a protein or nucleic acid structure.

Binding site: Region of a protein in which a ligand can bind. These regions often correspond to the cavities and pockets of the protein, though there are examples of binding sites that sit at the surface of the protein.

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.

STATISTICS OF PROTEIN STRUCTURE GEOMETRY

The experimental determination of a protein structure at the atomic level remains a difficult problem. There is hope however that theoretical and computational techniques will supplement experimental methods and enable protein structure prediction at the near atomic level [KL99]. Many of these techniques rely on the knowledge derived from the analysis of the geometry of known protein structures. Such an analysis requires an objective definition of atomic packing within a molecular structure. The alpha shape theory has proved a useful approach for deriving such a definition. For example, Singh *et al.* [STV96] used the Delaunay complex to define nearest neighbors in protein structures and to derive a four-body statistical potential. This potential has been used successfully for fold recognition, decoy structure determination, mutant analysis, and other studies; see [Vai12] for a full review. The potentials considered in these studies rely on the tetrahedra defined by the Delaunay triangulation of the points representing the atoms. In parallel, Zomorodian and colleagues have shown that it is possible to use the alpha shape theory to filter the list of pairwise interactions to generate a much smaller subset of pairs that retains most of the structural information contained in a proteins [ZGK06]. The alpha shape theory has also been used to compute descriptors for the shapes [WBKC09] and surfaces [TDCL09, TL12] of proteins.

SIMILARITY AND COMPLEMENTARITY

The alpha shape theory allows for the detection of simplices characterizing the geometry of a protein structure. Those simplices include points, edges, triangles, and tetrahedra connecting atoms of this protein structure. It is worth mentioning that it is possible to use those simplices to compare two protein structures and even to derive a structural alignment between them [RSKC05].

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

As the function of a protein is related to its geometry and as function usually involves binding to a partner protein, significant efforts have been put into characterizing the geometry of protein-ligand interactions, where *ligands* include small molecules, nucleic acids, and other proteins. Among these efforts, a few relate to the applications of the alpha shape theory. They have recently been extended to characterize binding sites at the surface of proteins [TDCL09, TL12]. The alpha shape theory has also been used to characterize the interfaces in protein-protein complexes [BER06] as well as protein-DNA interactions [ZY10].

We mention a geometric parallel between finding a structural alignment between two proteins and predicting the structure of their interactions. While the former is based on the identification of similar geometric patterns between the two structures, the latter is based on the identification of complementary patterns between the surfaces of the two structures. As mentioned above, geometric patterns based on the Delaunay triangulation have been used for structural alignment. In parallel, similar patterns have recently been used to predict protein-protein interactions [EZ12].

CHARACTERIZING MOLECULAR DYNAMICS

All the applications described above relate to the static geometry of molecules. Bio-molecules however are dynamics. A *molecular dynamics simulation* is designed to capture this dynamics: it follows the Newtonian dynamics of the molecule as a function of time, generating millions of snapshots over the course of its trajectory. The alpha shape theory has proved useful to characterize the geometric changes that occur during such a trajectory. For example, using the concept of topological persistence [ELZ02], Kasson et al. characterized structural changes in membrane fusion over the course of a simulation [KZP+07].

65.8 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; [Fej72], 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: [Ale61], 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 24: Persistent homology
Chapter 27: Voronoi diagrams and Delaunay triangulations
Chapter 29: Triangulations and mesh generation

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.

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

[Ale61] P. Alexandroff. *Elementary Concepts of Topology.* Dover, New York, 1961.

[Anf73] C.B. Anfinsen. Principles that govern protein folding. *Science*, 181:223–230, 1973.

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

[BEKL04] R. Bryant, H. Edelsbrunner, P. Koehl, and M. Levitt. The weighted area derivative of a space filling diagram. *Discrete Comput. Geom.*, 32:293–308, 2004.

[BER06] Y.E.A. Ban, H. Edelsbrunner, and J. Rudolph. Interface surfaces for protein-protein complexes. *J. ACM*, 53:361–378, 2006.

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

[Bli82] J.F. Blinn. A generalization of algebraic surface drawing. *ACM Trans. Graph.*, 1:235–256, 1982.

[BLMP97] C. Bajaj, H.Y. Lee, R. Merkert, and V. Pascucci. NURBS based B-rep models for macromolecules and their properties. In *Proc. 4th ACM Sympos. Solid Modeling Appl., 1997*, pages 217–228.

[BS01] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93–96, 2001.

[BWF+00] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, *et al.* The Protein Data Bank. *Nucl. Acids. Res.*, 28:235–242, 2000.

[BWZ08] P.W. Bates, G.-W. Wei, and S. Zhao. Minimal molecular surfaces and their applications. *J. Comp. Chem.*, 29:380–391, 2008.

[CCW06] T. Can, C.-I. Chen, and Y.-F. Wang. Efficient molecular surface generation using level set methods. *J. Molec. Graph. Modeling*, 25:442–454, 2006.

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

[Cec93] T.R. Cech. The efficiency and versatility of catalytic RNA: implications for an RNA world. *Gene*, 135:33–36, 1993.

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

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

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

[CKL11] F. Cazals, H. Kanhere, and S. Loriot. Computing the volume of union of balls: a certified algorithm. *ACM Trans. Math. Software*, 38:3, 2011.

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

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

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

[CP13] S.W. Chen and J.-L. Pellequer. Adepth: new representation and its implications for atomic depths of macromolecules. *Nucl. Acids Res.*, 41:W412–W416, 2013.

[CP53] R.B. Corey and L. Pauling. Molecular models of amino acids, peptides and proteins. *Rev. Sci. Instr.*, 24:621–627, 1953.

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

[DB07] R. Das and D. Baker. Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.*, 77:363–382, 2008.

[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*, 12:771–784, 1995.

[Del34] B. Delaunay. Sur la sphère vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk*, 7:793–800, 1934.

[Dil07] K.A. Dill, S.B. Ozkan, T.R. Weikl, J.D. Chodera, and V.A. Voelz. The protein folding problem: when will it be solved? *Curr. Opin. Struct. Biol.*, 17:342–346, 2007.

[DO93a] B.S. Duncan and A.J. Olson. Approximation and characterization of molecular surfaces. *Biopolymers*, 33:219–229, 1993.

[DO93b] B.S. Duncan and A.J. Olson. Shape analysis of molecular surfaces. *Biopolymers*, 33:231–238, 1993.

[Doo13] W.F. Doolittle. Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. (USA)*, 110:5294–5300, 2013.

[DOT+06] J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, and J. Liang. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nuc. Acids Res.*, 34:W116–W118, 2006.

[DP11] J.J. de Pablo. Coarse-grained simulations of macromolecules: from DNA to nanocomposites. *Ann. Rev. Phys. Chem.*, 62:555–574, 2011.

[DSUS08] A.K. Dunker, I. Silman, V.N. Uversky, and J.L. Sussman. Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, 18:756–764, 2008.

[DW05] H.J. Dyson and P.E. Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, 6:197–208, 2005.

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

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

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

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

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

[EH07] A. Elofsson and G. von Heijine. Membrane protein structure: prediction versus reality. *Annu. Rev. Biochem.*, 76:125–140, 2007.

[EHZ03] H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchy of Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete Comput. Geom.*, 30:87–107, 2003.

[EK03] H. Edelsbrunner and P. Koehl. The weighted volume derivative of a space-filling diagram. *Proc. Natl. Acad. Sci. (USA)*, 100:2203–2208, 2003.

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

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

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

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

[ENC12] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

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

[EZ12] L. Ellingson and J. Zhang. Protein surface matching by combining local and global geometric information. *PLoS One*, 7:e40540, 2012.

[Fej72] L. Fejes Tóth. *Lagerungen in der Ebene auf der Kugel und im Raum.* Second edition, Springer-Verlag, Berlin, 1972.

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

[Gil86] W. Gilbert. The RNA world. *Nature*, 319:618, 1986.

[GP95] J.A. Grant and B.T. Pickup. A Gaussian description of molecular shape. *J. Phys. Chem.*, 99:3503–3510, 1995.

[GR01] M. Gerstein and F.M. Richards. Protein geometry: distances, areas, and volumes. In M.G. Rossman and E. Arnold, editors, *The International Tables for Crystallography*, Vol. F, Chapter 22, pages 531–539. Kluwer, Dordrecht, 2001.

[Gue07] M.G. Guenza. Theoretical models for bridging timescales in polymer physics. *J. Phys. Condens. Matter*, 20:033101, 2007.

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

[KDS+60] J.C. Kendrew, R.E. Dickerson, B.E. Strandberg, R.G. Hart, D.R. Davies and D.C. Philips. Structure of myoglobin: a three dimensional Fourier synthesis at 2 angstrom resolution. *Nature*, 185:422–427, 1960.

[KL99] P. Koehl and M. Levitt. A brighter future for protein structure prediction. *Nature Struct. Biol.*, 6:108–111, 1999.

[Kol65] W.L. Koltun. Precision space-filling atomic models. *Biopolymers*, 3:665–679, 1965.

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

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

[Kun04] T.A. Kunkel. DNA replication fidelity. *J. Biol. Chem.*, 279:16895–16898, 2004.

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

[KV07] N. Kruithof and G. Vegter. Meshing skin surfaces with certified topology. *Comput. Geom.*, 36:166–182, 2007.

[KZP+07] P.M. Kasson, A. Zomorodian, S. Park, N. Singhal, L. Guibas and V.S. Pande. Persistent voids: a new structural metric for membrane fusion. *Bioinformatics*, 23:1753–1759, 2007.

[LEF+98] J. Liang, H. Edelsbrunner, P. Fu, P.V. Sudhakar, and S. Subramaniam. Analytical shape computation of macromolecules. I. Molecular area and volume through alpha shape. *Proteins: Struct. Func. Genet.*, 33:1–17, 1998.

[LEW98] J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Prot. Sci.*, 7:1884–1897, 1998.

[LFSB03] M.S. Lee, M. Feig, F.R. Salsbury and C.L. Brooks. New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comp. Chem.*, 24:1348–1356, 2003.

[LLY+15] H. Liu, F. Lin, J.-L. Yang, H.R. Wang, and X.-L. Liu. Applying side-chain flexibility in motifs for protein docking. *Genomics Insights*, 15:1–10, 2015.

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

[Mat02] Y. Matsumoto. *An Introduction to Morse Theory.* Amer. Math. Soc., Providence, 2002.

[MG88] N.L. Max and E.D. Getzoff. Spherical harmonic molecular surfaces. *IEEE Comput. Graph. Appl.*, 8:42–50, 1988.

[MHS11] D.S. Marks, T.A. Hopf, and C. Sander. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30:1072–1080, 2011.

[Mil63] J. Milnor. *Morse Theory.* Princeton University Press, 1963.

[MK08] S.D. McCulloch and T.A. Kunkel. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Research*, 18:148–161, 2008.

[MK11] P. Mach and P. Koehl. Geometric measures of large biomolecules: surface, volume, and pockets. *J. Comp. Chem.*, 32:3023–3038, 2011.

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

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

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

[OF03] S. Osher and R. Fedkiw. *Level Sets Methods and Dynamic Implicit Surfaces.* Spinger-Verlag, New York, 2003.

[OMY+09] C.J. Oldfield, J. Meng, J.Y. Yang, M.Q. Yang, V.N. Uversky, *et al.* Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics*, 9:S1, 2009.

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

[PG14] A.F. Palazzo and T.R. Gregory. The case for junk DNA. *PLoS Genetics*, 10:e1004351, 2014.

[PRC+60] M. Perutz, M. Rossmann, A. Cullis, G. Muirhead, G. Will, and A. North. Structure of hemoglobin: a three-dimensional Fourier synthesis at 5.5 angstrom resolution, obtained by X-ray analysis. *Nature*, 185:416–422, 1960.

[PS10] M. Pertea and S.L. Salzberg. Between a chicken and a grape: estimating the number of human genes. *Genome Biology*, 11:206–212, 2010.

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

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

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

[Ric85] J.S. Richardson. Schematic drawings of protein structures. *Methods in Enzymology*, 115:359–380, 1985.

[Row63] J.S. Rowlinson. The triplet distribution function in a fluid of hard spheres. *Molecular Phys.*, 6:517–524, 1963.

[RSKC05] J. Roach, S. Sharma, K. Kapustina, and C.W. Carter. Structure alignment via Delaunay tetrahedralization. *Proteins: Struct. Func. Bioinfo.*, 60:66–81, 2006.

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

[SDD+07] D.E. Shaw, M.M. Deneroff, R. Dror, *et al.* Anton, a special-purpose machine for molecular dynamics simulation. *ACM SIGARCH Computer Architecture News*, 35:1–12, 2007.

[SM12] M.G. Seetin and M.H. Mathews. RNA structure prediction: an overview of methods. *Methods Mol. Biol.*, 905:99–122, 2012.

[SPF+07] J.E. Stone, J.C. Philipps, P.L. Freddolino, D.J. Hardy, L.G. Trabuco and K. Schulten. Accelerating molecular modeling applications with graphics processors. *J. Comp. Chem.*, 28:2618–2640, 2007.

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

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

[STV96] R.K. Singh, A. Tropsha, and I.I. Vaisman. Delaunay tessellation of proteins: Four body nearest-neighbor propensities of amino acid residues. *J. Comput. Biol.*, 3:213–221, 1996.

[TDCL09] Y.Y. Tseng, C. Dupree, Z.J. Chen and W.H. Li. SplitPocket: identification of protein functional surfaces and characterization of their spatial patterns. *Nucl. Acids Res.*, 37:W384–W389. 2009.

[Tho49] R. Thom. Sur une partition en cellules associée à une fonction sur une variété. *C. R. Acad. Sci. Paris*, 228:973–975, 1949.

[TL12] Y.Y. Tseng and W.H. Li Classification of protein functional surfaces using structural characteristics. *Proc. Natl. Acad. Sci. (USA)* 109:1170–1175, 2012.

[Vai12] I.I. Vaisman. Statistical and computational geometry of biomolecular structure. In J.E. Gentle, W.K. Hardle, and Y. Mori, editors, *Handbook of Computational Statistics*, pages 1095–1112, Springer-Verlag, New York, 2012.

[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*, pages 36–43.

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.

[VD11] M. Vendruscolo and C.M. Dobson. Protein dynamics: Moore's law in molecular biology. *Current Biology*, 21:R68–R70, 2011.

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

[WBKC09] J.A. Wilson, A. Bender, T. Kaya and P.A. Clemons. Alpha shapes applied to molecular shape characterization exhibit novel properties compared to established shape descriptors. *J. Chem. Inf. Model.*, 49:2231–2241, 2009.

[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*, 171:737–738 and 964–967, 1953.

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

[XW14] K. Xia and G.-W. Wei. Persistent homology analysis of protein structure, flexibility, and folding. *Int. J. Numer. Meth. Biomed. Engng.*, 30:814–844, 2014.

[XW15a] K. Xia and G.-W. Wei. Multidimensional persistence in biomolecular data. *J. Comput. Chem.*, 20:1502–1520, 2015.

[XW15b] K. Xia and G.-W. Wei. Persistent homology for cryo-EM data analysis. *Int. J. Numer. Meth. Biomed. Engng.*, 31, 2015.

[ZBX11] W. Zhao, C. Bajaj and G. Xu. An Algebraic spline model of molecular surfaces for energetic computations. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 8:1458–1467, 2011.

[ZGK06] A. Zomorodian, L. Guibas, and P. Koehl. Geometric filtering of pairwise atomic interactions applied to the design of efficient statistical potentials. *Comput. Aided Graph. Design*, 23:531–544, 2006.

[Zha08] Y. Zhang. Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.*, 18:342–348, 2008.

[Zha09] Y. Zhang. Protein structure prediction: when is it useful? *Curr. Opin. Struct. Biol.*, 19:145–155, 2009.

[ZY10] W. Zhou and H. Yan. A discriminatory function for prediction of protein-DNA interactions based on the alpha shape modeling. *Bioinformatics*, 26:2541–2548, 2010.

Preliminary version (August 10, 2017). To appear in the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.