BIOMOLECULAR MODELING

DANIEL B. DIX

Contents

1. Introduction

Molecular biology is currently one of the most exciting and challenging application frontiers for practitioners of applied mathematics. Mathematical models have for a long time and successfully been developed for complex biological systems at the cellular, organismal, and population levels [34]. And recently there has been the realization that engineering principles governing complex systems of human origin are applicable to various subsystems of the cell, which has resulted in a quest for a *systems biology* [4]. Physicists have long asserted that the underlying laws governing the motions and arrangements of atoms and molecules are known, and yet the well-known difficulty of using these mathematical laws to derive the behavior of complex systems together with the conviction that the emergent "self-organizing principles" will be largely independent of the atomic details has lead many to abandon the "reductionistic" approach, and even to deny the existence of a deductive link between the known laws and the postulated organizing principles [21]. In some cases organizing principles have been deduced from the underlying theory, such as the derivation of hydrodynamics as an asymptotic limit of molecular dynamics [5]. Rigorously deducing asymptotic limits of complex dynamical theories is the work of mathematicians, and though the work is hard we are not yet willing to give up.

Date: January 29, 2003.

We are interested in such limits as are important for the behavior of large biomolecular systems. Since our ability to directly observe this “mesoscopic” realm is quite limited [22], the “reductionistic” approach to understanding, especially using computational tools to manage the increasing complexity, seems to be competitive, even promising.

This proposal is concerned with the effort to construct a mathematical and computational modeling environment which connects all the levels of organization from atomic level structure to biological function. It will by no means provide all the components needed to achieve this worthy goal. Our approach is basically to start at the structural level and work upwards toward the cellular level. We highly value the efforts of those who can contribute insight through experiment, empirical law, and other nonreductionistic approaches [22].

We present three types of proposals. First we discuss mathematical concepts which are useful to model structures and assemblies of biomolecules. At this stage it is extremely challenging to formulate complete well-posed and tractable mathematical models with important and experimentally correct predictions. In many cases the “correct” mathematical concepts still need to be defined so that the main problems can be stated as mathematics problems. Often the resulting mathematics problems appear to be very difficult to solve, but this does not mean that the definitions were not helpful. A “correct” definition is one which matches reality well and yet also possesses mathematical elegance. We will first describe a very natural mathematical theory of biomolecular structure, thereby laying a firm foundation at that level. We will also describe some subsidiary concepts which build on this foundation, and point out some concepts which yet await definition.

Secondly we propose certain mathematical and computational problems which seem to be necessary to solve if our grander aims are to be achieved. In some cases the exact biologically relevant formulation is still not clear, and we propose to collaborate with molecular biologists to resolve the ambiguities.

Thirdly, we propose certain software development projects which are closely related to our mathematical concepts and are guided by our larger aims. Software allows human beings to extend their ability to comprehend complex systems through graphical interfaces, hierarchy, and detail hiding. A relevant example is the software systems used by computer chip designers. We view the software in a supporting role as facilitating exploration of the geometry and motions of biomolecular systems, leading to new mathematical concepts and to a greater understanding of modeling concepts at higher levels of organization.

The proposed research will significantly enhance the mathematical infrastructure of molecular biology, and hence it participates in the broader impacts of that discipline on health and biotechnology.

2. Definitions of Mathematical Concepts

2.1. **Generalized Z-systems.** GZ-systems are combinatorial structures useful for biomolecular geometry description. The theory of n-dimensional Z-systems and all the basic geometric concepts surrounding them is developed in [6]. Earlier related work is [3], [7], [47]. Here we will describe a useful generalization, but in the 3-dimensional case, which will form the foundation of our approach to biomolecular geometry.

Suppose $\mathcal{N}$ is the set of the atom names for all the atoms in a molecular system. We suppose $\mathcal{N}$ has $N \geq 3$ elements. If $k \geq 0$ then let $\binom{\mathcal{N}}{k+1}$ denote the set of all $k+1$ element subsets of $\mathcal{N}$. Each element of $\binom{\mathcal{N}}{k+1}$ will be called an (abstract) *k-simplex.* If $\Gamma \subset \mathcal{P}(\mathcal{N})$ and $\mathcal{C} \subset \mathcal{N}$ then let $\Gamma^k_{\mathcal{C}}$ denote the set of all abstract k-simplices in Γ which are subsets of $\mathcal{C}$; abbreviate $\Gamma^k_{\mathcal{N}}$ by Γ^k and $\cup_{k=0}^{\infty}\Gamma^k_{\mathcal{C}}$ by $\Gamma_{\mathcal{C}}$. $\Gamma \subset \cup_{k=0}^{3}\binom{\mathcal{N}}{k+1}$ is called a *unoriented generalized Z-system (GZ-system) on $\mathcal{N}$* if the following conditions hold.

1. $\Gamma^0 = \binom{\mathcal{N}}{1}$.
2. If $e \in \Gamma^k$, $k \geq 1$, then $|\Gamma^{k-1}_e| = 2$. (We may think of $e \in \Gamma^k$ as an edge incident on the two vertices $v_1, v_2 \in \Gamma^{k-1}$, where $\Gamma^{k-1}_e = \{v_1, v_2\}$.)
3. (Γ^0, Γ^1) is an acyclic graph.
4. If $\mathcal{C} \subset \mathcal{N}$ is such that $(\Gamma^0_{\mathcal{C}}, \Gamma^1_{\mathcal{C}})$ is a connected component of (Γ^0, Γ^1) then $(\Gamma^1_{\mathcal{C}}, \Gamma^2_{\mathcal{C}})$ and $(\Gamma^2_{\mathcal{C}}, \Gamma^3_{\mathcal{C}})$ are (possibly empty) trees. ($\mathcal{C}$ is called a *component.*)
5. $\Gamma^2 \neq \emptyset$.
6. If $v_1, v_2 \in \Gamma^2$ and $v_1 \cup v_2 \in \Gamma^3$ then $v_1 \cap v_2 \in \Gamma^1$.

If (Γ^0, Γ^1) is also connected then we say Γ is an *unoriented Z-system.* If $w = \{A_0, A_1, A_2, A_3\} \in \Gamma^3$ then define *oriented* 3-simplices to be equivalence classes

$$[A_0, A_1, A_2, A_3] = \{(A_{\pi(0)}, \ldots, A_{\pi(3)}) \mid \pi \text{ is an even permutation of } \{0,1,2,3\}\}.$$

Let υ map oriented 3-simplices to their underlying 3-simplices, i.e. $\upsilon(w^*) = w$. Let Γ^3_* be a collection of oriented 3-simplices such that $\upsilon\colon \Gamma^3_* \to \Gamma^3$ is a bijection. $(\Gamma^1, \Gamma^2, \Gamma^3_*)$ is called simply a *(G)Z-system.* If $\mathcal{C}$ is a component of Γ with $|\mathcal{C}| \geq 3$ then $\Gamma_{\mathcal{C}}$ is a Z-system; $\Gamma_{\mathcal{C}}$ is called *monatomic* (resp. *diatomic*) if $|\mathcal{C}| = 1$ (resp. $|\mathcal{C}| = 2$). Elements of $\Gamma^0, \Gamma^1, \Gamma^2, \Gamma^3_*$, are called *atoms, bonds, angles,* and *wedges* respectively. Components typically describe covalently bound molecules (or monatomic ions); a system will typically consist of many components. A pictorial view of a Z-system for the molecule methanol is in figure 1. If we keep in mind that edges $e \in \Gamma^k$ in the graph (Γ^{k-1}, Γ^k) become vertices of the graph (Γ^k, Γ^{k+1}) then we have another view of the same Z-system in figure 2. The orientation of a 3-simplex can be chosen canonically when the bonds involved form a chain (*dihedral* case), but can be indicated by an arrow when the bonds share a common atom (*improper* case).

If Γ is a Z-system we can specify a particular geometry of the molecule described by Γ by defining three mappings: $L\colon \Gamma^1 \to (0, \infty)$, $C\colon \Gamma^2 \to (-1, 1)$, and $Z\colon \Gamma^3_* \to S^1$. If $\mathbf{R} \in (\mathbb{R}^3)^{\mathcal{N}}$ is a molecular configuration (giving the position of each of the atoms) and $b \in \Gamma^1$ then $L_b(\mathbf{R})$ is the distance between the two atoms of the bond (*bond length*) in the configuration $\mathbf{R}$. If $a \in \Gamma^2$ is the angle incident on the two bonds $b_1, b_2 \in \Gamma^1$, $b_j = \{A_j, A\}$, $j = 1, 2$, then

$$C_a(\mathbf{R}) = \frac{\mathbf{R}_{A_1} - \mathbf{R}_A}{\|\mathbf{R}_{A_1} - \mathbf{R}_A\|} \cdot \frac{\mathbf{R}_{A_2} - \mathbf{R}_A}{\|\mathbf{R}_{A_2} - \mathbf{R}_A\|}$$

is the cosine of the geometrical angle (*bond angle*) between these two bonds in the configuration $\mathbf{R}$. If $w^* = [A_0, A_1, A_2, A_3] \in \Gamma^3_*$, where $\{A_0, A_1, A_2\}, \{A_0, A_1, A_3\} \in$

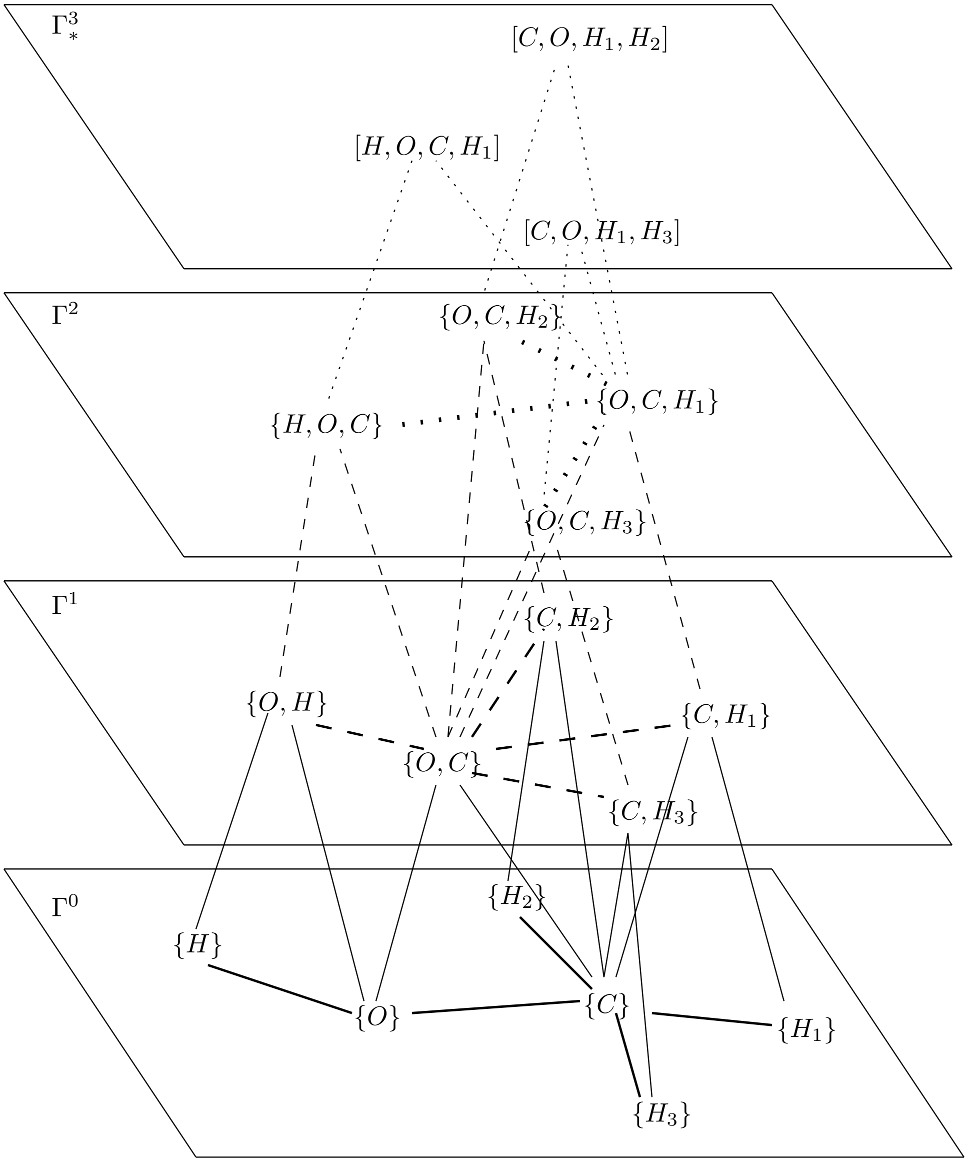

FIGURE 1. A 3-dimensional Z-system Γ for Methanol. The set $\mathcal{N} = \{C, H_1, H_2, H_3, O, H\}$ contains the atom names. The tree for (Γ^{k-1}, Γ^k), $k = 1, 2, 3$, is indicated on the part labeled Γ^{k-1}, where the edges are indicated by heavier lines of various styles. Above each such line is the element of Γ^k which is the edge, and it is connected to its two vertices by lighter lines of the same style.

Γ^2, then define

$$Z_{w^*}(\mathbf{R}) = \mathbf{v}\cdot\mathbf{w} + i\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}, \qquad \text{where} \qquad \mathbf{u} = \frac{\mathbf{R}_{A_1} - \mathbf{R}_{A_0}}{\|\mathbf{R}_{A_1} - \mathbf{R}_{A_0}\|},$$

$$\mathbf{v} = \frac{(1-\mathbf{u}\mathbf{u}^T)(\mathbf{R}_{A_2} - \mathbf{R}_{A_0})}{\|(1-\mathbf{u}\mathbf{u}^T)(\mathbf{R}_{A_2} - \mathbf{R}_{A_0})\|}, \qquad \mathbf{w} = \frac{(1-\mathbf{u}\mathbf{u}^T)(\mathbf{R}_{A_3} - \mathbf{R}_{A_0})}{\|(1-\mathbf{u}\mathbf{u}^T)(\mathbf{R}_{A_3} - \mathbf{R}_{A_0})\|}.$$

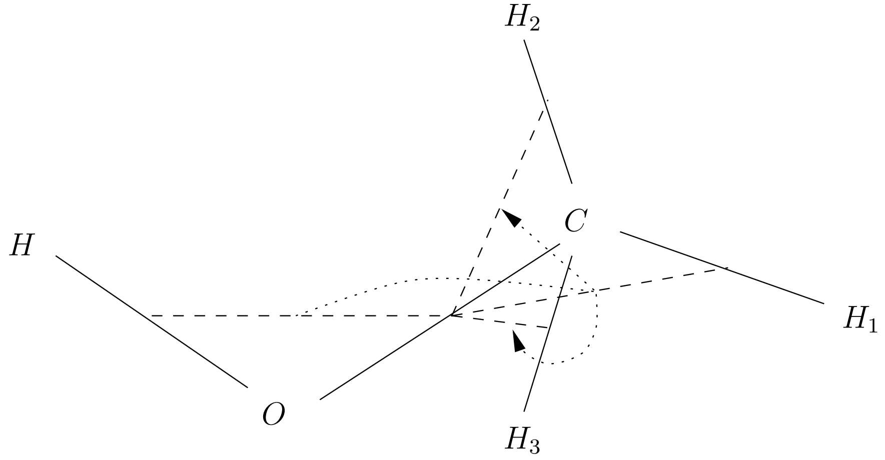

FIGURE 2. The iterated line graph form (used by IMIMOL) of the same Z-system Γ for Methanol as in Figure 1. Bonds (1-simplices) are indicated by solid lines. Angles (2-simplices) are indicated by dashed lines from one vertex bond to the other. Wedges (3-simplices) are indicated by curved or straight dotted lines from one vertex angle to the other.

There are four permutations $(A_{\pi(0)}, \dots, A_{\pi(3)})$ of $w = \{A_0, A_1, A_2, A_3\} \in \Gamma^3$ such that $\{A_{\pi(0)}, A_{\pi(1)}, A_{\pi(2)}\}, \{A_{\pi(0)}, A_{\pi(1)}, A_{\pi(3)}\} \in \Gamma^2$, namely

$$\begin{array}{ll} (A_0, A_1, A_2, A_3) & (A_1, A_0, A_2, A_3) \\ (A_1, A_0, A_3, A_2) & (A_0, A_1, A_3, A_2). \end{array}$$

The two in the first column are in the same equivalence class and determine the same number $Z_{w^*}(\mathbf{R})$ (since $\mathbf{u'} = -\mathbf{u}$, $\mathbf{v'} = \mathbf{w}$, $\mathbf{w'} = \mathbf{v}$) whereas the two in the second column yield the complex conjugate of $Z_{w^*}(\mathbf{R})$. Thus $Z_{w^*}(\mathbf{R}) = e^{i\varphi}$ determines a signed angle (*wedge angle*) between the half-plane containing $\{\mathbf{R}_{A_0}, \mathbf{R}_{A_1}, \mathbf{R}_{A_2}\}$ and the half-plane containing $\{\mathbf{R}_{A_0}, \mathbf{R}_{A_1}, \mathbf{R}_{A_3}\}$, the common boundary of the half-planes being the line containing $\{\mathbf{R}_{A_0}, \mathbf{R}_{A_1}\}$. The sign of φ and the orientation of w^* are related by the right-hand rule, as the above formulae show. These definitions of $L_b(\mathbf{R}), C_a(\mathbf{R})$, and $Z_{w^*}(\mathbf{R})$ make sense and fall in the ranges $(0, \infty), (-1, 1)$, and S^1, respectively provided the configuration $\mathbf{R}$ is such that for every $\{A_0, A_1, A_2\} \in \Gamma^2$ the points $\mathbf{R}_{A_0}, \mathbf{R}_{A_1}, \mathbf{R}_{A_2}$ are not collinear in $\mathbb{R}^3$. This set of molecular configurations is invariant under rigid motions. An orbit in this set of configurations with respect to the action of the group of rigid motions will be called a *conformation* of Γ; $L_b(\mathbf{R}), C_a(\mathbf{R})$, and $Z_{w^*}(\mathbf{R})$ are constant on orbits. These definitions establish a diffeomorphism between the manifold of conformations of Γ and the manifold (of *labelled Z-systems*) $(0, \infty)^{\Gamma^1} \times (-1, 1)^{\Gamma^2} \times (S^1)^{\Gamma^3_*}$. This diffeomorphism is called the *polyspherical coordinate chart* associated to the Z-system $(\Gamma^1, \Gamma^2, \Gamma^3_*)$.

In order to specify the conformation of a system of molecules described by a GZ-system Γ we make the following definitions. A *site* is a set $r = \{s_0, \dots, s_k\}$ of elements of Γ, linearly ordered by inclusion $s_0 \subset \cdots \subset s_k$, which is maximal in

$\Gamma \setminus \Gamma^3$. Thus $s_0 \in \Gamma^0, \dots, s_k \in \Gamma^k$; the value of k depends on the site; we call r a k-site. Each site r is associated with a component $\mathcal{C}$ of Γ where $r \subset \Gamma_{\mathcal{C}}$. Monatomics have 0-sites, diatomics have 1-sites, and Z-systems have 2-sites. Let $\mathcal{D}_C(\Gamma)$ denote the set of all configurations $\mathbf{R} \in (\mathbb{R}^3)^{\mathcal{N}}$ such that the geometric simplex $\mathbf{R}_e \subset \mathbb{R}^3$ corresponding to every maximal element e of $\Gamma \setminus \Gamma^3$ is geometrically independent. If $\mathbf{R} \in \mathcal{D}_C(\Gamma)$ and r is a k-site of Γ then define $E_r(\mathbf{R}) = (\mathbf{e}_0, \dots, \mathbf{e}_{2^k-1})$ to be a 3×2^k matrix, where $\mathbf{e}_0 = \mathbf{R}_{A_0}$, $\mathbf{e}_1 = \frac{\mathbf{R}_{A_1} - \mathbf{R}_{A_0}}{\|\mathbf{R}_{A_1} - \mathbf{R}_{A_0}\|}$ (when $k \geq 1$), and $\mathbf{e}_2 = \frac{(1-\mathbf{e}_1\mathbf{e}_1^T)(\mathbf{R}_{A_2} - \mathbf{R}_{A_0})}{\|(1-\mathbf{e}_1\mathbf{e}_1^T)(\mathbf{R}_{A_2} - \mathbf{R}_{A_0})\|}$ and $\mathbf{e}_3 = \mathbf{e}_1 \times \mathbf{e}_2$ (when $k = 2$). $E_r(\mathbf{R})$ is called the *k-pose* at r conformed to the configuration $\mathbf{R}$. If r is a 2-site and r' is a k-site of Γ, then there exists a unique 4×2^k matrix $\mathcal{A}_{(r,r')}(\mathbf{R})$, whose first row is $(1, 0, \dots, 0)$, such that $E_{r'}(\mathbf{R}) = E_r(\mathbf{R})\mathcal{A}_{(r,r')}(\mathbf{R})$. We have that

$$\mathcal{A}_{(r,r')}(\mathbf{R}) \overset{k=2}{=} \begin{pmatrix} 1 & \theta^T \\ \mathbf{b} & A \end{pmatrix}, \quad \mathcal{A}_{(r,r')}(\mathbf{R}) \overset{k=1}{=} \begin{pmatrix} 1 & 0 \\ \mathbf{b} & \mathbf{u} \end{pmatrix}, \quad \mathcal{A}_{(r,r')}(\mathbf{R}) \overset{k=0}{=} \begin{pmatrix} 1 \\ \mathbf{b} \end{pmatrix}$$

where $\mathbf{b} \in \mathbb{R}^3$, $A \in \mathrm{SO}(3)$, $\mathbf{u} \in S^2 \subset \mathbb{R}^3$. If sites r and r' are associated with different components of Γ, then the matrix $\mathcal{A}_{(r,r')}(\mathbf{R})$ specifies the relative position (and orientation) of the two components in the configuration $\mathbf{R}$, and is invariant with respect to rigid motions on $\mathbf{R}$.

To fix the conformation of the entire system we introduce another graph $\mathcal{S}$ whose vertices are the components of Γ and whose oriented edges are ordered pairs (r, r'), where r is a 2-site in one component $\mathcal{C}$ and r' is a k-site in another component $\mathcal{C}'$. We require that the graph $\mathcal{S}$ be a tree. The conformation of the system is fixed when one is given mappings $L \colon \Gamma^1 \to (0, \infty)$, $C \colon \Gamma^2 \to (-1, 1)$, $Z \colon \Gamma^3_* \to S^1$, and an assignment $\mathcal{A}$ of a matrix (satisfying the above restrictions) to each pair in $\mathcal{S}$.

Z-systems and their associated polyspherical coordinates are very similar to *Z-matrix*, or *valence* internal coordinates used by chemists for many years, although without the formal structure and rigor exhibited here. However Z-matrices [12], [30] can be understood as being Z-systems with additional structure added [6]. This additional structure imposes an ordering on the atoms and also defines a preferred Cartesian coordinate system fixed to the molecule. Also Z-matrices are well-suited to serving as a file format for storing molecular conformations in a computer. But Z-matrices for two molecules cannot easily be combined to form a Z-matrix for the result of a chemical combination of the two molecules. This can now be understood as resulting from the imposition of the extra structure, because Z-systems *can* be cleanly *glued* to yield the Z-system of the chemical product. This gluing operation on Z-systems is very much like the manipulations one performs on plastic models of molecules. Also it is possible to *tether* any component to a Z-system, forming a single component out of the two. The details are straightforward, and can be found in [6].

2.2. **Hierarchical Structure.** Although it would be theoretically possible to study a biomolecular system in the form that is present in one specific species, it would not be advisable because so much information can be obtained more easily by comparison of the different forms which are present in different species. One of the main differences in form of a biomolecular system between species is in the monomer sequences of the biopolymers in the system. Hence we need to add more structure to our system description formalism so that such differences can be naturally discussed. We have been assuming that distinct atoms have distinct names, and the simplest way to accomplish this is to name the atoms hierarchically, much like the full path name of a file in a file/directory hierarchy. Hierarchy seems to be a universal method of dealing with complexity, and it applies well enough to the complexity of biomolecular systems. Hierarchy implies the existence of a rooted tree $\mathcal{T}$, where we assume the root vertex is not a leaf. The set of leaf vertices of $\mathcal{T}$ will be in one-to-one correspondence with the set $\mathcal{N}$ of all the atom names in the system. Vertices of $\mathcal{T}$ will be called *nodes*. Each non-root node has a *parent node* (the next node on the path connecting it to the root node) and a set of *children* nodes; a node is the parent node of each of its children. Each node has a label; the labels of each of the children of a given node must be distinct from each other. The *path name* of a node is given by concatenating the node labels (separated by dots) along the unique path connecting that node to the root node. Each node X of $\mathcal{T}$ can be assigned the set $\mathcal{C}_X \subset \mathcal{N}$ of all leaf nodes which are descendants of X. If X is the root node then $\mathcal{C}_X = \mathcal{N}$. For each component $\mathcal{C}$ there will be a node X such that $\mathcal{C}_X = \mathcal{C}$. If $\mathcal{C}$ describes a biopolymer then it is natural that $\mathcal{T}$ should contain nodes corresponding to each of the monomers comprising $\mathcal{C}$. For linear polymers like proteins one might have the children of $\mathcal{C}$ labeled with the numbers $1, 2, \ldots, M$, where M is the number of amino acids in the protein. Then under the node k one might have a single child labeled with the name of one of the 20 types of amino acids. Thus the hierarchy contains species specific information in a form which is also useful when addressing higher levels of organization (such as complexes of proteins, functional units, organelles, etc.).

Biomolecular systems usually consist of multiple biopolymers together with sufficient solvent (water) molecules to fill the spatial domain to the appropriate density or pressure. Thus $\mathcal{T}$ will contain a node, labeled “solvent”, which will have many children, each child representing a water molecule. In the case of systems containing nucleic acids we will also have a node labeled “counterions”, whose many children will be monotomic positive ions (Mg^{2+}) to neutralize the strong negative charge of the phosphate groups along the backbone. If the spatial domain is not a 3 dimensional torus $\mathcal{T}$ will contain a node, labeled “cage”, which will usually consist of a rigid cage of partially charged centers statically simulating a surrounding body of liquid water. It could have a “hydrophobic band” if one wanted to include part of a lipid bilayer in the system. This cage will be impermeable to water molecules or monatomic ions, so it will confine the other contents of the system to the interior of the cage. The most interesting nodes will be multi-protein complexes and assemblies, which will vary from example to example.

2.3. **Dynamical and Energetic Aspects.** Let $\mathbf{q} = (L, C, Z, \mathcal{A})$ be a conformation of a GZ-system. There is a real-valued function $V_{qm}(\mathbf{q})$ which gives the quantum mechanical potential energy of the conformation $\mathbf{q}$. It is not difficult to define precisely but to save space we will not do so. Using V_{qm} yields a very accurate physical model, but for large systems V_{qm} is very difficult to compute accurately. Alternatively one can approximate $V_{qm}(\mathbf{q})$ by $V_{mm}(\mathbf{q})$, a molecular mechanics expression which is much easier to compute but makes covalent chemistry impossible and may not treat electronic polarization effects accurately. Let $V(\mathbf{q})$ denote a potential energy function which we will assume is accurate enough for our purposes.

If the atom names in $\mathcal{N}$ are numbered from 1 to N then let the configuration $\mathbf{R}$ be represented by a $3 \times N$ matrix whose jth column vector is the position vector $\mathbf{R}_j$ of the jth atom. Let M be an $N \times N$ diagonal matrix whose jth diagonal entry is the mass m_j of the jth atom. Let $m = \text{tr } M$ and let $\bar{\mathbf{R}} = \frac{1}{m}\sum_{j=1}^N m_j \mathbf{R}_j$ denote the center of mass. The total kinetic energy is $\mathbb{T} = \frac{1}{2}\text{tr } (\dot{\mathbf{R}} M \dot{\mathbf{R}}^T)$, where $\dot{\mathbf{R}}$ denotes the time derivative of $\mathbf{R}$. Let r denote a distinguished 2-site in Γ and let $E_r(\mathbf{R}) = (\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, where $B = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) \in \text{SO}(3)$ denotes a set of basis vectors "attached" to the molecule (component) containing the site r. Let $\mathbf{r}$ be the $3 \times N$ matrix whose jth column vector $\mathbf{r}_j$ satisfies $\mathbf{R}_j = B\mathbf{r}_j + \bar{\mathbf{R}}$. the 6 degrees of freedom of $(\bar{\mathbf{R}}, B)$ determine the overall position and orientation of the system. $\mathbf{r}$ is a function of the $(3N-6)\times 1$ column vector $\mathbf{q}$ of internal coordinates. Let $\nabla_{\mathbf{q}}\mathbf{r}_j$ denote the $3 \times (3N-6)$ matrix whose kth column vector is $\frac{\partial \mathbf{r}_j}{\partial q_k}$. Define $\mathbf{i} = \text{tr } (\mathbf{r}M\mathbf{r}^T)\mathbf{1} - \mathbf{r}M\mathbf{r}^T$ to be the 3×3 *moment of inertia matrix*. For $\mathbf{v} \in \mathbb{R}^3$ let $[\mathbf{v}\times]$ denote the 3×3 antisymmetric matrix such that $[\mathbf{v}\times]\mathbf{w} = \mathbf{v} \times \mathbf{w}$ for all $\mathbf{w} \in \mathbb{R}^3$. Let $\mathbf{A}$ denote the $3 \times (3N-6)$ matrix, called the *mechanical connection 1-form*, $\mathbf{A} = \mathbf{i}^{-1}\sum_{j=1}^N m_j[\mathbf{r}_j\times]\nabla_{\mathbf{q}}\mathbf{r}_j$ [24]. Define $\tau(\mathbf{q})$ to be the $(3N-6)\times(3N-6)$ symmetric matrix $\sum_{j=1}^N m_j(\nabla_{\mathbf{q}}\mathbf{r}_j)^T\nabla_{\mathbf{q}}\mathbf{r}_j - \mathbf{A}^T\mathbf{i}\mathbf{A}$, which will be called the *Riemannian metric on conformation space*. It is possible to restrict attention to motions where $\sum_{j=1}^N m_j\dot{\mathbf{R}}_j = \mathbf{0}$ and $\sum_{j=1}^N \mathbf{R}_j \times m_j\dot{\mathbf{R}}_j = \mathbf{0}$ and for which initially we have $(\bar{\mathbf{R}}, B) = (\mathbf{0}, \mathbf{1})$. (We cannot impose vanishing angular momentum in the case of periodic spatial boundary conditions, but there are modifications of this discussion to cover that case.) It follows that $\bar{\mathbf{R}} = \mathbf{0}$ for all time and the evolution of $(\mathbf{q}, \dot{\mathbf{q}})$ is uncoupled from the evolution of B. There is no kinetic energy of overall translational or rotational motion, and we have $\mathbb{T} = \frac{1}{2}\dot{\mathbf{q}}^T\tau(\mathbf{q})\dot{\mathbf{q}}$. If $B(\mathbf{v}) = \mathbf{1}\cos(\theta) + \mathbf{v}\mathbf{v}^T\frac{1-\cos(\theta)}{\theta^2} + [\mathbf{v}\times]\frac{\sin(\theta)}{\theta}$, where $\mathbf{v} \in \mathbb{R}^3$, $\theta = \|\mathbf{v}\|$, then one can show that $\frac{d}{dt}B(\mathbf{v}) = B(\mathbf{v})[\{S(\mathbf{v})\dot{\mathbf{v}}\}\times]$, where $S(\mathbf{v}) = \mathbf{1}\frac{\sin(\theta)}{\theta} + \mathbf{v}\mathbf{v}^T\frac{1}{\theta^2}(1 - \frac{\sin(\theta)}{\theta}) + [\mathbf{v}\times]\frac{1-\cos(\theta)}{\theta^2}$ and $S(\mathbf{v})^{-1} = \mathbf{1}\frac{\theta/2}{\tan(\theta/2)} + \mathbf{v}\mathbf{v}^T\frac{1}{\theta^2}(1 - \frac{\theta/2}{\tan(\theta/2)}) + [\mathbf{v}\times]\frac{1}{2}$. Thus after the evolution of $(\mathbf{q}, \dot{\mathbf{q}})$ is determined $B(\mathbf{v})$ may be found by solving the equation $\dot{\mathbf{v}} = S(\mathbf{v})^{-1}\mathbf{A}(\mathbf{q})\dot{\mathbf{q}}$.

The set of all conformations $\mathbf{q}$ is a smooth Riemannian manifold (Q, τ), and the phase space for the dynamics is the cotangent space T^*Q. The generalized momentum conjugate to $\mathbf{q}$ is $\mathbf{p} = \tau(\mathbf{q})\dot{\mathbf{q}}$. The metric $\tau(q)^{-1}$ on the vector bundle $T^*Q \to Q$ is naturally associated to the metric $\tau(q)$ on $TQ \to Q$. Adding the kinetic and potential energies $H(\mathbf{q}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^T\tau(\mathbf{q})^{-1}\mathbf{p} + V(\mathbf{q})$ gives the Hamiltonian $H\colon T^*Q \to \mathbb{R}$, which determines the dynamics on T^*Q. We will restrict attention to dynamics on a level set $X = \{(\mathbf{q}, \mathbf{p}) \in T^*Q \mid H(\mathbf{q}, \mathbf{p}) = E\}$, where E is chosen so that X is a compact smooth manifold. T^*Q is equipped with the standard symplectic volume form which can be "divided by" the 1-form dH to yield a volume form on X, which after normalization makes (X, μ) into a probability space. μ is the *microcanonical equilibrium state* of the system (assuming the measure μ is ergodic with respect to the flow). For reasons of efficiency most molecular dynamics programs (such as CHARMM, or NAMD [35]) do not compute trajectories in the $\mathbf{q}$ variables (see however [28]). Here we are defining this dynamical system for theoretical purposes; the system clearly would not be ergodic in Cartesian coordinates.

Most questions in molecular biology, such as "Is the protein folded?" or "Are subunits A and B stably associated?" or "Is the ligand bound in the active site?", concern the conformations of specific macromolecules and small molecules of interest, but not the conformations of the solvent, ions, or other components not of interest. Let y denote a typical list of only the conformational variables of interest, and let Y denote the space of all such y and $\rho\colon X \to Y$ the projection mapping. Let dy denote Lebesgue measure in the conformational variables y. If the projection $\rho_*\mu$ of the probability μ on X down to Y is absolutely continuous with respect to dy and $d(\rho_*\mu) = e^{-F(y)}\, dy$ then the function $F\colon Y \to \mathbb{R}$ could be called the *conformational free energy* associated to ρ and E (it also depends implicitly on the volume of the spatial domain and the composition of the system). So if y_0 represents a conformation of biochemical interest, like a folded protein or an assembled complex of proteins, then at equilibrium we expect the system phase point $x = (\mathbf{q}, \mathbf{p})$ to be such that $\rho(x)$ is near y_0 if the thermodynamic variables (E, volume of spatial domain, number of water molecules, number of ions, etc.) are such that $F(y)$ is minimized when $y = y_0$. F defines the "energy landscape" of the system, and many biological mysteries are hidden in this mathematical object.

The above dynamical formalism is inadequate to treat nonequilibrium processes such as the conversion of glucose to ATP and pyruvate via glycolysis. The challenge which is outstanding is to formulate an atomistic model which allows the situation of constant flux through this system to be studied as a time independent state.

2.4. **Ramachandran Spaces.**

GZ-systems provide standard internal coordinates for the description of the geometry of the system. But the biologically interesting geometries form a small subset of the set of all possible geometries. One of the main virtues of GZ-system internal coordinates is that they simplify the description of this subset (more or less) as much as possible. Since approximations are inevitable, it is fortunate that certain reasonable ones are geometrically natural and fairly easy to describe. All bond lengths, all bond angles (except in furanose rings in nucleic acids), all improper wedge angles, and all dihedral wedge angles spanning bonds which are part of rigid five or six membered covalently bound rings we may assume to hold fixed constant values in all low energy conformations. All biologically relevant motion (excluding covalent chemistry processes) takes place in the remaining dihedral wedge angles and in the coordinates relating different components. However, not all combinations of values of these active coordinates make chemical sense. One source of restrictions is the presence of larger (more than six chemical bonds) covalently bound rings, such as in proteins with disulfide bonds. But the most important source of restrictions is the fact that any two atoms A and A' in Γ^0 such that the path in (Γ^0, Γ^1) from A to A' is four or more edges long cannot be allowed to be too close to each other in any reasonable conformation. This restriction is called *steric exclusion.* We will call the set of the active coordinates satisfying the various ring constraints and all steric exclusions the *Ramachandran space* (or $\mathcal{R}$-space) $\mathcal{R}(\Gamma) \subset Q$ of the system. We can interpret the $\mathcal{R}$-space as a subset of Q such that as $\mathbf{q} \in Q$ moves away from $\mathcal{R}(\Gamma)$ the potential energy $V(\mathbf{q})$ increases drastically. $\mathcal{R}(\Gamma)$ is a rough approximation of the image of X under the projection mapping $T^*Q \to Q$. $\mathcal{R}$-spaces have extremely intricate structure which is difficult to analyze. But they are a natural mathematical home for many questions in molecular biology.

3. Mathematical and Computational Problems

3.1. **Computing Points in Ramachandran Spaces.**

A common task is to compute points in an $\mathcal{R}$-space which satisfy certain constraints. For example, consider the $\mathcal{R}$-set for a segment of duplex DNA, described by a GZ-system Γ. It is desirable to compute points $\mathbf{q} \in \mathcal{R}(\Gamma)$ that describe regular double helix structures with standard base pair geometry [36], [23], standard helical parameter (rise, twist, inclination, and propeller twist) values for B-form DNA [23], ring parameter (Cremer-Pople puckering amplitude and phase, Marzec-Day elongation amplitude and phase) values for a $2'$-*endo* furanose ring [27], and standard bond lengths and angles [9]. The primary remaining degrees of flexibility are described by the χ torsion angle between the furanose ring and the base, and the torsion angles $\alpha, \beta, \gamma, \epsilon, \zeta$ [26] along the backbone (δ is fixed by the ring parameters and bond angles). If we require these torsion angles to be the same for all the nucleotides (insuring regularity of the helix) then none of these are independently variable, i.e. they are constrained by something analogous to the presence of a covalently bound ring [10]; we will give an abstract formulation below. Furthermore we are only interested in the values $(\chi, \alpha, \beta, \gamma, \epsilon, \zeta)$ which also satisfy all the steric constraints. These constraints are essentially geometric in nature.

The usual way this problem is addressed by biochemists and crystallographers is to perform an energy minimization where certain coordinates have constrained values and others are variable subject to an "energy" function which penalizes deviations from ideal bond lengths and angles. The starting conformation for the energy minimization is arbitrary. Unfortunately this minimization procedure may result in a local minimum where the bond lengths and angles do not assume their ideal values. Multiple attempts at minimization are needed to find the global minima and to feel confident that all the important minima have been found. An alternate approach, which we prefer, sets up and solves the constraint equations exactly, finding all the solutions, and then selects those which also satisfy the steric constraints. We propose to utilize GZ-systems to provide a systematic way to do this, so that new software need not be written for every individual case.

We have applied such a systematic procedure to the above problem of computing regular B-form DNA conformations by means of the following "bridging" algorithm. Other "bridging" algorithms were applied to this same problem in [44], [31], [48]. Suppose the 2-sites $r = \{\{A_1\}, \{A_1, A_2\}, \{A_1, A_2, A_3\}\} = (A_1, A_2, A_3)$ and $r' = (A'_1, A'_2, A'_3)$ are in the GZ-system Γ. Let $\mathcal{A}_{(r,r')}(\chi)$ be the 4×4 matrix of the form described in section 2.1 such that $E_{r'}(\mathbf{R}) = E_r(\mathbf{R})\mathcal{A}_{(r,r')}(\chi)$ for all configurations $\mathbf{R} \in \mathcal{D}_C(\Gamma)$. This matrix is a function of the adjustable parameter χ and can be computed explicitly and automatically because of the GZ-system formalism (see section 4.2). We wish to build a bridge between r and r' by adding atoms $\{A_0\}, \{A'_0\}$ such that the bond lengths of $\{A_1, A_0\}, \{A_0, A'_0\}, \{A'_0, A'_1\}$ are given and the bond angles associated to the bond pairs

$$\{\{A_2, A_1\}, \{A_1, A_0\}\} = \tau_1,$$
$$\{\{A_1, A_0\}, \{A_0, A'_0\}\} = \tau_0,$$
$$\{\{A_0, A'_0\}, \{A'_0, A'_1\}\} = \tau'_0,$$
$$\{\{A'_0, A'_1\}, \{A'_1, A'_2\}\} = \tau'_1$$

are also given (see figure 3). With χ and the given data fixed the triangle $\{A_1, A_0, A'_1\}$ is determined up to congruence, and if it is rotated about the line through A_1 and A'_1 then $\{A_0\}$ can occupy at most two positions relative to $E_r(\mathbf{R})$ because of the constraint on the measure of τ_1 (assuming the angle $\{\{A_2, A_1\}, \{A_1, A'_1\}\}$ is neither 0 nor π). Let $\mathcal{I} \subset S^1$ denote the set of χ values where $\{A_0\}$ can occupy exactly two positions. For $\chi \in \mathcal{I}$ let the two positions of $\{A_0\}$ be indexed by $\sigma \in \{1, -1\}$. Similarly $\{A'_0\}$ will occupy two positions indexed by $\sigma' \in \{1, -1\}$ for all $\chi \in \mathcal{I}'$.

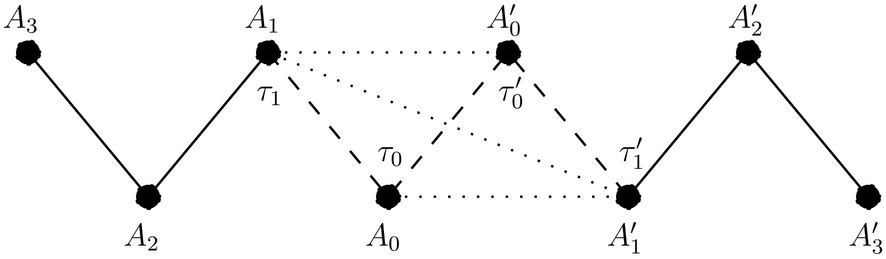

FIGURE 3. Bridging between 2-sites $r = (A_1, A_2, A_3)$ and $r' = (A'_1, A'_2, A'_3)$. Atoms $\{A_0\}$ and $\{A'_0\}$ are added with distance constraints shown as darker dashed lines. Constrained angles $\tau_1, \tau_0, \tau'_0, \tau'_1$ are also shown.

By imposing the constraint on the distance between $\{A_0\}$ and $\{A'_0\}$ one obtains four equations in the one unknown $\chi \in \mathcal{I} \cap \mathcal{I}'$ indexed by $(\sigma, \sigma') \in \{1, -1\}^2$. All solutions in $\mathcal{I} \cap \mathcal{I}'$ of each of these four equations can easily be found, and they determine all of the solutions of the the bridging problem.

In the case of backbone conformations of regular B-form DNA double helices the 2-sites are $r = (O3'.1, C3'.1, C4'.1)$ and $r' = (C5'.2, C4'.2, C3'.2)$ and the bridging atoms are $A_0 = P.2$ and $A'_0 = O5'.2$ (a suffix $.j$ on these atom names indicates that they are part of nucleotide j, $j = 1, 2$). We propose to continue this study by allowing both χ and the Cremer-Pople pseudorotation phase angle P to vary, and to map the solution curves in (χ, P) space, together with the boundaries of the $\mathcal{R}$-space. The parameters χ and P both have multimodal distributions with large variance in statistical studies of DNA crystals [9].

This approach and extensions of it can be applied in many situations. For example it is easy to use this bridging algorithm to map the space of ring conformations for covalently bound rings with six, seven and eight bonds [3], [2], [8]. One can also compute protein secondary structural elements (α, π and 3_{10} helices, parallel and antiparallel β-sheets, γ and β turns, etc. [1]) and could attempt to compute RNA tertiary structural motifs (such as U-turns, tetraloops, cross-strand purine stacks, bulged G, A-platforms, ribose zippers, etc. [32]). This mathematical approach to biomolecular structure teaches us much more than if we simply receive structures from crystallographers or molecular simulations.

3.2. **Computing Geodesics in Ramachandran Spaces.**

$\mathcal{R}$-spaces are equipped with a well-defined Riemannian metric which it inherits from being a subspace of Q. We propose to devise an algorithm to compute this metric (and related quantities, such as $\frac{\partial[\tau(\mathbf{q})^{-1}]}{\partial q_l}$, which appears in the equation of motion for p_l) in terms of a general GZ-system. Expressions using a special internal coordinate system were derived in [29], and might be useful as a starting point.

One application is to be able to compute geodesic curves of this metric starting at any given point of the unit (co)tangent bundle of $\mathcal{R}$-space. Geodesics of this metric are "inertial motions", i.e. solutions of Hamilton's equations with the inertial Hamiltonian $H = \mathbb{T}$. We are interested in finding those directions in the unit (co)tangent space at $\mathbf{q} \in \mathcal{R}(\Gamma)$ for which the resulting geodesic curve stays in $\mathcal{R}(\Gamma)$ for a long time, since these correspond to large "cooperative" motions of the system. We propose to study the hinge motions of the protein T4 lysosyme to see if they are well described by such geodesics. We intend to consult with David and Jane Richardson (Duke University, Biochemistry) on this project.

Another use for this Riemannian metric is to define a distance d between any two points in $\mathcal{R}$-space. This distance could be used to define a notion of an "average conformation". Suppose one is given a finite set of conformations $\{\mathbf{q}^1, \dots, \mathbf{q}^L\}$ which are not too spread out in $\mathcal{R}(\Gamma)$. These could be crystal structures of the same system from different experiments. Rather than averaging each conformational variable (which is the usual procedure) it seems better to define the average conformation to be the element $\bar{\mathbf{q}}$ of $\mathcal{R}(\Gamma)$ such that $\sum_{l=1}^{L} d(\mathbf{q}^l, \bar{\mathbf{q}})^2$ is minimized. This will insure that the conformational coordinates of the "average structure" together satisfy all the usual bond length, bond angle, and steric constraints required of any reasonable structure. We propose to develop algorithms for computing $\bar{\mathbf{q}}$ from the sample $\mathbf{q}^1, \dots, \mathbf{q}^L$. We intend to consult with Ralph Howard (University of South Carolina, mathematics) on the question of the uniqueness of $\bar{\mathbf{q}}$, which is related to strict convexity [11].

4. Software Development Projects

4.1. **Building Large Systems.** As indicated in the introduction we propose to develop software systems to allow the user to create, organize, manipulate, simulate, and otherwise analyze very large systems of biomolecules. Our purpose is not to reproduce in freeware expensive proprietary software systems like INSIGHT II [17], or SCULPT [45], but to go far beyond them with an extendable mathematically integrated system designed to connect cell level modeling with atomic level structural details. The software is to be a means of learning the new and appropriate mathematical ideas needed for molecular biology.

We see a benefit in combining two types of user interfaces. A two dimensional (2D) interface is for system creation, editing and organization. A three dimensional (3D) interface is for visualization and manipulation. IMIMOL [15] is a free program we have developed, together with graphics programmer Scott Johnson (funding from the Industrial Mathematics Institute at the University of South Carolina), with a 2D interface. Visual Molecular Dynamics (VMD) [46] is a free program for all sorts of 3D visualization of molecular systems, including movies of molecular dynamics (MD) trajectories. It is produced and supported by the Theoretical Biophysics Group at the University of Illinois Urbana-Champaign with the intention that it be extended in various ways. We propose to extend and integrate both of these programs (see the next two subsections for details).

An essential test to any software system of the type we are proposing is to apply it to actual large and complex biomolecular systems. Currently a Masters-level mathematics graduate student, Haruna Katayama, is using IMIMOL and RASMOL [40] to build the Light Harvesting Complex II (LH2) from the purple bacterium *Rhodobacter Sphaeroides* [43], [14], [13]. A more or less complete model of the entire Photosynthetic Unit (PSU), which includes multiple copies of LH2, from that species has been built in the lab of Klaus Schulten, and pictures of it are available on his web site [20]. However the details of the structure are available for investigation only to a relatively few members of his lab or collaborators.

LH2, which functions (and looks like) an antennae, is composed of a ring formed from 9 identical subcomplexes (called *heterodimers*). Each heterodimer has two large surfaces, which we will call a and b. The a surface of one heterodimer is closely complementary in shape to the *b* surface of the next heterodimer. Two heterodimers fit together along these surfaces. When 9 heterodimers are put together in this manner they form a continuous ring. Each heterodimer is composed of two (approximately 50 amino acid) protein chains, the alpha chain and the beta chain. These assume primarily alpha helix secondary structure which is aligned roughly perpendicular to the membrane, and roughly parallel to the central axis of the LH2 ring. To each alpha chain is associated two bacteriochlorphyll molecules, which are (roughly planar) porphrin macrocycles. The beta chain is associated with an additional bacteriochlorophyll. Between the alpha and beta chains lies a long carotenoid molecule, spheroidene. Another carotenoid is loosely anchored to the beta chain and one of the bacteriochlorophylls. Photons are absorbed by the bacteriochlorophylls and carotenoids. The arrangement of these elements in the species *Rhodopseudomonas acidophila* can be seen in the file 1KZU.pdb (in the Protein DataBase [38]; use RASMOL). It is difficult to imagine how these flexible molecules assemble themselves in such a beautiful and intricate geometric relationship.

The PSU is mostly a static system (unless one wonders how it assembles) but we must eventually face the complexity of dynamic systems, such as the ribosome. Recently crystal structures have become available [33], [25]. We propose to build an all-atom model of the ribosome as a means of challenging our software and our mathematical descriptions. For this project especially we will need the help of talented undergraduate and graduate students, the support for which we have requested in the budget. These projects will of course enhance their education, which is another type of broader impact of the research. We will also seek the advice of David and Jane Richardson, as well as other structural and molecular biologists.

4.2. **IMIMOL.** As can be seen in figure 1 we do not wish to write down the mathematical components of larger Z-systems. The graphical notation of figure 2 allows calculations (gluings) to be done with larger Z-systems, but this can easily become cumbersome on paper. IMIMOL was created to perform all sorts of Z-system calculations. There is a canvas where atoms can be placed by clicking, and bonds between them can be defined by dragging. The geometry on the canvas is completely adjustable, and has no intrinsic relation to the geometry in space of the molecule described. One must omit bonds from covalently bound rings so that one obtains a tree. The program automatically generates the line graph of the atom/bond tree, and the user selects certain edges of this line graph to be deleted so that one obtains the bond/angle tree. The line graph of this tree is automatically generated as well, and the user again selects the edges of the line graph to be deleted so that one obtains the angle/wedge tree. The orientations of improper wedges can be switched from the default. Various levels of detail can be displayed or hidden. The atom names are hierarchical: "O.His.129.A" might be the name of the Oxygen attached to the backbone in the residue Histidine 129 in the A chain of some protein. Gluing can be accomplished in a "merge" mode in two steps by dragging and dropping A_1 onto the leaf atom B_0, which it will replace, and then by dragging and dropping B_1 onto A_0. The new wedge must be added to the tree by double clicking on its ghost. To assign coordinates one clicks on a bond, angle, or wedge, and edits the value (of type 'string') in the property box. A root site can be defined by clicking on an atom, a bond containing it, and an angle containing it. Z-systems can be exported in an IMIMOL readable (ascii) format which can then be imported into other Z-systems. A rooted Z-system can be exported as a Z-matrix and viewed using various molecular visualization programs, such as RASMOL. Various facilities for building large Z-systems are included such as panning (translating the Z-system on the canvas), zooming in or out, focussing on a part of a Z-system, recentering the molecule, rotating, reflecting, etc. XYZ coordinates from structure files available on the internet can be imposed on a Z-system for the same molecule, essentially accomplishing a conversion from Cartesian to (user defined!) internal coordinates. Likewise the XYZ coordinates of a (numerically) labeled and rooted Z-system can be exported to a file, and then imposed on a new Z-system for the same molecule, accomplishing a conversion from one internal coordinate system to another. One can also define an alternate site (besides the root site) and export a MAPLE procedure file which computes symbolically the 4×4 coordinate transformation matrix from the root system to the alternate system. This feature is extremely useful for studying complex geometric questions about molecular conformations (see subsection 3.1). The WINDOWS executable is available free of charge on the proposer's web site. A UNIX version is also avilable. As the program is still under development, only a sketch of documentation is included in the help menu. A console window displays error and other messages. This program makes Z-systems practically useful.

We propose to extend IMIMOL to fully implement the GZ-system formalism. In particular we want to facilitate building, editing, and visualization (in 2D) of the hierarchy $\mathcal{T}$ (see subsection 2.2), as well as the component tree $\mathcal{S}$ (see subsection 2.1). We intend to allow easy gluing or tethering of components, so that large complex systems can be easily built and organized in a 2D map.

The hierarchical structure of LH2 has motivated the introduction of the tree $\mathcal{T}$. However other modeling features are also suggested which go beyond the organization of nodes in a tree structure. Individual nodes usually have specific patches of their molecular surfaces which serve as functional interfaces with other nodes. For example the a surface patch of one heterodimer and the b surface patch of another heterodimer. The chlorophyll molecules are in contact with one another and this allows electronic excitation to propagate as electric current in wires. These functional relationships should be able to be mapped out using a 2D graphical interface similar to the tools used by silicon chip designers to design computer circuitry. We propose, in collaboration with James Davis (a specialist in chip design software), to extend IMIMOL to allow the mapping out of these physical interfaces and their connections. Eventually we intend to integrate IMIMOL with qualitative modeling environments, such as the one described in [39].

4.3. **VMD.** VMD already can already read molecular structures in most of the common file formats, such as pdb or Z-matrix, but we have extended it to be able to read Z-system specifications written by IMIMOL. This work was done by an undergraduate student Matthew Hielsberg. This extension of VMD does not simply compute the Cartesian coordinates of the atoms and then discard the bonds, angles and wedges; rather these become new data structures within VMD. This gives the user the ability to specify and use a particular internal coordinate system.

VMD has the ability to display a molecular system using a CAVE, which is a sophisticated set of screens with polarized light projectors, polarized sunglasses, a light-weight headset which tells the computer where the viewer is and in which direction he or she is looking, as well as a six-dimensional mouse able to point at and select objects in three dimensions. The two types of polarized light convey the two distinct images to be viewed by the left and right eyes of the viewer, thereby conveying a strong three dimensional image responsive in a natural way to movements of the headset. The Industrial Mathematics Institute (IMI) in the Mathematics Department of the University of South Carolina has a CAVE and has VMD running in CAVE mode and producing this sort of interface. VMD is being further extended to allow selecting and full editing of internal coordinate values for bonds, angles and wedges defined in IMIMOL. One uses the 6D mouse to point (in 3D) to a region of the molecule and the selectable items are highlighted. Once an item (such as a wedge) is selected then the coordinate can be continuously varied using mouse buttons and the effect of this change is immediately fed back to the viewer in real time. One can edit $\mathcal{A}$ matrices by selecting a component and positioning it by hand.

We propose to continue enhancing VMD to provide more advanced features to be made freely available to researchers through the VMD website. One desirable feature is visual feedback of steric clashes such as in the program PROBE [42] produced in the lab of David and Jane Richardson. PROBE is designed to work with their visualization program MAGE [41], but we propose to incorporate it into VMD.

Another feature concerns the animation of biomolecular motions. VMD can already visualize MD trajectories, but we propose to enhance it to allow the visualization of user specified geodesics in the Rammachandran space of the system. We are interested in hinge motions, but also more complex motions such as those involved in the ribosome during protein synthesis.

The standard means of maneuvering in a 3D scene in VMD is not adequate for dealing with a truly large biomolecular system such as a ribosome. Thus we propose to extend VMD to allow “fly through” of the scene to be displayed, similar to the VRML viewer [16]. The objects in the field of view should change their mode of representation depending on the distance from the viewer. Molecules which are close by could have an all atom representation (if desired) whereas molecules much further away could be given an abbreviated polyhedral representation.

Another proposed enhancement involves the computation and display of the complete hydrogen bonding network in a biomolecular system. INSIGHT II can do this, and even the freeware RASMOL [40] can do this for pdb files of proteins. We are interested in applying graph theoretical rigidity tests as in [19]. This display could be used to score and evaluate points of the Rammachandran set as a substitute for attempting to compute the conformational free energy. Taken together the enhanced VMD system would become a powerful tool for “by hand” folding and manipulating biomolecules.

Department of Mathematics, University of South Carolina
E-mail address: dix@math.sc.edu