ITR/ACS+IM
Computational Geometry for Structural Biology and Bioinformatics

Response to Program Announcement NSF 99-167

DUKE UNIVERSITY: Pankaj K. Agarwal*, Herbert Edelsbrunner* (PI), Homme W. Hellinga†,
STANFORD UNIVERSITY: Leonidas J. Guibas*, Jean-Claude Latombe*, Michael Levitt‡,
UNIVERSITY OF NORTH CAROLINA: Frederick P. Brooks, Jr.*, Charles W. Carter§, Jack S. Snoeyink*,
NORTH CAROLINA AGRICULTURE AND TECHNICAL STATE UNIVERSITY: Solomon Bililign¶

PROJECT SUMMARY

The function of all life forms depends on organization in space and time, and the effect of one part of a biological system on another is generally much greater when the two parts are in close proximity in space and/or time. In themselves, these two observations would seem to indicate that geometric methods should be an essential component of any attempt to understand and simulate biological systems. Existing techniques in computational structural biology and bioinformatics, however, rely primarily on sequence information and use statistical and energy-based methods to analyze biological structure and function. They have been developed over three decades and have their roots in methods first applied by computational chemists to much smaller molecular systems. Although there have been significant advancements in the field, a systematic solution of many of the most important biological problems is still elusive, including *ab initio* protein structure prediction, the protein folding process, and ligand to protein docking. It is widely believed that the geometry of molecules plays a crucial role in these processes, yet geometrical methods are relatively uncommon in computational biology because several unresolved representational and algorithmic issues remain.

We propose to develop new computational techniques and paradigms for representing, storing, searching, simulating, analyzing, and visualizing biological structures. We will rely on geometry, but combine it with statistics and physics. We will aim for methods that have practical, predictive power and validate them by comparison with the best existing techniques. In order to transfer the technology in an effective way and have real impact on research in biology, the proposed project will create software whose aim is to help structural biologists with their research and integrate with their current tools.

Ideas from a wide range of areas of computer science and mathematics, including algorithms, geometry, topology, graphics, robotics, and databases will be needed to accomplish our goals. Some of the problem areas to be addressed represent great challenges for computer science itself. These include building and querying large libraries of three-dimensional and possibly flexible shapes, exploring hierarchical representations of deformable geometry, integrating geometry and physics in modeling, and properly sampling systems with many degrees of freedom. Although we focus on computational structural biology and bioinformatics applications, the research carried out under the proposed project will have a wider impact and will advance knowledge in several areas of computer science, including geometric modeling, shape analysis, 3D geometry databases, physical simulation, robotics, and visualization. The project will also foster integration of research and education in the proposed domains. We have put together a team of researchers with strong credentials who have already made significant contributions to these and related areas. To ensure that we address the real problems in computational structural biology and bioinformatics and that we have a fruitful collaboration, biologists, chemists and physicists will participate and be involved in the research from the beginning.

This proposal has aspects of both the Advanced Computational Science and the Information Management parts of the ITR solicitation, as it addresses the need for new geometric representations and algorithms in structural biology.

*Computer Science
†Biochemistry
‡Structural Biology
§Biochemistry and Biophysics
¶Physics

PROJECT DESCRIPTION

In this proposal we have assembled an interdisciplinary team to address fundamental computational problems in the representation of molecular structures and the simulation of biochemical processes important to life. Among these are ligand-to-protein docking, *ab initio* structure prediction, and protein folding. We also plan to consider engineering tasks, such as drug and protein design. Through a novel focus on geometric and topological representations, we have an opportunity to enable new scientific breakthroughs and more generally to transform the way we represent, analyze, communicate, and teach these fundamental structures and processes in molecular biology. In the process of doing so we will need to make advances in several areas of information technology that will be scientific accomplishments in themselves as well as being potentially relevant to other natural sciences and engineering disciplines dealing with computer modeling of the physical world.

1 Research Goals

We derive the structure of this proposal from the research flow diagram

$$\text{Biology} \rightarrow \text{Modeling} \rightarrow \text{Algorithms} \rightarrow \text{Software}$$

with Software feeding back into Biology. As mentioned in the PROJECT SUMMARY, our approach to biological problems is geometric and targets the molecular level of life. In

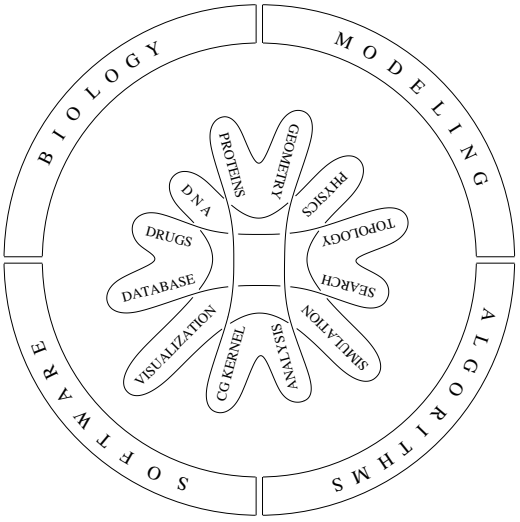

Figure 1: The circle starts with problems in biology motivating the proposed geometry work.

the first four sections, we outline the research opportunities and our geometry-based approaches in each of these four research areas. In Section 1.5, we point out some of the specific challenges for Information Technology posed by these investigations, and addressed by our project.

1.1 Biology

A deep but possibly surprising insight that emerged from decades of research in biology is that function follows form on the molecular level of life. To give a glimpse of how this connection between biology and geometry unfolds, we identify three classes of research activity in structural biology and derive tangible computational challenges in geometry from each.

Observation includes the experimental determination of structures, their analysis, and their classification. By many peoples' definition, bioinformatics is the science of determining, classifying, and organizing structures for the purpose of storage and retrieval; this is discussed in the paragraph on SHAPE ORGANIZATION in Section 1.3. The determination and analysis of structure requires geometric and numerical methods whose effectiveness depends on appropriate representations, as addressed from different angles throughout this proposal.

Simulation establishes the link between structure and function. Central to simulation is the analysis of motion. As discussed in the paragraph on MOLECULAR DYNAMICS in Section 1.4, motion is generated following the laws of statistical mechanics. There is obviously a strong connection to MOTION PLANNING as practiced in computer science and discussed in Section 1.3. During motion it is important to have efficient algorithms that track the geometry and topology as it varies with the changing conformation, as described in the paragraph on DEFORMATION.

Design seeks linear amino-acid sequences that fold into predetermined three-dimensional structures, or small molecules that bind to predetermined sites in proteins. Protein design is a top-down process using geometric modeling and analysis at every level. To identify primitives that are helpful in design is by and large an open problem, and we expect that the proposed work will lead to important contributions. There are also connections to more specific and well-defined computational geometry questions, such as measuring the quality of fit between two or more molecules, as discussed in the paragraphs on DRUG SHAPE CLASSIFICATION and on SHAPE DESCRIPTORS.

The remainder of this section discusses three areas in biology that will serve as motivation and focus for the geometric work proposed in this project.

Protein Design. The general goal of PROTEIN DESIGN is the *ab initio* creation of a sequence that will form a predetermined structure. Protein design creates a natural link between theory and experimentation by providing the testing ground for general hypotheses on protein structure, function

and evolution. It also has significant biotechnological applications. If it is possible to design functional proteins at will, a large number of different avenues for the development of important technologies will become available, including the development of catalysts for asymmetric chemical synthesis (enzymes), biosensors (receptors), macromolecular therapeutics (inhibitors, activators), and biomaterials.

The protein design problem can be attacked in two stages: inverse folding and full design. The *inverse folding* problem is to predict amino-acid sequences compatible with a given three-dimensional backbone structure. This involves both the choice and the placement of amino-acid residues. The resulting combinatorial problem is extremely large, since 20 amino-acids and their corresponding thousand side-chain conformations can be placed at each position along the backbone. Nevertheless, a number of algorithms have been developed over the last decade, several of which yield predictions that give reasonably folded small proteins when tested experimentally [18, 21]. Most of the inverse folding calculations and experiments conducted so far have focused on backbones that were derived from natural proteins. The *full design problem* requires that the backbone is also designed *de novo*. In order to design biologically active proteins, it will be necessary to couple the design of ligand-binding sites into the protein. This requires the design of regions of the protein surface that are complementary to a ligand of choice [43] — a difficult problem, but one in which recent progress has been made for metal-binding centers.

Not surprisingly, the full design problem is still in its infancy, but two important areas in which geometric algorithms and software play critical roles are clearly delineated:

1. Top-down design of a hierarchy of shape representations used to generate structure and sequences.
2. Independent validation tools.

Top-down design envisages a small number of high-level geometric structural or shape elements, which in appropriate combination and deformation can be used to represent any known protein structure. Methods for generating combinatorial assemblies of these shape elements, together with a set of constraints and cost functions, need to be established. The constraints will be hierarchical in nature, ranging from topological rules to geometric relationships. Finally, the hierarchy of shape representations needs to be traversed to introduce increasing levels of detail, until ultimately an atomic representation of a protein backbone is generated. This backbone is then used to calculate a sequence by solving the associated inverse folding problem.

Independent validation tools are necessary to analyze the designed structure, and thus determine compliance with quantifiable factors that are currently understood to contribute to folding and stability. These include physicochemical parameters, represented by heuristically determined force fields, as well as relationships like amino-acid secondary structure preferences or pairwise potentials. Figure 2 illustrates what kind of tool we imagine. Of particular

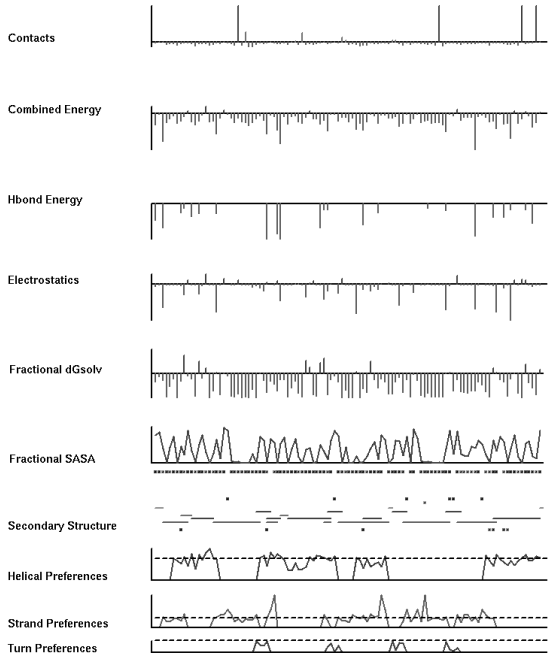

Figure 2: Graphical display of validation parameters as applied to an inverse folding solution generated for a natural protein fold. The display shows a superimposition of structural classification (secondary structure, and topological information) together with statistically derived local structural preferences, as well as semi-empirical force-field parameters such as hydrogen-bonding, van der Waals (steric), and electrostatic energies.

interest is the development of tools that measure compliance with geometrical features such as the number and distribution of internal cavities, and surface roughness. This work requires the development of a variety of tools, partially based on computational geometry work described in more detail in the paragraphs on SHAPE REPRESENTATION in Section 1.2, and on DEFORMATION and SHAPE ORGANIZATION in Section 1.3.

Drug Shape Classification. Drugs and other ligands bind to proteins because the free energy of the bound state is lower than of the unbound state. In many cases, this favorable binding involves a good fit of the ligand into the protein active site, leading to an emphasis on shape complementarity as a primary determinant for drug binding. Shape is defined broadly to include the van der Waal's surface, the electrostatic potential surface, the variable/conserved surface, etc. This is illustrated in Figure 3, which shows the biotin vitamin that strongly binds to streptavidin, a protein excreted by Streptomyces avidinii. Its overall geometry is shown here as a collection of balls. The fitness or quality of its interaction with streptavidin can be quantified by taking in account the excluded volume of the balls, their properties (such as elec-

trostatics), as well as their surface areas (which provide a measure of hydrophobic interactions). We can use appropri-

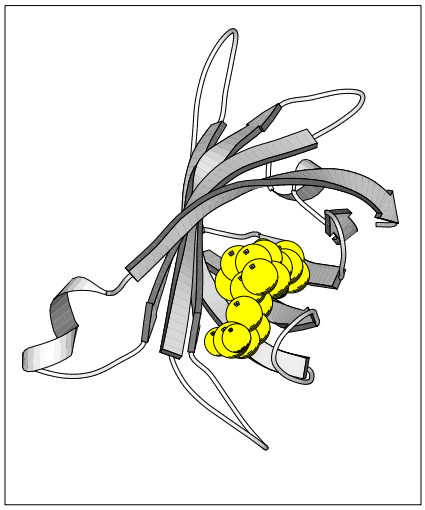

Figure 3: Understanding drug-protein interactions through shape complementarity.

ate representations to classify binding effectiveness. One example representation is the molecular skin, described in the paragraph on SHAPE REPRESENTATION, which is symmetric with respect to inside and outside. It is the only known surface representation that permits a mathematically perfect fit, at least when the criterion is limited to geometric complementarity. Although a perfect fit is unlikely to occur in nature, it is an ideal configuration against which we can measure imperfect fits. We expect to be able to use advanced measures of fit to classify drug molecules better than possible hitherto. Improvements on this front will lead to more accurate ways of capturing and differentiating shapes, leading to better SHAPE ORGANIZATION as discussed in Section 1.3.

Identifying binding sites for small molecules is closely related to the docking problem that studies the path along which a molecule finds its binding site. We will extend the applicability of motion planning from studying protein folding to following the path of ligand binding. Ligand motion occurs in a six-dimensional space (three translational and three rotational degrees of freedom), so it is an easier analysis problem than protein folding, even with simple representations. We refer to the paragraph on MOTION PLANNING in Section 1.3 for preliminary results. Docking and path searching methods could also prove to be useful in the study of enzymatic reactions. This is an important and difficult subject in the area of simulation. Although enzymatic reactions depend on a complicated quantum mechanical energy function, there are some simplifying principles that may allow geometric methods to be applied. The transition state in the reaction has to be complementary to the shape of the active site. Here again shape is defined in the broad sense. In many reactions, the structures of initial, intermediate, and final states are known to differ, making the use of motion planning appropriate. Collaborations between computer science and biology as represented on our team are essential for such work.

Structure Determination. When doing geometric computation for applications such as drug design and shape classification, it is of crucial importance to understand the source of the data and its characteristic errors. X-ray diffraction techniques remain the principle experimental access to high quality solved structures. The recent explosion of structural genomics and expanded access to X-ray sources at national synchrotron facilities necessitates the development of automated structure determination methods. Geometric tools can help in automation and in improving the quality of the extracted structure. Structure determination proceeds by first establishing the electron density in the crystal and then interpreting it in terms of an atomic model. The Fourier transform of the electron density in the crystal is a polynomial in sine and cosine with complex coefficients. An X-ray diffraction experiment measures the amplitudes (magnitudes) of these coefficients, but to reconstruct the electron density it is necessary to determine phase angles for the measured amplitudes. The quality of these phases determines the correctness of the electron density map. Available methods for estimating and refining phases introduce biases, which are difficult to detect and nearly impossible to eliminate. Rapid structure determination methods will therefore benefit from algorithms that ensure optimal phase choices before the electron density map is interpreted.

Results of Hauptman [42], culminating in the Shake and Bake algorithm, illustrate the synergy to be obtained by manipulating both the electron density in three-dimensional real space and the phases of its Fourier transform in reciprocal space. Figure 4 shows the essential features of a generic cycle of phase improvement. Modification of electron density

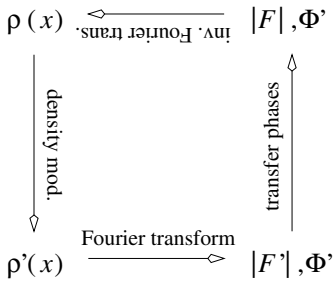

Figure 4: Generalized phase-improvement scheme in X-ray crystallography. It iterates between the density map ϱ in real space and its Fourier transform with amplitudes $|F|$ and phases Φ in reciprocal space.

in real space and phase refinement in reciprocal space are complementary because continuous properties in one space become discrete in the other, and vice versa. Rudimentary geometric constraints on the electron density and its deriva-

tives, such as the density histogram matching and solvent flattening, have already proven their value in improving electron density maps [53, 79, 80]. These methods are local and thus sacrifice much of the potential benefits of a global geometric constraint. Efforts to automate electron density map interpretation [33] suggest that Morse theoretic ideas can transcend these limitations. Since phase determination methods cope with inherently noisy data, we expect that global persistence measurements of critical points in this Morse theoretic view will further improve phase determination methods. Some of the details of this idea are explained in the paragraphs on HIERARCHIES and SHAPE ORGANIZATION in Sections 1.2 and 1.3. We plan to develop these ideas using novel representations of density maps, such as contour trees [14] and filtrations [27].

To fully exploit the iterative process outlined in Figure 4, it is necessary to refine the phases in reciprocal space in a manner complementary to the real space manipulations. Preventing the atoms comprising the electron density map from interpenetrating greatly improves the phase quality. The amount of interpenetration is measured in reciprocal space by the failure of the Sayre squaring relation [67, 68]. Potentially more robust refinement processes are possible by generalizing these relations [61, 62, 63]. A novel and powerful aspect of this particular iterative scheme is that crystallographic symmetry information can be seemlessly integrated into both sides of the computation. Morse theory makes no assumptions on the particular nature of the space where the electron density occurs. Therefore the orbifold generated by the crystallographic group can replace the usual unit cube and thus significantly reduce computational cost. Analogously, the calculation of convolutions necessary in the reciprocal space refinement can be reduced to fast Fourier transforms specifically written to take advantage of symmetry [11].

1.2 Modeling

Modeling issues are fundamental as they determine the feasibility and efficiency of algorithms. In this section we discuss some of the representations of molecules we plan to study and their properties. Building shape hierarchies and supporting motion and deformation are high priority items in this context. We also discuss representations for families of molecules, which we organize into shape spaces by providing concrete notions of conformational distance or similarity.

Shape Representation. A direct link between biology and geometry was developed almost thirty years ago by Frederic Richards and his collaborators [17, 50, 59]. They introduced a variety of *space-filling diagrams,* all based on the idea of viewing an atom as a sphere in three-dimensional space. The sphere center coincides with the location of the atom, while its radius depends on the atom type and is calculated from experimental observations of inter-atomic van der Waals forces. The envelope of the spheres is the *van der Waals* surface of the molecule. The interaction between the molecule and a solvent is studied by growing the spheres (as in the *solvent accessible* surface) and by filling up cavities and crevices that are too small for the solvent to enter (as in the *molecular* or *Connolly* surface). The historic success of these models proves their usefulness, and it comes despite seemingly serious shortcomings, including a rather low level of physical realism and the omission of flexibility and deformability in the representation. The former shortcoming will be addressed in the paragraph on HIERARCHIES below, and the latter can in part be remedied by fast data structures and maintenance strategies as discussed in the paragraph on DEFORMATION in Section 1.3.

Interesting biological molecules range from small ligands to macromolecules such as proteins and DNA strands consisting of hundreds of thousands of atoms. We believe there are contributions to be made to representations on all scale levels. We need to choose high-level representations based on the context and the function for which they will be used. The most effective representations grossly suppress some information or features, while preserving or possibly exaggerating other features of interest in the molecules. The significance of different types of information typically varies even over a single molecule, and carefully arranged combinations of different representations seem to give the best results. Figure 5 taken from [52] illustrates this point by displaying the pocket structure of the Staphylococcal nuclease protein using a sketch of the backbone, a space-filling diagram for the walls, and a dual complex for the pocket itself.

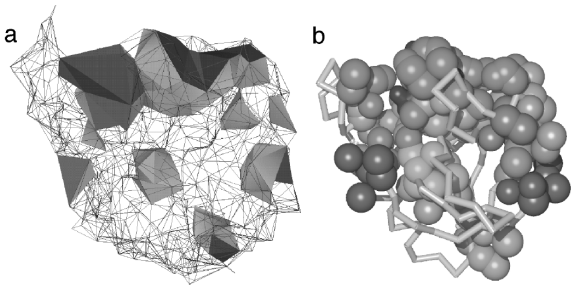

Figure 5: Staphylococcal nuclease with pockets and rest of Delaunay tetrahedrization in (a) and pocket walls and backbone in (b).

We spend the rest of this paragraph on one of our more mature ideas for an improvement of the traditional space-filling diagrams. Recall that the molecular surface attempts to smooth sharp edges and corners in the van der Waals surface by adding toroidal and inverse sphere patches that blend between the atom spheres. We propose a different blending mechanism using hyperboloids of revolution (1-sheeted and 2-sheeted) and inverse sphere patches; we call the resulting surface the *molecular skin* [25]. One of the advantages of the skin over the molecular surface is that it avoids the occasional self-intersections that happen when the rolling motion of the solvent sphere sweeps out a self-intersecting torus.

While the trouble created by the self-intersections may seem minor, they complicate the computation and are obstacles to constructing and maintaining surface meshes reliably and automatically.

Another useful property of the molecular skin is that it deforms gracefully and predictably under atomic motion. As described in [16], we can even give an unambiguous meaning to any partial deformation of one molecular skin to another or to any constant collection of others. In mathematical language, we can give meaning to any convex combination

$$B \quad = \quad \sum_{i=1}^{k} \gamma_i \cdot A_i,$$

where the A_i are different molecular skins and the γ_i are non-negative reals that sum to 1. Because convex combinations are unambiguous, we can use deformation to create a continuous space of molecular skins. We call this the *vector space* approach to shape space. A shape is indexed by the γ_i, which monitor the relative importance of the basic shapes A_i. This is reminiscent of work in computer vision that describes shapes by a fixed basis of spherical harmonics [70], or describes images by a combination of given images [76]. Dimension reduction by computing significant eigenvalues and their eigenvectors has proved useful in these applications, and is to be considered for molecular shapes as well. We contrast this construction of shape space with the more conventional landmark-based approach discussed in the next paragraph.

Shape Descriptors. Even though the concept of molecular shape is ubiquitous in molecular biology, it is not as crisply defined as several other molecular attributes. The descriptors for molecular shape usually depend on the intended application. Commonly used shape descriptors range from nuclear geometry (coordinates of atom centers) to electron density functions, from single numbers describing size or anisometry to vectors describing various geometric and topological properties. Also common is connectivity information, typically given in terms of the interatomic distance matrix and/or the distribution of dihedral angles (Ramachandran diagrams).

The description of shape through a collection of invariants can be regarded as a variant of *landmark-based* shape representations, in which points are used to mark important features. In this approach a shape is mapped to a point in a high dimensional space [12, 72]. In this area, work is needed to develop fitting shape descriptors, to enlarge the class of described shapes, and to improve measures of similarity. Some molecules produce interesting topological structures when in cyclic conformation, such as the circular double-stranded DNA. In these cases we might want to use topological invariants, such as the linking number, the tunnel number, or the knot polynomials of Alexander and of Jones. However, some of these invariants are rather difficult to compute and finding faster algorithms or good descriptors that are easy to compute remains a challenging open problem. As mentioned in the paragraph on ALPHA SHAPES in Section 1.4, we plan to write software computing the linking and the tunnel number of molecules so that the effectiveness of these invariants as shape descriptors can be assessed. Most of the early work on molecular shapes considers rigid structures only and does not take into account the conformational flexibility that characterizes many molecules. Recent progress has been made in this direction, but more sophisticated models are warranted.

The problem of measuring the *similarity* between molecules arises in both PROTEIN DESIGN and DRUG SHAPE CLASSIFICATION, discussed in Section 1.1. In some applications one wants to determine the similarity between two molecular shapes. In others, one wants to determine whether a substructure of a shape is similar to (or complementary to) another shape; this is analogous to the global vs. local distinction in sequence alignment algorithms. In the last few years there has been much work on matching shapes in computational geometry [5, 77], computer vision [38], databases [31], and other computer science areas, and this is still a major research topic.

A commonly used measure of the similarity between two molecules, or portions of molecules, is the *cRMS distance*, which is defined as the square root of the average square distances between corresponding atoms in the molecules, when the two molecules have been optimally aligned. A few other similar functions have also been proposed. In the absence of the optimal correspondence, finding the optimal alignment is hard, and approximation algorithms must be used. Taking steric and chemical properties of molecules and flexibility of molecular surfaces into account poses additional challenges to the matching problem. If molecules are described by electron density functions, the similarity between two molecules can be measured by the overlap of their density functions. Work in the database and computer vision communities measures the similarity of two-dimensional images using histogram or feature-based analysis [31, 57], or by considering morphing cost measures such as the earth mover's distance [66]. We plan to study extensions of these approaches to measure the similarity between electron density measures. If molecules are, instead, described by a space-filling diagram, the similarity between two molecules can be measured by the overlap of the molecular surfaces. Since spherical surfaces are rather difficult to compare, they are often replaced by a set of points sampled on the surface or by a triangulation of the surface. Weights can be assigned to different regions depending on their importance. There is work on combinatorial algorithms for matching point sets in $\mathbb{R}^3$ and on matching rigid polygonal objects in the plane and on numerical algorithms for matching polyhedral surfaces in $\mathbb{R}^3$ [10]. We plan to improve these results and combine the advantages of the combinatorial and numerical approaches to the problem.

Hierarchies. Hierarchical representations are a computational necessity when dealing with the vast amounts of data in molecular biology. Simplified models of proteins were introduced in the earliest simulations of protein folding [51] and have proved popular and effective until the present time [48]. They are built bottom-up through stepwise simplification, which is appropriate whenever we start at a fine level of detail. We will work on simplification procedures that combine knowledge about geometry, physics, and topology into a single hierarchical representation.

Efficient methods for geometric simplification for triangulated surfaces embedded in $\mathbb{R}^3$ have been developed during the last few years in the computer graphics community. A number of algorithms have been suggested, and the most versatile among these simplifies the surface through repeated edge contractions [44], with the sequence controlled by a locally evaluated quadratic error measure [35]. Local combinatorial tests as described in [22] are used to disallow edge-contractions that would change the global topology of the surface. The error measure can be modified to take into account attribute functions given over the surface [36], which can be chosen to represent physically meaningful attributes, such as energy. Each attribute is interpreted as another dimension and worked into the error function as such. All ingredients of the method generalize from triangulated surfaces to tetrahedrized volumes, which is significant in our work where we deal with energy functions over domains that are more often three- than two-dimensional, see the paragraphs on SHAPE ORGANIZATION in Section 1.3 and on MOLECULAR DYNAMICS in Section 1.4.

A significant weakness of the above geometric simplification method is the inability to deal with topological features in a meaningful way. In three-dimensional shapes such features manifest themselves as perfect or imperfect holes in the shape or its complement [26, 56], and in density functions they show up as critical points [54]. We elaborate on the latter case, which has direct applications to STRUCTURE DETERMINATION as discussed in Section 1.1 and further ramifications in medical imaging and in disciplines where shape information is derived from experimental density data. Consider the 1-parameter family of iso-surfaces of the density function defined by varying the density threshold. The global topology of the iso-surface changes whenever the threshold passes the density value of a critical point. As shown in [27], each topological feature can be marked by a critical point that creates and another that destroys it. The difference in density values of the two markers is the life-time, which we call the *persistence* of the feature. It is a measure of the set of iso-surfaces that contain a given feature, and we can use it to identify noise within the set of topological features. We eliminate the noise by smoothing the density function, which effectively moves paired markers towards each other until they collide and disappear. We plan to work (1) on identifying the numerically most effective way of smoothing and (2) on obtaining density dependent decompositions of the domain that can be used to control and localize the effect of smoothing. For this we need a good understanding of the obstructions in making pairs of critical points collide and of methods that can get around them.

The effect of simplification based on topological persistence is illustrated in Figure 6. The protein Gramicidin is represented by a sequence of simplicial complexes (technically a *filtration*) modeling its appearance at different scale levels. The filtration records the evolution of a growth process. During that process, the topology changes and we can again measure persistence by identifying the times when a feature is created and when it is destroyed. To define what exactly that means is straightforward for components and voids but more challenging for tunnels, which are counted by the first Betti number, $\beta_1 \geq 0$. The dominant feature of

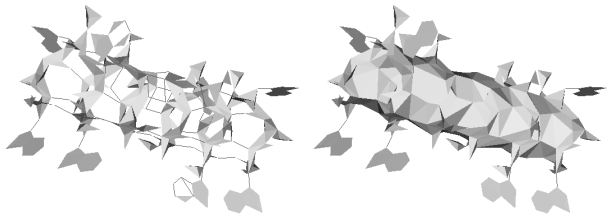

Figure 6: Alpha shapes of Gramicidin protein computed with zero and substantially positive persistence thresholds.

Gramicidin is a tunnel whose biological function is the transportation of ion particles along the channel. We monitor β_1 during the growth process and observe a fairly complicated evolution represented by a rather noisy function from time to the non-negative integers. As described in the paragraph on ALPHA SHAPES in Section 1.4, that function is made up by the contribution of the persistent ion channel and a large amount of noise counting tunnels created and destroyed in quick succession. The tunnels with low persistence are eliminated by local rearrangement of the growth evolution. The effect is illustrated in Figure 6, where the left picture shows the dual shape of Gramicidin at a certain scale level and the right picture shows the shape at the same scale level but topologically simplified so that all tunnels other than the ion channel are eliminated.

1.3 Algorithms

In order to build the infra-structure of geometric software tools for molecular work, we plan to design and study a variety of algorithms. These include meshing surfaces and volumes in space, motion planning, maintaining representations through motion and deformation, and organizing shapes for fast shape similarity queries.

Meshing. *Meshes* of geometric domains or spaces are decompositions into simple pieces usually referred to as *elements*. Common uses of such decompositions are the nu-

merical solution of differential or integral equations expressing physical reactions on these domains, their visualization, etc. There is a great variety of meshes in use, distinguished by the dimension of the domain, the shape of the elements, and the way the elements are connected. We plan to work on algorithms that mesh surfaces with triangles and mesh spatial domains with tetrahedra. We will base our three-dimensional algorithms on weighted and unweighted Delaunay tetrahedrizations for which we already have fast and reliable codes.

For a two-dimensional molecular skin surface F, we construct a triangle mesh D which adapts to the shape, curvature, and topology of F. Using results in [29] we show how to construct D so it is homeomorphic to F. Besides having matching topology, we require that D approximates F geometrically, and that the mesh is denser where the surface has higher curvature. Figure 7 illustrates the latter condition by showing the triangulation of a point set sampled from the skin. We see an accumulation of points around the narrow neck connecting the less curved surface pieces on its two ends. A uniform sampling strategy follows the right distribution but does not have any local spacing guarantees. Using edge contractions to eliminate short edges and point insertions to fill gaps we can turn the uniform into a semi-regular sample of predictable local mesh size. The triangulation of that sample can have neither small nor large angles and is thus well suited for finite element methods computing electrostatic charges or other physical functions over the surface.

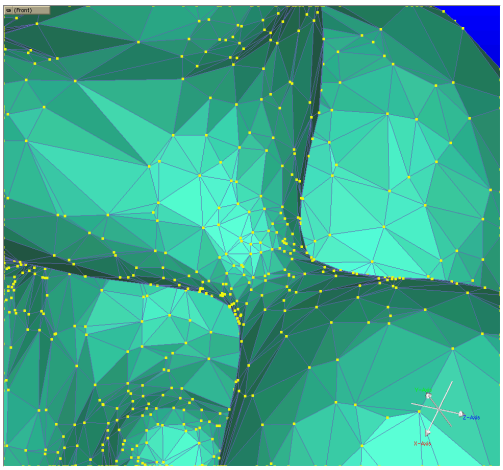

Figure 7: The points are sampled uniformly from a density inversely proportional to the local maximum normal curvature.

Dynamic meshes are needed to capture motion, flexibility, or changing scales levels. We can compute the mesh D for various scale levels by continuously growing the spheres representing the atoms. During this process F changes its shape, its local curvature, and occasionally also its topology. As described in [15] we can use point movement and flipping to adapt D to the changing shape, edge contractions and point insertions to adapt D to changing curvature, and local adjustments of connectivity to adapt D to changing topology. We envision an implementation where, instead of using fixed time steps, the operations are scheduled in time and executed in that sequence. Each operation affects only a local portion of D and thus takes little time to perform. Of course, we may get time-warp effects where one portion of D develops faster than another, but the difference in local time can only vary slowly. One of the difficulties of the scheduling approach to mesh maintenance is that determining the time at which an operation will mature is a somewhat expensive root computation. We can avoid it by executing the operations in a partial rather than a total time order. This increases the time-warp effect, and, more seriously, it may lead to inconsistencies that jeopardize the correctness of the algorithm. A mixed strategy where, for example, we execute topological changes according to time and all other operations lazily might be the right way to proceed.

What exactly we can implement and what we can prove depends not only on the meshing algorithm but also on the type of motion we allow the atoms to perform. Deformation by growth is the simplest type and the only one we understand in detail. More general deformations will require extra mechanisms looking out for events to come and forewarning the algorithm early enough so it can schedule and execute all necessary actions before another event accesses the same portion of the surface. All these issues lead us to the study of kinetic data structures, which are precisely such event-scheduling mechanisms discussed in the next paragraph.

Deformation. Molecules move and deform. Chemical processes essential to life critically depend on the ability of biomolecules to adopt different shapes over time. If we could describe realistic molecular motion and execute that motion efficiently with our representations, we would greatly aid the understanding of these temporal processes. At the atomic level, the physical principles underlying such motions and deformations are fairly well understood: proteins have primarily torsional degrees of freedom, the forces on an individual atom can be expressed as sums of well-known potentials, etc. Yet *ab initio* physical or Monte-Carlo simulation of such molecular systems, as practiced in molecular dynamics, is still far too expensive computationally to allow us to follow these processes over the time period during which the phenomena of interest take place. In order to make progress we must develop novel ways to represent deforming shapes, find techniques for describing the physics in terms of higher-level units than individual atoms, and efficiently track the spatial conformation of the molecule.

In classical molecular dynamics, proximity relations among atoms are determined by overlaying a regular voxel grid on the space and then tracking individual atoms as they move among the grid cells [39]. This can work well for moving objects (the atoms in this case) which are all of roughly similar size, but no grid size would fit well if we were trying to do the simulation at the level of, say, alpha helices and beta

strands, which are elongated and exhibit themselves much greater shape variability. Instead of using a fixed spatial partition, such as the voxel grid, we might do better inventing new spatial partitions that are *conforming* or *adapted* to the moving geometry. We propose to develop new techniques for representing and simulating deformable shapes based on such conforming partitions and adaptable surface and spatial meshes (or topological equivalents thereof), a topic already discussed in the paragraph on MESHING above. As mentioned there, a crucial aspect of such deforming meshes is that the mesh topology will change at certain discrete time instances, as the mesh has to adapt to the moving or deforming geometry.

The issue of detecting when remeshing is needed is best studied in the context of *kinetic data structures*, developed by members of this team [7, 37]. These are event-based data structures aimed at efficiently tracking attributes of interest in an environment of moving or deforming objects. A kinetic data structure certifies the identity of the attribute of interest, say the identity of the closest pair of atoms in a MOLECULAR DYNAMICS simulation, by asserting a set of elementary conditions about the world, called *certificates*, whose validity mathematically proves the correctness of the attribute value. A well-designed kinetic data structure exploits continuity to maintain the certificate set and the corresponding attribute value by small incremental steps. Several kinetic data structures for maintaining closest pairs of moving points in all dimensions are known [8, 7]. We developed a new structure for tracking the closest pair among moving balls or arbitrary radii based on their power diagram in $\mathbb{R}^3$. These balls may be atoms, or bounding volumes for molecules or parts thereof. Although it is too early to know how this method compares to tracking the balls using a fixed voxel grid, we believe it will outperform the grid in cases where atoms form aggregates that are themselves static but move relative to other aggregates. In support of this belief, we present results from a related experiment: we compare an axis-aligned box method for motion tracking [55] with the new kinetic data structure. The goal is to track a set of balls or boxes that move linearly and bounce against each other inside a cube. Figure 8 shows the performance of the two methods for data sets ranging up to a thousand items and followed for sufficiently long simulation time to obtain a reliable number of events and collisions per second. As we can see from the graph, the number of events processed by the kinetic data structure is roughly linear in the size of the data, while the number processed by the box method is roughly quadratic.

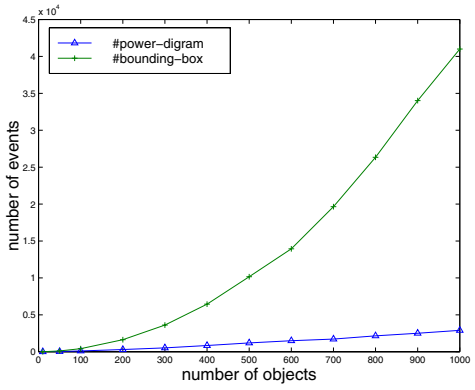

Figure 8: Comparison between the kinetic data structure based on the power diagram and a bounding box method.

Motion Planning. In PROTEIN DESIGN and DRUG SHAPE CLASSIFICATION, understanding which motions a molecule is able or likely to perform will support prediction of ligand binding and dissociation rates, which are critical parameters of drug activity. Considering the deformations of the protein (both of its backbone and of its side chains) as the ligand comes close to the binding site can help us discover more potent drugs with significantly smaller dissociation rates. Molecule motions are guided by potentials, such as electrostatic and van der Waals, which are used to determine minimum-energy conformations of proteins and ligands. Tools that represent and compute motion during molecular activity are still imperfect. There are several reasons for this: most available energy models are imperfect and/or expensive to compute; classical techniques, such as steepest gradient descent, do not work well due to the imperfect models available; the problem is complicated by the large number of degrees of freedom, making for huge spaces of possible motion. We have proposed some new tools to represent and compute motion in the paragraph on DEFORMATION.

We believe that the framework of *probabilistic roadmaps* developed for robot motion planning [46, 45] offers a promising approach to predict plausible molecular motions. These roadmaps have been successful in solving complex motion planning problems for robots and animated characters with many degrees of freedom in the presence of tight dynamic constraints and moving obstacles. The principle is to capture the connectivity of the robot's collision-free space M by sampling it at random. While computing an explicit representation of M is impractical for robots with more than 4 or 5 degrees of freedom, testing if a given robot placement (or configuration) is collision-free is fast, even in geometrically complex environments. A roadmap planner typically samples thousands or hundreds of thousands of robot configurations at random, retains those that are collision-free, and tries to connect every pair of configurations that are close to each other by a simple collision-free path. The end result is a network, called a probabilistic roadmap R. Under reasonable assumptions, it has been shown that with high probability R correctly captures the connectivity of M. This means that every point in M can be connected by a simple path to a node of R and that M and R have corresponding connected components. It was proven that the probability that R fails to correctly capture the connectivity of M goes to zero exponentially in the number of nodes of the roadmap.

By viewing a molecule as an articulated object and mapping each degree of freedom to a dimension, we can map each conformation of the molecule to a point in an appropriate configuration space. Random sampling a conformation is equivalent to randomly picking a point in that space. Such sampling techniques offer a viable approach for dealing with both the high dimensionality of configuration spaces and the complexity of the probability distributions modeling energy functions. However, applying the framework of probabilistic roadmaps to predict molecular motions will require several key developments, including the extension of binary to continuous functions over the configuration space. In contrast to robot configurations, which are either colliding or not, the configuration space of a molecule is the domain of a continuous energy function. Likely motions correspond to curves traversing regions of low energy. This suggests that in the molecular case the sampling process of a planner should attempt to place many samples in low-energy regions and fewer samples in high-energy ones. The plausibility of a local path should evaluate the likelihood of the molecule to transit from one node of the roadmap to another through that path.

An energy-based roadmap in a molecule's configuration space would form a compact representation of energetically and sterically possible docking motions. In fact, a version of such a planner described in [71] has been used to generate plausible ligand motions and successfully predict binding sites in three known cases:

- receptor: mutant of tyrosyl-transfer-RNA synthetase (2423 atoms, 319 residues), ligand: L-lencyl-hydroxylamine (13 atoms, 9 degrees of freedom);
- receptor: Lactate Dehydrogenance (2386 atoms, 309 residues), ligand: Oxamate (6 atoms, 7 dof);
- receptor: Streptavidin (901 atoms, 121 residues), ligand: Biotin (16 atoms, 11 dof).

Figure 9 shows a docking path generated by this planner. Our research will aim at biasing the sampling toward the most interesting regions of such a space, and at using a hierarchical representation of distributions. The hierarchy can, for example, be used to build a large roadmap using simplified and thus computationally cheap energy models. The roadmaps can then be pruned back and refined using more accurate and complex energy models.

Shape Organization. Three-dimensional structures of small and large molecules at atomic resolution are being generated at a record rate, both by X-ray crystallography and by NMR spectroscopy, and large databases (PDB) have been compiled [3, 9]. The more urgent issue now is to design efficient procedures that search in these structure databases. For example in PROTEIN DESIGN we will frequently search for structure elements that satisfy given constraints, and in DRUG SHAPE CLASSIFICATION we will often need ligands

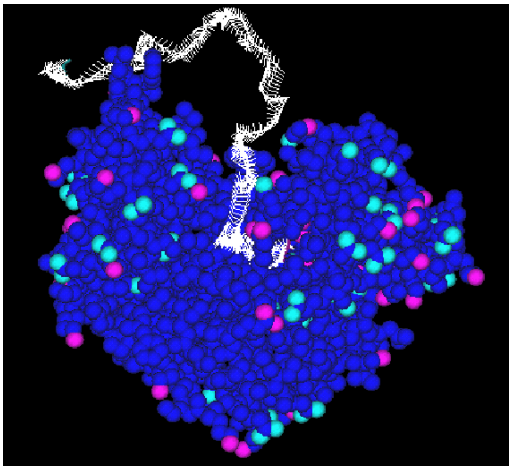

Figure 9: Low-energy path of ligand planned by probabilistic roadmap algorithm.

with given geometric and steric properties. In order to answer these queries efficiently, we need indexing schemes to store the molecules. Despite much work on indexing point sets, two-dimensional images, and polygonal chains [1], little is known about indexing three-dimensional shapes [47]. Standard indexing schemes such as R-trees and others are limited in their effectiveness as the dimension of the data increases, as it does in molecules. Effective reductions of the dimension can be achieved by identifying clusters along affine subspaces of the shape space. Recent efforts are encouraging [2, 4, 32], but developing fast algorithms for finding such clusters remains a challenging problem.

As a step toward more effective organization of proteins we propose to systematically study the conformational correlation among adjacent residues. Our goals are:

- Define the structurally significant fragments of protein structures. By clustering all protein fragments in a database, we will generate a library of non-redundant fragment conformations from which it is expected that virtually all protein structures can be constructed. Conversely, it will be possible to describe a protein structure via its content of such fragment conformations. This will generate a simplified representation of protein topology and may drastically speed up protein structure comparisons.
- Understand the relationship between the local sequence of a fragment and its conformation. A protein amino-acid sequence defines a unique three-dimensional structure under physiological conditions. This is not true, however, for short fragments, which have been observed to adopt different conformations in proteins from their intrinsic conformation in solution [41]. A systematic study of the diversity in sequence of fragments with similar conformation will generate valuable information on the relationship between protein sequence and structure.

Motivated by indexing and classifying molecules, we propose to develop efficient algorithms for clustering molecules, which not only take geometric but also steric properties into account. Since the data sets are large, efficiency of the algorithms is of utmost importance. Suppose that we want to cluster protein fragments using the cRMS or another notion of distance between shapes. Since a single distance computation takes a noticeable amount of time, it will be expensive to compare every pair of molecules. Worst, storing all pairwise distances among the quarter-million or so fragments in the PDB is prohibitive. One of the challenges is to find linear-size data structures that support the fast and sufficiently accurate evaluation of distances between molecules. We propose to build a low-dilation graph spanner of the fragments, that is a sparse graph whose vertices are the molecules and edges are weighted by cRMS distance such that the distance between any two molecules is roughly the length of the shortest path in the graph. Such graph spanners have been studied for discrete sets of points in Euclidean spaces [13, 19, 65] but not yet for more complicated 3D shapes described above.

Preliminary results have been derived on classifying fragments of protein structures in a database that contains 1,500 protein structures, corresponding on average to 250,000 different fragments. The fragment sizes vary from 7 to 20 residues. We first apply a hierarchical clustering on random subsets of the database, and then perform a supervised classification of the remaining fragments using a fuzzy version of the k-near neighbor heuristic. Examples of clusters found for fragments of size 7 are shown in Figure 10. As expected, the most common structural motifs such as helices and strands were identified (clusters A and B in Figure 10, respectively). Other relevant fragments include turns (cluster C), as well as fragments at the junction between turn and standard secondary structure elements (cluster D).

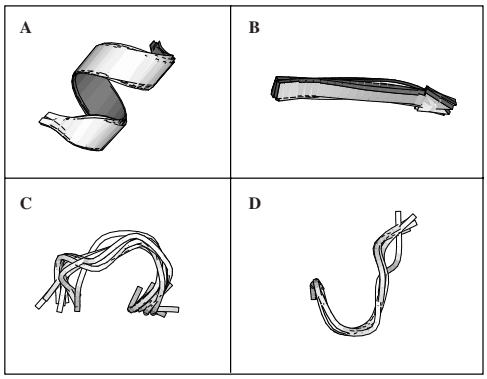

Figure 10: Four significant clusters found for protein fragments of length 7. For the sake of clarity, only a small sample of the clustered fragments is shown. The protein representations were generated using MOLSCRIPT [49].

1.4 Software

One of the more tangible results of this project will be the creation and distribution of integrated and user-friendly software. While industrial strength software seems beyond the scope of this proposal, we plan to release tools that can compete well in academic and industrial research environments.

Alpha Shapes. Several versions of the alpha shapes software have been released for public consumption in the past by the principal investigator [28] and the software has reached a great number of users. We plan a complete overhaul, aiming at a modular structure achieved with the help of modern software design practices. It should be possible to rearrange the modules without much programming effort and to decouple modules for integration in other packages. We will add new functionality and, through collaboration with the natural scientists on the team, bring the tool closer to the wider community of scientists. There are five major new functionalities that will take the alpha shapes software to the next stage:

- topological persistence and simplification,
- meshing of molecular skin,
- motion and deformation,
- computation of volume and surface area derivatives,
- detection of knots and links.

Some of these functionalities will require new visualization techniques and virtual reality methods to reach their full potential. All but the last extension are discussed in other sections of this proposal. Here we address a typical user-interface challenge that arises in the interaction with sophisticated geometric data, such as the family of alpha shapes indexed by scale and persistence. We also discuss the detection of knots and links in molecular data.

Topological simplification adds a hierarchy orthogonal to the scale hierarchy of shape representation. We use Gramicidin from Figure 6 to illustrate what we mean. Figure 11 plots the number of tunnels as a function of a scale level α and persistence p. The cross-section for persistence $p = 0$ is the 1-dimensional histogram computed and displayed in the current alpha shapes software, which records the number of tunnels as α increases from left to right. We see a noisy function that hides the existence of the one dominant shape feature, the tunnel passing long-side through the protein shown in Figure 6. As we move the cross-section forward by increasing the persistence, we decrease the noise until at about $p = 2,700$, the trace of the dominant tunnel is the only feature left. The difference between $p = 0$ and $p = 2,700$ is also illustrated in Figure 6, which reconstructs the protein for the same scale and two different persistence levels. The user interaction with 2-dimensional histograms in the selection of shapes is a new paradigm with interesting design questions.

Figure 11: The number of tunnels in Gramicidin in Figure 6 depends on scale (coarsening from left to right) and persistence (increasing from back to front).

Knots and links are topological features that cannot be detected with homological information alone. Although biomolecules are unlikely to have such features, nobody knows for sure. On the other hand, knots and links could be unwanted by-products of protein design steps that may remain undetected without an automatic tool as described. We base our approach on the tunnel number, which vanishes if, and only if, the knot is the unknot. Intuitively, this is the number of times the knot must pass through itself to get unknotted. We generalize this number to complexes embedded in $\mathbb{R}^3$ and determine knotting by deciding whether or not the tunnel number vanishes. It does iff the exterior of the complex is a union of handlebodies, each obtained by thickening a 1-dimensional graph or equivalently by gluing handles to a solid ball. While we have no polynomial time algorithm for recognizing handlebodies, there is hope that even for large molecules we can compress the exterior into a very small complex. Whether or not the exterior is a handlebody can then be decided by an algorithm based on the theory of normal surfaces [40]. We do not expect a fast algorithm but one that can compute valuable information that is otherwise difficult or impossible to obtain. The function that finds knots and links will thus not be generally executed and, if invoked by the user, it will be applied only to the current complex or molecule and not to the entire filtration.

Molecular Dynamics. As mentioned in the paragraph on DEFORMATION in Section 1.3, dynamics is essential for a macromolecule to function. Recent improvements in computer architectures and in computational techniques have enabled more accurate simulations of molecular motions in realistic environments [23]. Most of these simulations involve models built from discrete interaction centers, be they atoms, residues, or protein domains. Motion is induced by solving Newtonian equations, by an analytical normal-mode approximation, or by random Monte Carlo moves. Normal-mode approximation assumes that the potential is perfectly quadratic, which is almost never true, whereas the other approaches allow the exploration of a huge portion of configuration space by wandering aimlessly through it. Similar problems, but on a much reduced scale, are encountered in robotic motion planning.

Accurate molecular dynamics simulations remain a major challenge in computational biology, since they involve thousands of degrees of freedom in the molecule of interest in addition to the need to account for the water environment. They need to span a large variety of functionally significant time scales, from nanoseconds to seconds. Computer simulations that include a large number of water molecules remain the state of the art in this field. They are, however, inefficient, since a large fraction of the computing time is spent calculating a detailed trajectory of the solvent molecules, even though it is primarily the solute behavior that is of interest. Within these limitations, it is remarkable that a one micro-second simulation has been done [24]; this study remains isolated however, and its application is limited to small bio-molecules. It is therefore desirable to develop different approaches in which the effect of the solvent is taken into account *implicitly*. Such treatment would make it possible to perform simulations covering much longer time intervals, and including much larger molecular systems. The difference between explicit and implicit representations is illustrated in Figure 12.

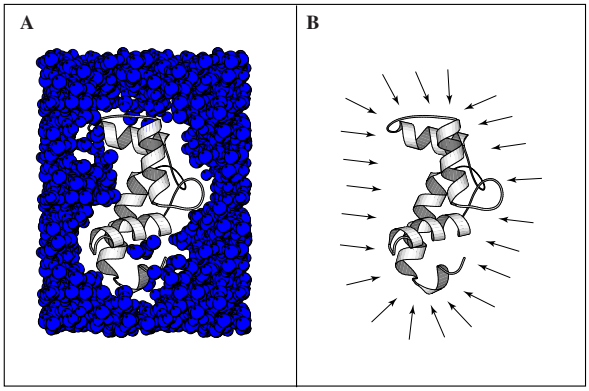

Figure 12: Molecular dynamics simulation of a small protein, the N terminal domain of Troponin C. (A) View of the inside of the initial box for a simulation that includes water molecules explicitly; the whole system consist of 1229 atoms in the protein and 6009 atoms for water. (B) In an implicit solvent simulation, the effect of water is included in a potential of mean force visualized as arrows. In this case, the simulation only includes the 1229 protein atoms.

All solvent effects on a molecule X can be included in an effective potential, or potential of mean force $W(X)$. Implicit solvent potentials commonly decompose $W(X)$ in terms of electrostatics and non-polar contributions:

$$W(X) \quad = \quad W_{elec}(X) + W_{np}(X). \qquad (1)$$

$W_{elec}(X)$ is usually represented by continuum electrostatics, for which several semi-analytical approximate treatments have been proposed [20, 69, 73]. All these treatments describe $W(X)$ as a pair potential, which can easily be incorporated in a molecular dynamics program. The situation is not so simple for $W_{np}(X)$. The contribution of non-polar interactions to the effective potential $W(X)$ is usually related to the solvent exposed surface area of each atom of the solute, $W_{np} = \sum_{i=1}^{n} P_i \cdot A_i$, where A_i is the accessible surface area of the i-th atom, P_i is the atomic solvation parameter, and the summation extends over all n atoms. This approximation is simple and physically sound [64], and consequently widely used in bio-computation [30, 34, 78]. It is, however, a multi-body potential.

Inclusion of $W_{np}(X)$ based on equation (1) in a molecular dynamics simulation requires the calculation of accurate molecular surface areas, as well as their analytical derivatives with respect to atomic position. For that purpose, we plan to combine the Alpha Shape software with the molecular dynamics software ENCAD both developed by members of this team. One of the new features this brings to ENCAD is the computation of the individual contributions A_i to the accessible surface area $A = \sum A_i$. By multiplying A_i with the solvent parameter P_i we then get the non-polar contribution to the mean force. Similarly, we can use the Alpha Shapes software to compute the area derivative efficiently. Given the molecule in a given state described as a point in $3n$-dimensional space, the software uses local operations within the dual complex to derive a vector $\mathbf{v} \in \mathbb{R}^{3n}$ so that for each motion vector $\mathbf{u} \in \mathbb{R}^{3n}$ the area derivative is the scalar product $\langle \mathbf{u}, \mathbf{v} \rangle$. Each atom directly determines three coordinates of $\mathbf{v}$, and we can again multiply with the solvent parameters to get the derivative of the non-polar contribution to the mean force. The method for computing the derivative of W_{np} is new and was discovered during the preparation of this proposal; details will be published.

Visualization. It is no surprise that computer graphics workstations and even virtual reality have been heavily used in biology. There is so much information in molecular data, that reliance on the visual system is necessary for sufficient bandwidth. The project team has a long track record of producing tools that help biologists visualize and analyze molecular structure and activity. For example, UNC Chapel Hill hosted one of two NIH Research Resources in Molecular Graphics from 1974 to 2000, under Brooks as PI. Although the Resource has recently changed direction to capitalize on opportunities in virtual-environment interfaces for probe microscopes, we continue to have considerable expertise, software, and unique facilities for real-time 3D interactive molecular visualization.

The UNC Resource built the first computer graphics molecular structure determination system on which proteins/nucleic acid structures were solved without the construction of physical models. David and Jane Richardson in 1974 solved bovine Cu, Zn superoxide dismutase using the UNC GRIP graphic system [60]. Petsko then solved sea-snake erabutoxin [75], and Kim solved yeast-phenalynine tRNA [74]. Over 40 teams of molecular structure scientists have come to Chapel Hill to use the unique facilities. All systems at the UNC Resource have exploited the expertise of molecular scientists by using real-time interactive graphics to facilitate interactive steering of high-performance computing, or conversely, using real-time high-performance scientific computing to aid interactive structure study. This has proved to be a powerful mode of investigation, and we propose to continue this tradition.

We will use our in-place visualization capabilities to speed the exploration and testing of many algorithms. These capabilities range from 3D wax-deposition printer at Duke to visualization stations such as the UNC Protein Interactive Theater (PIT). This two-surface, rear-projection system provides high quality stereo realizations of molecular structures for one or two seated users. Figure 13 shows the configuration, including control devices such as 6 degrees of freedom flying mice, force-feedback joysticks, and auxiliary keyboards and screens. The latter display the scientists' plans for experimental sessions and serve as laboratory notebooks for observations and collection of result files.

Figure 13: The Protein Interactive Theater (PIT) supports collaboration.

The structures under study are displayed in the viewing space over the workbench, hung out in front of the viewing screens. Density maps and other imagery can be superimposed. Two viewers can choose to see orthogonal views of the same three-dimensional structure, allowing each to point out locations to his/her colleague. Alternatively, two users can choose to have identical views and virtual laser pointers. These types of interactions between a biochemist and a computer scientist are necessary to develop effective tools for such geometric tasks as electron density modification in real and reciprocal space, or for protein design.

Real-time, interactive visualization of new algorithms operating on real datasets has always proved to be illuminating. We plan to use, rather than build new, virtual reality equipment and engage in content-providing activities. For

example, the PIT system and elaborate user interface software is separate from the molecular applications that have been developed. We plan to use the system software as is, and develop specialized application software as the research in the project proceeds.

1.5 Summary of Novel IT Research

The PITAC report emphasized the transforming nature of Information Technology. This project seeks to build computational tools that can transform the way we determine, represent, manipulate, simulate, reason, and act upon the essential molecular structures and processes of life. In the process of doing so we will have to create new representations and algorithms for the analysis of dynamic structure and form, develop spatial and temporal hierarchies to make feasible the manipulation of complex shapes and processes, integrate discrete and continuous methods, and build new computational bridges between physics, geometry, and probability. We have already mentioned the following computer science challenges that we plan to address:

- Topology persistence and simplification, discussed in the paragraph on HIERARCHIES and ALPHA SHAPES.
- Shape classification and indexing, discussed in the paragraphs on SHAPE DESCRIPTORS and SHAPE ORGANIZATION.
- Energy-based sampling, discussed in the paragraph on MOTION PLANNING.
- Deformation and flexible shape representation, discussed in the paragraphs on MESHING, DEFORMATION, and SHAPE DESCRIPTORS.
- Advanced physical simulation, discussed in the paragraphs on DEFORMATION and MOLECULAR DYNAMICS.

Progress in these fundamental computational areas can enable many other activities related to Information Technology. Examples are: modeling of human organs and tissue; modeling of fabric and other non-rigid structures, as well as the robotic manipulation of such; building and querying 3D mechanical parts databases; simulation of biological processes at the cellular level; and many more.

2 Impact

By taking a geometric approach to the research flow-diagram of Figure 1, this project can have a significant impact on the tools for molecular biology and, in the long term, on the health and well-being of American people and business. The problems of drug design and molecular analysis motivate our research into studying models and representations for molecular structure. On these we can base well-founded algorithms that underly software tools for simulation and analysis of fundamental molecular activity, such as folding and docking. Dissemination of the results of this geometric research in the form of software tools makes them accessible outside of the discipline of computer science, and complements the usual academic means of dissemination by the education of students. The success of this project depends upon collaboration: linking the computer science researchers who study geometric form with biology and chemistry researchers who study biochemical function.

2.1 Collaboration

In a project of scope as broad as we are proposing, multi-institute and multi-disciplinary collaboration is crucial, as no single department or university has expertise in all the areas. The project rests upon strong interdisciplinary collaboration within the three lead institutions, and upon a long history of collaboration between the PI's in computational geometry, as well as a long history of personal interactions between the PI's in structural biology. These collaborations include the following groups and topics: Agarwal, Edelsbrunner, Snoeyink, Zomorodian on density modification based on topological persistence; Agarwal, Guibas, Snoeyink on kinetic data structures for deformable meshes; Brooks, Levitt on harmonic analysis of molecular vibration modes; Carter, Latombe, Levitt on robot planning to plot enzymatic reaction paths; Bryant, Edelsbrunner, Koehl, Levitt on area and volume derivatives; Guibas, Koehl, Kolodny, Levitt on clustering algorithms to analyze protein fragments; Hellinga and Koehl on sequence design methods in active site re-design. These on-going efforts, together with the potential for collaboration on topics such as electron density modification, metal-binding sites, and flexible shapes, made this large team a natural choice fitting with the PITAC recommendation [58] for "projects of broader scope, longer duration, and a renewed emphasis on research carried out in teams."

Specific measures to facilitate and encourage collaborations will include the following.

- Live interaction that will include a monthly meeting at which all groups will participate directly, by video or telephone conference call. This will be augmented by an annual meeting of all PIs. The annual meeting will be arranged to precede or follow an annual workshop, designed to meet education and outreach goals.
- Regular visits by PIs, post-doctoral fellows, and graduate students among the participating institutions, and exchanges of students and post-doctoral fellows. There is already a history of such activities between the computational geometry PIs.
- Installation of a Molecular Geometry web site that will contain the proposal, any papers in preparation, white papers, and serve as a means to exchange files. Large parts of the site will be publicly accessible.

- Shared teaching encouraged in two ways. At the same institute people in computer science and computational biology will team up to give joint courses. These courses will be recorded and broadcast over the web as is already done with many courses currently given at Stanford. Once we have material for different courses, we will attempt to team-teach by mixing lectures from the different institutions.

2.2 Education

This project seeks to address education and training goals at undergraduate, graduate, and post-doctoral levels. The project will produce students with broad knowledge in both biology and computation, a combination that is in high demand. Teaching will be integrated with research, and students will be encouraged to develop their creative capabilities as early as possible. Undergraduate research is not only a strong catalyst for student development, it also attracts bright young people to the field. It can be a major factor in career decisions. We will seek for REU supplements to support this activity. Efforts will be made to increase the involvement of under-represented members of society in the research. This will be achieved through the partnership with NCA&T in setting admission criteria and in recruitment of minority students.

At the level of programs, the faculty in this project see opportunities to create combined undergraduate major programs or strong major/minor programs in biology and computer science. One effort is a NCA&T proposal for a computational sciences program including faculty from Physics, Chemistry, and Mathematics, which we support. The proposed Centers for Genomics at Duke and for Bio-X at Stanford will play a pivotal role in developing similar programs at the corresponding institutes. At Duke, Stanford, and UNC a number of biocomputation courses are already offered this Spring semester. There are plans to develop undergraduate majors in biocomputation with a core curriculum comprising mathematics, physics, chemistry, biology, and computer science.

2.3 Outreach

It is often said that biology is becoming an information science. Unfortunately, the sciences of biology and information have historically developed in separate departments, which is a barrier to applying sophisticated geometric representations and advanced algorithms to problems in computational structural biology. We envision establishing a stronger link between the two communities as one of the major goals of this project. In order to bring the people from two disciplines together, we propose to organize two workshops, each inviting about 25 people, early in the project. The first workshop, which we plan to hold in the Research Triangle area of North Carolina in Spring 2001, will focus on the state of and need for software for structural analysis. The second workshop, which we plan to hold in the Bay Area of California in Spring 2002, will focus on metrics for molecular conformation that are suitable for statistical linkage of function to form. External funding will be sought to support the travel expenses of visitors and the local-arrangement expenses. Later, we plan to organize open workshops and special sessions attached to conferences such as Intelligent Systems in Molecular Biology (ISMB), International Conference on Computational Molecular Biology (RECOMB), and ACM Symposium on Computational Geometry (SoCG).

The focus of the transfer of knowledge from this project is through the people and the software that the project produces. The Alpha Shape software is already widely used and we will build on the popularity of that product. The presence of biotech industries in the Research Triangle and the Bay areas provides excellent opportunities for collaborating with and transferring technology to industry.

The goals of this project fit well with new institutes in the planning at Duke and at Stanford, both aimed at bringing together multi-disciplinary research and innovative teaching in their areas. As already mentioned, Stanford University is implementing a new program in the Biosciences (Bio-X) that will overlap the Schools of Engineering, Medicine, and Humanities and Sciences. Forty-five faculty members will be hosted in a new building, the Clark Center, which will include new high-performance computing and virtual reality facilities available for research in biocomputation. Three members of our team are on the Bio-X Advisory and Education Committees. The goals of the Center for Genomics at Duke are similar. At UNC we plan to continue the tradition of serving outreach and educational goals through regularly scheduled demonstration days in the visualization laboratories. We offer both 'Introduction to Virtual Reality' demos, targeted at middle and high school students and non-technical visitors, and more technical demos for professional colleagues, industry and press guests.

3 Conclusions

We have presented a set of novel tools based on computational geometry and topology that we feel can make a large impact in molecular biology and its applications to medicine and engineering. To build these tools we will have to address computer science challenges and make fundamental advances in multi-scale and flexible shape modeling, energy-based sampling of conformations and pathways, and the organization and search of large shape databases. We expect to contribute significantly to biocomputation curricula and to build solid research and educational links between computer science and biology, and the corresponding economic forces of information and biotechnology that will shape our future.

PRIOR NSF SUPPORT

P. K. Agarwal. The work supported by the following three grants focused on developing simple and efficient algorithm for a variety of geometric problems that arise in computer graphics, robotics, geographic information systems, and spatial databases. Approximation and randomized techniques were extensively used for developing simpler algorithms than existed before. Techniques were developed for handling moving objects, for dealing with massive datasets, and for analyzing the combinatorial structure underlying some of the problems.

'Simple and efficient geometric algorithms and their applications', 1998–2001. 'Geographic information systems on high-speed clusters: a vertically integrated approach', 1998–2002. 'Efficient geometric algorithms and their applications' (with J. Chase and others), 1993–1998.

H. Edelsbrunner. The earliest of the four listed grants supported the creation of the Alpha Shapes software. The other three supported the study of algorithms, geometry, and topology questions that arise in computational biology on the molecular level. This research lead to the new geometric concepts (pockets, molecular skin, geometry-based deformation) and fast algorithms for construction, analysis and visualization.

'Deformable smooth surface and volume design', 1997–99. 'Computational geometry and biomolecular docking' (with K. Schulten), 1997–99. 'Computation of shape and topology in proteins' (with S. Subramaniam), 1994–96. 'Shapes for modeling and visualization' (with P. Fu), 1992–94.

H. Hellinga. His research is supported by grants of the NIH, Office of Naval Research, Juvenile Diabetes Foundation, as well as industrial sources. He has not received any NSF support in the last five years.

L. J. Guibas. Work under the following four grants addressed fundamental questions in geometric computing, including the computation of arrangements and parts thereof, randomized algorithms, visibility problems, and robustness issues. Progress was made on applications to matching problems in computer vision, to path tracing and partition trees in computer graphics, and to a variety of motion and assembly planning tasks in robotics. The kinetic data structures framework for mobile data described in the proposal was developed with support from these grants.

'Geometric algorithms and applications', 1993–1996. 'Computational geometry in the physical world', 1996–1999. 'Visibility-based motion planning' (with J.-C. Latombe), 1997–2000. 'Geometric structures for mobile data', 1999–2002.

J.-C. Latombe. Under the following four grants, we designed, implemented, and analyzed efficient algorithms for motion planning in the presence of a variety of constraints and for modeling deformable objects. The first type of algorithms are based on computing an arrangement of regions in a certain geometric space, such that a property of interest remains fixed over each region. The second type of algorithms attempt to capture the connectivity of a complex high-dimensional space by random sampling techniques.

'Automated reasoning about assembly processes and products' (with L. Guibas), 1993–1997. 'The intelligent observer: a tool for collaborative experimental research among geographically dispersed groups' (with C. Tomasi), 1995–1997. 'Visibility-based motion planning' (with L. Guibas), 1997–2000. 'Modeling human-body soft tissues for surgical applications', 1999–2002.

M. Levitt. His research is supported by grants from NIH and DOE. He has not received any NSF support in the last five years.

F. P. Brooks, Jr. Developed a nanoWorkbench that allows users to see and feel a surface as if it is floating in front of them, magnified one million times. Developed two promising and complementary techniques for the visual display of multiple scalar fields on a surface. Designed a scalable nanomanipulator system that runs on everything from PC-based graphics cards to the PixelFlow graphics supercomputer, which is being used to study the mechanical, electrical and interfacial properties of nanotubes.

'Application of high-performance graphics supercomputers and communication to provide improved interfaces to scanning probe microscopes', 1995–2000.

C. W. Carter. The listed grant supported applications of Gerard Bricogne's conditional probability formulation for direct methods. We focussed on using the methods to minimize model bias in difficult structures, on preparing the software for use by other groups, and on developing tutorials.

'Bayesian methods for macromolecular phase determination', 1993–1995.

J. S. Snoeyink. He returned to the US from Canada in August 1999, and has not received prior NSF support. In Canada, the National Science and Engineering Research Council (NSERC), three National Centres of Excellence (MITACS, GEOIDE, and IRIS), the BC Advanced Systems Institute, and local industry supported his research into computational geometry and its application to Geographic Information Systems (GIS).

S. Bililign. Both grants listed supported *ab initio* studies of metal-rare gas and metal-hydrogen interaction potentials.

'Quantum calculations', 1996–1999. 'Effects of electronic orbital alignment in laser induced metal-H2 and metal-CH4 reactions'(CAREER), 1998–2002.

References

[1] P. K. AGARWAL AND J. ERICKSON. Geometric range searching and its relatives. In *Advances in Discrete and Computational Geometry*, (B. Chazelle, J. E. Goodman and R. Pollack, eds.), AMS Press, 1999, 1–56.

[2] P. K. AGARWAL AND C. M. PROCOPIUC. Approximation algorithms for projective clustering. *In* "Proc. 11th ACM-SIAM Sympos. Discrete Alg., 2000", 538–547.

[3] F. H. ALLEN ET AL. The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Crystallogr.* **B35** (1979), 2331.

[4] C. C. AGGARWAL, C. M. PROCOPIUC, J. L. WOLF, P. S. YU AND J. S. PARK. Fast algorithms for projective clustering. *In* "Proc. ACM-SIGMOD Intl. Conf. Management Data, 1999" 61–72.

[5] H. ALT AND L. J. GUIBAS. Discrete geometric shapes: matching, interpolation, and approximation. In *Handbook of Computational Geometry* (J.-R. Sack and J. Urrutia, eds.), North-Holland, 2000, 121–154.

[6] F. AURENHAMMER. Power diagrams: properties, algorithms and applications. *SIAM J. Comput.* **16** (1987), 78–96.

[7] J. BASCH, L. J. GUIBAS AND J. HERSHBERGER. Data structures for mobile data. *J. Algorithms* **31** (1999), 1–28.

[8] J. BASCH, L. J. GUIBAS AND L. ZHANG. Proximity problems on moving points. *In* "Proc. 13th Ann. Sympos. Comput. Geom., 1997", 344–351.

[9] F. C. BERNSTEIN ET AL. The protein data bank: A computer based archival file for macromolecular structures. *J. Mol. Biol.* **112** (1977), 532–542.

[10] P. J. BESL AND N. D. MCKAY. A method for registration of 3-D shapes. *IEEE Trans. Pattern Analysis Mach. Intell.* **14** (1992), 239–256.

[11] G. BRICOGNE. (1993). Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives. *Acta Crystallogr.* **D49** (1993), 37–60.

[12] F. L. BROOKSTEIN. *Morphometric Tools for Landmark Data: Geometry and Biology.* Cambridge Univ. Press, Cambridge, 1991.

[13] P. B. CALLAHAN. Dealing with higher dimensions: the well-separated pair decomposition and its applications. Ph.D. thesis, Dept. Comput. Sci., Johns Hopkins Univ., Baltimore, Maryland, 1995.

[14] H. CARR, J. SNOEYINK AND U. AXEN. Computing contour trees in all dimensions. *In* "IEEE Sympos. Discrete Algor., 2000", 918–926.

[15] H.-L. CHENG, T. K. DEY AND H. EDELSBRUNNER. The breathing skin triangulation. Manuscript, 2000.

[16] H.-L. CHENG, H. EDELSBRUNNER AND P. FU. Shape space from deformation. *In* "Proc. 6th Pacific Conf. Comput. Graphics Appl., 1998", 104–113.

[17] M. L. CONNOLLY. Analytical molecular surface calculation. *J. Appl. Cryst.* **16** (1983), 548–558.

[18] B. DAHIYAT AND S. L. MAYO. De novo protein design: fully automated sequence selection. *Science* **278** (1997), 82–87.

[19] G. DAS AND G. NARASIMHAN. A fast algorithm for constructing sparse Euclidean spanners. *Int. J. Comp. Geom. Appl.* **7** (1997), 297–315.

[20] M. E. DAVIS. The inducible multipole solvation model: a new model for solvation effects on solute electrostatics. *J. Chem. Phys.* **100** (1994), 5149–5159.

[21] J. DESMET, M. DE MAEYER, B. HAZES AND I. LASTERS. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356** (1992), 539–542.

[22] T. K. DEY, H. EDELSBRUNNER, S. GUHA AND D. V. NEKHAYEV. Topology preserving edge contraction. *Publ. Inst. Math. (Beograd) (N. S.)* **66** (1999), 23–45.

[23] S. DONIACH AND P. EASTMAN. Protein dynamics simulations from nanoseconds to microseconds. *Cur. Opin. Struct. Biol.* **9** (1999), 157–163.

[24] Y. DUAN AND P. A. KOLLMAN. Pathways to a protein-folding intermediate observed in a 1-microsecond simulation in aqueous-solution. *Science* **282** (1998), 740–744.

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

[26] H. EDELSBRUNNER, M. FACELLO AND J. LIANG. On the definition and the construction of pockets in macromolecules. *Discrete Appl. Math.* **88** (1998), 83–102.

[27] H. EDELSBRUNNER, D. LETSCHER AND A. ZOMORODIAN. Topological persistence and simplification. Manuscript, 2000.

[28] H. EDELSBRUNNER AND E. P. MÜCKE. Three-dimensional alpha shapes. *ACM Trans. Graphics* **13** (1994), 43–72.

[29] H. EDELSBRUNNER AND N. R. SHAH. Triangulating topological spaces. *Internat. J. Comput. Geom. Appl.* **7** (1997), 365–378.

[30] D. EISENBERG AND A. MCLACHLAN. Solvation energy in protein folding and binding. *Nature (London)* **319** (1986), 199–203.

[31] C. FALOUTSOS, R. BARBER, M. FLICKER, J. HAFNER, W. NIBLACK AND D. PETKOVIC. Efficient and effective querying by image content. *J. Intell. Inform. Sys.* **3** (1994), 231–262.

[32] C. FALOUTSOS AND K.-I. LIN. FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia databases. *In* "Proc. ACM-SIGMOD Conf. Management Data, 1995", 163–173.

[33] S. FORTIER, A. CHIVERTON, J. GLASGOW, AND L. LEHERTE. (1997). Critical point analysis in protein electron-density map interpretation. *Methods in Enzymology* **277** (1997), 131–157.

[34] F. FRATERNALI AND W. F. VAN GUNSTEREN. An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. *J. Mol. Biol.* **256** (1996), 939–948.

[35] M. GARLAND AND P. S. HECKBERT. Surface simplification using quadratic error metrics. *Computer Graphics*, Proc. SIGGRAPH 1997, 209–216.

[36] M. GARLAND AND P. S. HECKBERT. Simplifying surfaces with color and texture using quadric error metrics. *In* "Proc. IEEE Sympos. Visualization, 1998", 263–269.

[37] L. J. GUIBAS. Kinetic data structures — a state of the art report. *In* "Proc. 3rd Workshop on Algorithmic Foundations of Robotics, 1998", 191–209.

[38] A. GUPTA AND R. JAIN. Visual information retrieval. *Comm. ACM* **40** (1997), 69–79.

[39] J. M. HAILE. *Molecular Dynamics Simulation : Elementary Methods.* John Wiley & Sons, 1997.

[40] W. HAKEN. Theorie der Normalflächen, ein Isotopiekriterium für den Kreisknoten. *Acta Math.* **105** (1961), 245–375.

[41] D. HAMADA, Y. KURODA, T. TANAKA AND Y. GOTO. High helical propensity of the peptide fragments derived from beta lactoglobulin, a predominantly beta-sheet protein. *J. Mol. Biol.* **254** (1995), 737–746.

[42] H. A. HAUPTMAN. Shake and Bake, an algorithm for the automatic solution, ab inito, of crystal structures. *Methods in Enzymology* **277**, in press.

[43] H. W. HELLINGA AND F. M. RICHARDS. Construction of new ligand binding sites in proteins of known structure. *J. Mol. Biol.* **222** (1991), 763–785.

[44] H. HOPPE, T. DEROSE, T. DUCHAMP, J. MCDONALD AND W. STÜTZLE. Mesh optimization. *Computer Graphics*, Proc. SIGGRAPH 1993, 19–26.

[45] D. HSU, J.-C. LATOMBE AND R. MOTWANI. Path planning in expansive configuration spaces. *Int. J. Comput. Geom. Appl.* **9** (1999), 495–512.

[46] L. E. KAVRAKI, P. SVESTKA, J.-C. LATOMBE, AND M. OVERMARS. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robotics Automation* **12** (1996), 566–580.

[47] D. KEIM. Efficient geometry based similarity search of 3D spatial databases. *In* "Proc. ACM-SIGMOD Conf. Management Data, 1999", 395–406.

[48] P. KOEHL AND M. LEVITT. A brighter future for protein structure prediction. *Nat. Struct. Biol.* **6** (1999), 108–111.

[49] P. J. KRAULIS. Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Applied Crystallogr.* **24** (1991), 946–950.

[50] B. LEE AND F. M. RICHARDS. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55** (1971), 379–400.

[51] M. LEVITT AND A. WARSHEL. Computer simulation of protein folding. *Nature* **253** (1975), 694–698.

[52] J. LIANG, H. EDELSBRUNNER AND C. WOODWARD. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Science* **7** (1998), 1884–1897.

[53] V. Y. LUNIN. Use of information on electron density distribution in macromolecules. *Acta Crystallogr.* **A44** (1988), 144–150.

[54] J. MILNOR. *Morse Theory.* Princeton Univ. Press, New Jersey, 1963.

[55] B. MIRTICH. *Impulse-based Dynamic Simulation of Rigid Body Systems*. Ph.D. thesis, Dept. Elec. Engin. Comput. Sci., Univ. California, Berkeley, CA, 1996.

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

[57] A. NATSEV, R. RASTOGI AND K. SHIM. WALRUS: a similarity retrieval algorithm for image databases. *In* "Proc. ACM SIGMOD Intl. Conf. Management of Data, 1999", 395–406.

[58] PRESIDENT'S INFORMATION TECHNOLOGY ADVISORY COMMITTEE. Information technology research: investing in our future. National Coordination Office for Computing, Information, and Communications, 1999.

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

[60] J. S. RICHARDSON, D. C. RICHARDSON, K. A. THOMAS, E. W. SILVERTON AND D. R. DAVIES. Similarity of three-dimensional structure between the immuno-globulin domain and the Cu, Zc superoxide dismutates subunit. *J. Mol. Biol.* **102** (1976), 221–235.

[61] J. M. ROACH, P. RETAILLEAU AND C. W. CARTER, JR. Using anomalous scattering data in phase refinement via Sayre equations. Abstract W0105, American Crystallographic Association Annual Meeting, St. Paul, MN, 2000.

[62] R. ROTHBAUER. On the separation of the unknown parameters of the problem of crystal-structure analysis. *Acta Crystallogr.* **A36** (1980), 27–32.

[63] R. ROTHBAUER. An extremal principle for the phase problem of structures consisting entirely of spherically symmetric non-overlapping parts. *Zeitschrift für Kristallographie* **215** (2000), 157–168.

[64] B. ROUX AND T. SIMONSON. Implicit solvent models. *Biophys. Chem.* **78** (1999), 1–20.

[65] J. RUPPERT AND R. SEIDEL. Approximating the d-dimensional complete Euclidean graph. *In* "Proc. 3rd Canad. Conf. Comput. Geom., 1991", 207–210.

[66] Y. RUBNER, C. TOMASI, AND L. GUIBAS. A metric for distributions with applications to image databases. *In* "Proc. IEEE Int. Conf. Computer Vision, 1998", 59–66.

[67] D. SAYRE. On least-squares refinement of the phases of crystallographic structure factors. *Acta Crystallogr.* **A28** (1972), 210–212.

[68] D. SAYRE. Least-squares phase refinement. II. High-resolution phasing of a small protein. *Acta Crystallogr.* **A30** (1974), 180–184.

[69] M. SCHAEFER AND M. KARPLUS. A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.* **100** (1996), 1578–1599.

[70] G. SZEKELY, A. KELEMEN, CH. BRECHBUEHLER AND G. GERIG. Segmentation of 3D objects from MRI volume data using constrained elastic deformations of flexible Fourier surface models. *Medical Image Analysis* **1** (1996), 19–34.

[71] A. P. SINGH, J. C. LATOMBE AND D. B. BRUTLAG. A motion planning approach to flexible ligand binding. *In* "Proc. 7th Internat. Conf. Intell. Sys. Mol. Biol., 1999", 252–261, AAAI Press, Menlo Park, California.

[72] C. G. SMALL. *The Statistical Theory of Shape.* Springer-Verlag, New York, 1996.

[73] W. C. STILL, A. TEMPCZYK, R. C. HAWLEY AND T. HENDRICKSON. Semi analytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112** (1990), 6127–6129.

[74] J. L. SUSSMAN AND S. H. KIM. Three-dimensional structure of a transfer RNA common in two crystal forms. *Science* **192** (1976), 853–858.

[75] D. TSERNOGLOU AND G. A. PETSKO. Three-dimensional structure of neurotoxin a from venom of the Philippines sea snake. *Proc. Nat. Acad. Sci. USA* **74** (1977), 971–974.

[76] G. TURK AND A. PENTLAND. Face recognition using eigenfaces. *In* "IEEE Proc. Comput. Vision Pattern Recogn., 1991, 586–592.

[77] K. R. VARADARAJAN AND P. K. AGARWAL. Approximation algorithms for bipartite and non-bipartite matching in the plane. *In* "Proc. 10th ACM-SIAM Sympos. Discrete Algorithms, 1999", 805–814.

[78] L. WESSON AND D. EISENBERG. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.* **1** (1992), 227–235.

[79] S. XIANG AND C. W. CARTER, JR. Representing stereochemical information in macromolecular electron density distributions by multi-dimensional histograms. *Acta Crystallogr.* **D52** (1996), 49–56.

[80] K. Y. J. ZHANG AND P. MAIN. Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. *Acta Crystallogr.* **A46** (1990), 41–46.