SOME PROBLEMS IN MATHEMATICAL BIOCHEMISTRY

Daniel B. Dix

Department of Mathematics
University of South Carolina

Contents

Molecular Geometry

Specification. In a recent preprint [D] a new theory for the specification of the internal geometry of biomolecules is studied. One starts with the *molecular graph* $\mathcal{G}$, whose vertices are the atoms and whose edges are covalent bonds between atoms. Suppose $\mathcal{G}$ is a molecular graph with $N \geq 3$ atoms. Let $L^1(\mathcal{G})$ be the *line graph* of $\mathcal{G}$, whose vertices are the edges of $\mathcal{G}$ and whose edges are unordered pairs of distinct edges of $\mathcal{G}$ which share a vertex. Let $L^2(\mathcal{G})$ denote the line graph of $L^1(\mathcal{G})$. Edges of $L^1(\mathcal{G})$ and vertices of $L^2(\mathcal{G})$ are called *angles*. If a is an angle then let $\alpha_1(A)$ denote the atom shared by the two bonds comprising the angle a. Define $AL^2(\mathcal{G})$ to be the subgraph of $L^2(\mathcal{G})\Box L^1(\mathcal{G})\Box\mathcal{G}$ (graph Cartesian product) induced by the subset $\{(a,b,A) \in \mathrm{vert}L^2(\mathcal{G}) \times \mathrm{vert}L^1(\mathcal{G}) \times \mathrm{vert}\mathcal{G} \mid A \in b \in a\}$. The internal geometry of the molecule whose graph is $\mathcal{G}$ is specified by labelling the edges of certain tree subgraphs Γ of $AL^2(\mathcal{G})$ with real numbers. The interpretation of this numerical label depends on the type of the edge. Edges in $AL^2(\mathcal{G})$ are unordered pairs $\{(a_1,b_1,A_1),(a_2,b_2,A_2)\}$, where one of the following three cases holds.

(0) $a_1 = a_2$, $b_1 = b_2$, $A_1 \neq A_2$, and $\{A_1, A_2\}$ is an edge in $\mathcal{G}$.
(1) $a_1 = a_2$, $b_1 \neq b_2$, $A_1 = A_2$, and $\{b_1, b_2\}$ is an edge in $L^1(\mathcal{G})$.
(2) $a_1 \neq a_2$, $b_1 = b_2$, $A_1 = A_2$, and $\{a_1, a_2\}$ is an edge in $L^2(\mathcal{G})$.

Let $\mathrm{edge}_0\Gamma$, $\mathrm{edge}_1\Gamma$, $\mathrm{edge}_2\Gamma$ denote the sets of all edges of Γ of these three types. In case (0) the label is interpreted as the length of the bond $b_1 = b_2$, i.e. the distance between atoms A_1 and A_2. In case (1) the label is interpreted as the angle between bonds b_1 and b_2. And in case (2) the label is interpreted as the angle through which the half plane of the angle a_1 must be rotated so that it coincides with the half plane of the angle a_2, where the axis of rotation is the bond $b_1 = b_2$, oriented from the atom $A_1 = A_2$ toward the other atom of the bond. In this case the sign of the label depends on an assigned orientation of the edge, which is determined by a choice of a root vertex in the tree Γ. A *molecular configuration* is the specification of a position in space for each of the atoms of the molecule. A *molecular conformation* is an orbit of a molecular configuration under the action of the six-dimensional Lie group G of all proper affine symmetries (i.e. translations and proper rotations) of three dimensional space. In that preprint a theorem was proved giving necessary and sufficient conditions on the rooted tree subgraph Γ such that the associated system of internal coordinates defines a diffeomorphism between the set of all molecular conformations in which all the angles in $\pi_1(\mathrm{vert}\Gamma)$ are nondegenerate (having bonds with positive bond lengths and being noncollinear) and the set $(0,\infty)^{\mathrm{edge}_0\Gamma} \times (0,\pi)^{\mathrm{edge}_1\Gamma} \times (S^1)^{\mathrm{edge}_2\Gamma}$. These conditions are as follows.

(1) $\pi_3(\mathrm{vert}\Gamma) = \mathrm{vert}\mathcal{G}$, i.e. for every atom $A \in \mathrm{vert}\mathcal{G}$ there exists a vertex $(a, b, A) \in \mathrm{vert}\Gamma$ based at that atom.
(2) For every $A \in \mathrm{vert}\mathcal{G}$ the subgraph Γ_A of Γ induced by $\mathrm{vert}\Gamma_A = \{(a, b, A') \in \mathrm{vert}\Gamma \mid A' = A\}$ is connected.
(3) For every $(a, b, A) \in \mathrm{vert}\Gamma$, where $A \neq \alpha_1(a)$, there exists $\{(a, b, \alpha_1(a)), (a, b, A)\} \in \mathrm{edge}_0\Gamma$.
(4) For every $a \in \pi_1\mathrm{vert}\Gamma$, where $a = \{b_1, b_2\}$, there exists $\{(a, b_1, \alpha_1(a)), (a, b_2, \alpha_1(a))\} \in \mathrm{edge}_1\Gamma$.
(5) For every bond $b \in \pi_2(\mathrm{vert}\Gamma)$, where $b = \{A_1, A_2\}$ there exists $a \in \pi_1(\mathrm{vert}\Gamma)$ such that $\{(a, b, A_1), (a, b, A_2)\} \in \mathrm{edge}_0\Gamma$.
(6) For every vertex $(a, b, A) \in \mathrm{vert}\Gamma$ the subgraph $\Gamma_{b,A}$ of Γ induced by $\mathrm{vert}\Gamma_{b,A} = \{(a', b', A') \in \mathrm{vert}\Gamma \mid b' = b, A' = A\}$ is connected.

Tree subgraphs Γ satisfying these necessary and sufficient conditions are called *GZ-trees*. If $\mathcal{G}$ has N vertices (atoms), then every GZ-tree Γ consequently has $3N - 6$ edges. When the edges of a GZ-tree are labelled by internal coordinates of the appropriate type we call it a *3D molecule*. This name is most appropriate when a further condition of being *sterically allowed* is imposed (see section 1.3).

GZ-trees then emerge as a very natural, interesting and important class of graphs. There are several problems concerning these graphs which are still unresolved. The most important of these from the point of view of applications is to find an algorithm which always generates a GZ-tree and which can be used to generate any GZ-tree. We call this a *generating algorithm*. An example which hopefully explains the idea of a generating algorithm is the definition of a *Z-tree*. Suppose Γ is a subgraph of $AL^2(\mathcal{G})$. We say Γ is a *Z-tree* if there is an increasing sequence $(\Gamma_1, \ldots, \Gamma_{N-2})$ of subgraphs of Γ satisfying the following conditions:

(1) Γ_1 is a linear graph (called the *trunk*) of three edges and four vertices based on

a single angle $a_1 \in \mathrm{vert}L^2(\mathcal{G})$. So there exists $a_1 = \{b_1, b_2\} \in \mathrm{vert}L^2(\mathcal{G})$, where $b_1 = \{A_1, A_2\}$ and $b_2 = \{A_2, A_3\}$, such that the vertices of Γ_1 are

$$(a_1, b_1, A_1)\text{———}(a_1, b_1, A_2)\text{- - -}(a_1, b_2, A_2)\text{———}(a_1, b_2, A_3),$$

and each pair of consecutive vertices in the above is an edge of Γ_1.

(2) For each $2 \leq j \leq N-2$, Γ_j is obtained by attaching a linear chain (called a *branch*) of three edges and three vertices arising from a single new atom to a single vertex of Γ_{j-1}. So there exists $(\tilde{a}_j, \tilde{b}_j, \tilde{A}_j) \in \mathrm{vert}\Gamma_{j-1}$ (called the *vertex of attachment*), and $A_{j+2} \in \mathrm{vert}\mathcal{G} \setminus \pi_3(\mathrm{vert}\Gamma_{j-1})$ (called the *new atom*) such that $b_{j+1} = \{\tilde{A}_j, A_{j+2}\} \in \mathrm{edge}\mathcal{G}$, such that if $a_j = \{\tilde{b}_j, b_{j+1}\}$, then $\mathrm{edge}\Gamma_j \setminus \mathrm{edge}\Gamma_{j-1}$ contains the three edges formed from pairs of consecutive vertices from the following:

$$(\tilde{a}_j, \tilde{b}_j, \tilde{A}_j)\cdots\cdots(a_j, \tilde{b}_j, \tilde{A}_j)\text{- - -}(a_j, b_{j+1}, \tilde{A}_j)\text{———}(a_j, b_{j+1}, A_{j+2}).$$

Also, $\mathrm{vert}\Gamma_j \setminus \mathrm{vert}\Gamma_{j-1}$ contains the last three vertices in the above.

(3) $\Gamma_{N-2} = \Gamma$.

Edges of type 0, 1, or 2 are denoted by solid, dashed, or dotted lines respectively.

Every Z-tree is a GZ-tree; in fact the name GZ-tree comes from the idea of a *generalized* Z-tree. Rooted Z-trees generate internal coordinate systems of *Z-matrix type*, which are well known to computational chemists (although not in this graph theoretical guise) [P]. Z-trees got their name from Z-matrices. Z-trees as a class of graphs has a generating algorithm: one starts with a trunk and successively adds branches corresponding to new atoms until one runs out of atoms. This generating algorithm has made Z-trees useful in applications since they can be generated easily for molecules in practice. The proposer plans to consult with his colleagues László Székely, Jerry Griggs, and David Sumner, and possibly collaborate with them on the answer to this question.

Linking. Not every GZ-tree is a Z-tree. Examples come from the consideration of the operation of linking two 3D molecules to form a new 3D molecule. Suppose $(\mathcal{G}_1, \Gamma_1, r_1, \mathcal{A}_1)$, $(\mathcal{G}_2, \Gamma_2, r_2, \mathcal{A}_2)$ are 3D molecules, i.e. $\mathcal{G}_i$ is a molecular graph with three or more atoms, (Γ_i, r_i) is a rooted GZ-tree in $AL^2(\mathcal{G}_i)$, and $\mathcal{A}_i \in (0,\infty)^{\mathrm{edge}_0\Gamma_i} \times (0,\pi)^{\mathrm{edge}_1\Gamma_i} \times (-\pi,\pi]^{\mathrm{edge}_2\Gamma_i}$ is the labelling of the edges of Γ_i with the appropriate internal coordinates. We assume that $\mathrm{vert}\mathcal{G}_1 \cap \mathrm{vert}\mathcal{G}_1 = \emptyset$, i.e. the *intermolecular* case. Molecules 1 and 2 will become linked through a chemical reaction in which some bonds will break and other new bonds will form. This leads us to consider a *supermolecular graph* $\mathcal{G}^\ddagger$ which is appropriate for specifying the internal geometry of the transition state of the reaction, i.e. it includes as edges both breaking and forming covalent bonds. Both $\mathcal{G}_1$ and $\mathcal{G}_2$ are subgraphs of $\mathcal{G}^\ddagger$. We assume that there exists an edge $b = \{A_1, A_2\} \in \mathrm{edge}\mathcal{G}^\ddagger$ where A_1 is an atom of molecule 1 and A_2 an atom of molecule 2. We chose vertices $(a_1, b_1, A_1) \in \mathrm{vert}\Gamma_1$ and $(a_2, b_2, A_2) \in \mathrm{vert}\Gamma_2$. These vertices can be connected by a six edge construct called a *linker*. If the edges of the linker are labelled appropriately then the relative position and orientation of the two molecules will be fixed. A linker consists of the following edges.

$$(a_1, b_1, A_1)\cdots\cdots(\{b_1, b\}, b_1, A_1)\text{- - -}(\{b_1, b\}, b, A_1)\text{———}(\{b_1, b\}, b, A_2),$$
$$(\{b_1, b\}, b, A_2)\cdots\cdots(\{b_2, b\}, b, A_2)\text{- - -}(\{b_2, b\}, b_2, A_2)\cdots\cdots(a_2, b_2, A_2).$$

The first line describes a branch starting at the vertex (a_1, b_1, A_1) of Γ_1, with new atom A_2. The second line starts at the same vertex with which the first line ended, so the two lines taken together define a connected linear graph. The second line is called a *swivel*. It ends at the vertex (a_2, b_2, A_2) of Γ_2. If we add the linker to the union of the two GZ-trees Γ_1 and Γ_2 then we obtain a subgraph $\Gamma^\ddagger$ of $AL^2(\mathcal{G}^\ddagger)$ which turns out to be a GZ-tree. If both Γ_1 and Γ_2 are Z-trees, then $\Gamma^\ddagger$ will be a GZ-tree which is *not* a Z-tree.

Thus the GZ-tree is seen to be not an idle generalization of the Z-tree. However, it might be true that every GZ-tree can be obtained by linking together some Z-trees. The proposer hopes to answer this question with the help of his colleagues (listed earlier).

If we drop the root r_2, then the edges of the linker acquire an orientation from the root r_1 and we can label these edges with appropriate numbers with unambiguous geometrical meaning. The labelling $\mathcal{A}_2$ may need to be adjusted slightly because the root r_1 may cause certain edges of type 2 to switch their orientation. This involves simply changing the signs of the labels of the affected edges of type 2. (Edges of types 0 and 1 are not affected.) Let $\tilde{\mathcal{A}}_2$ denote this adjusted labelling. When we enlarge $\mathcal{A}_1$ and $\tilde{\mathcal{A}}_2$ by including the labels of the edges of the linker, we obtain $\mathcal{A}^\ddagger$. Thus we arrive at $(\mathcal{G}^\ddagger, \Gamma^\ddagger, r_1, \mathcal{A}^\ddagger)$, which is a special type of 3D molecule called a *linked species*.

Suppose we decided to link the two 3D molecules in the reverse order to the above, but using the same vertices to attach the linker. This would yield the linked species $(\mathcal{G}^\ddagger, \Delta^\ddagger, r_2, \mathcal{B}^\ddagger)$. If Γ and Δ are two GZ-trees in $AL^2(\mathcal{G})$ we write $\Gamma \sim \Delta$ if $\pi_1 \mathrm{vert}\Gamma = \pi_1 \mathrm{vert}\Delta$. It follows from $\Gamma \sim \Delta$ that $(0,\infty)^{\mathrm{edge}_0\Gamma} \times (0,\pi)^{\mathrm{edge}_1\Gamma} \times (S^1)^{\mathrm{edge}_2\Gamma}$ is diffeomorphic to $(0,\infty)^{\mathrm{edge}_0\Delta} \times (0,\pi)^{\mathrm{edge}_1\Delta} \times (S^1)^{\mathrm{edge}_2\Delta}$. In our case it is clear that $\Gamma^\ddagger \sim \Delta^\ddagger$ and we can choose the labels of the linker in $\mathcal{B}^\ddagger$ such that $\mathcal{A}^\ddagger$ and $\mathcal{B}^\ddagger$ correspond to one another in the associated diffeomorphism. (Geometrically this is simple.) These two 3D molecules would describe the same conformation of the linked species. This gives us a very natural equivalence relation in the set of 3D molecules. Thus the operation of linking two 3D molecules is commutative modulo this equivalence relation. Linking is also associative provided the appropriate pairs of 3D molecules are linkable, i.e. we do not try to link molecule 3 to one of the non-endpoint vertices of the linker of molecules 1 and 2. The details of this need to be worked out.

Chemical reactions are symmetric in the sense that theoretically they can go in either direction, reactants to products or products to reactants. What we have done so far is to give a way of describing the geometry of the transition state (a linked species) either from the perspective of the reactants or from that of the products. However these two perspectives will in general yield different descriptions of the transition state geometry, although the underlying conformation of the transition state will be the same. If the transition state conformation is nondegenerate with respect to both GZ-trees then the main theorem of [D] gives us a diffeomorphism which will map one description into the other. The proposer intends to study this diffeomorphism carefully, so as to write it down as explicitly as possible. This theory will formalize the mathematical study of intermolecular chemical reactions. This theory will be applied to the formation of biopolymers: proteins, nucleic acids, and polysaccharides. Labelled Z-trees for the monomers have already been given in [D].

Rings. There are many interesting reactions which are *intramolecular*, in that two parts of the same molecule react with each other to produce a new molecule with a ring of covalent bonds. The examples we have in mind are the ring closure reactions in sugars, and the formation of disulfide bonds in proteins. Since bond lengths of covalently bonded pairs of atoms are stongly constrained under normal biological conditions, the existence of a ring involves constraints on the labels of the edges of Γ which are more complicated than merely fixing the value of a few labels. These issues have been the subject of a good amount of study [CH], [GS], but many interesting questions remain and connections to other fields of mathematics, such as algebraic geometry, can be made [EM]. Our thoughts have centered mainly around the problem of small rings such as the cyclopentane ring and the furanose ring of nucleotides (both five membered) and the pyranose ring of glucose (six membered). Larger rings have have greater flexibility and more complicated issues arise which we cannot address (see however the next section).

In five membered rings one considers the five bond lengths to be fixed, but the five endocyclic bond angles are not fixed, or even equal to each other [C]. This gives rise to a four dimensional manifold of ring conformations, which we call the *ring manifold*, which is embedded in a ten dimensional space of conformational variables—five bond angles and five torsion angles. In cyclopentane there is a one dimensional submanifold of the ring manifold, diffeomorphic to the circle, on which the molecular potential energy is nearly constant, and away from which the energy increases markedly. The exact definition of this manifold depends on the approximation one uses for the potential energy function, but perhaps there is a geometrically natural definition of this submanifold. Motion along this path is called *pseudorotation*. This situation implies the existence of two natural coordinates on the ring manifold, a radial coordinate and an angular coordinate tuned to the *pseudorotation submanifold*. These are usually called *puckering coordinates*, and competing definitions of them exist [AS], [CP]. It would be quite interesting to find a natural geometrical definition of the pseudorotation submanifold, and also a natural pair of complementary coordinates to the two puckering coordinates, so that taken together we would have a set of four coordinates intrinsic to the ring manifold.

Molecules containing rings, and especially those with fused rings are important in biochemistry. Examples include the bases of nucleotides, amino acids histidine, phenylalanine, tryptophan and tyrosine, steroids such as cholesterol, and porphyrin structures such as heme. If the structure can be treated as rigid, then any difficulty in setting up the labelled GZ-tree initially is not so important. In fact most of the atomic details are also not important. What is important is usually how the larger rigid structure interacts with other polyatomic structural elements. For example, in DNA the complementary bases pair up with a certain hydrogen bonding pattern and then these base pairs stack on top of each other to form the center of the double helix. The important variables are not individual bond angles or torsion angles, but propeller twist angles, tilt angles, and roll angles of the planes of the bases with respect to the helix axis or its normal plane [S]. These variables are complicated functions of the individual internal coordinates labelling the edges of the GZ-tree of the molecule. Similar issues arise in secondary structural elements of proteins, such as α helices or β sheets. These examples suggest that we pass from the labelled GZ-

tree description to a description involving linked rigid bodies. This is a common subject in the robotics literature, and some formalism has already been developed [MZW]. However, we propose to show how the two levels of description relate to one another in the important examples from biochemistry. In the process we hope to refine the formalism so that it will be an adequate foundation on which to build even higher levels of structure.

Low Energy Molecules. One thing we have been mostly ignoring in our discussion so far has been the fact that molecular potential energy is much lower for some conformations than others. For example, bond lengths are so tightly constrained that we should consider them as being constant. Bond angles are also constant to within a few degrees. At chiral centers there are significant energetic barriers to epimerization (i.e. a reaction where the chirality is reversed). Finally atoms occupy space, and do not share that space without a steep energetic penalty except in the case of a pair of atoms which are covalently bonded. It appears that one can restrict attention to *low energy molecules* by imposing purely geometric constraints on the conformation.

Let X denote three dimensional space, and $\mathcal{X}\colon \mathrm{vert}\mathcal{G} \to X$ a molecular configuration. Let G denote the group of all proper affine symmetries of X. A molecular conformation $G\mathcal{X}$ can be described using a rooted labelled GZ-tree $(\mathcal{G}, \Gamma, r, \mathcal{A}_{\mathcal{X}})$.

(1) The edge $e = \{(a, b, A_1), (a, b, A_2)\} \in \mathrm{edge}_0\Gamma$ is associated to the function $\mathcal{X} \mapsto \mathcal{A}_{\mathcal{X}}(e) = \|\mathcal{X}(A_1) - \mathcal{X}(A_2)\|$.

(2) The edge $e = \{(a, b_1, A), (a, b_2, A)\} \in \mathrm{edge}_1\Gamma$ is associated to the function

$$\mathcal{X} \mapsto \mathcal{A}_{\mathcal{X}}(e) = \frac{\mathcal{X}(A_1) - \mathcal{X}(A)}{\|\mathcal{X}(A_1) - \mathcal{X}(A)\|} \cdot \frac{\mathcal{X}(A_2) - \mathcal{X}(A)}{\|\mathcal{X}(A_2) - \mathcal{X}(A)\|},$$

where $b_1 = \{A_1, A\}$ and $b_2 = \{A_2, A\}$.

(3) The oriented edge $\tilde{e} = ((a_1, b, A), (a_2, b, A))$ whose underlying unordered pair is in $\mathrm{edge}_2\Gamma$ is associated to the function

$$\begin{aligned} \mathcal{X} \mapsto \mathcal{A}_{\mathcal{X}}(\tilde{e}) &= \mathbf{e}_1 \cdot \mathbf{e}_2 + i\mathbf{e}_1 \times \mathbf{e}_2 \cdot \mathbf{e}_3, \\ \mathbf{e}_3 &= \frac{\mathcal{X}(A') - \mathcal{X}(A)}{\|\mathcal{X}(A') - \mathcal{X}(A)\|}, \\ \mathbf{U}_j &= \mathcal{X}(A_j) - \mathcal{X}(A), \qquad j = 1, 2, \\ \mathbf{e}_j &= \frac{\mathbf{U}_j - \mathbf{e}_3[\mathbf{e}_3 \cdot \mathbf{U}_j]}{\|\mathbf{U}_j - \mathbf{e}_3[\mathbf{e}_3 \cdot \mathbf{U}_j]\|}, \qquad j = 1, 2, \end{aligned}$$

where $a_j = \{b_j, b\}$, $b_j \setminus b = \{A_j\}$, $j = 1, 2$, and $b = \{A, A'\}$.

We will identify a low energy molecule with the set of its low energy conformations. We consider only conformations which are nondegenerate with respect to Γ. Since we do not yet have a geometric way to describe the all the low energy conformations of puckered five membered rings (such as the furanose ring in nucleotides or the pyrolidine ring in proline) we will assume for this discussion that all five or six membered rings are rigid. Thus for all $e \in \mathrm{edge}_0\Gamma$ we impose the condition that $\mathcal{A}_{\mathcal{X}}(e) = l(e) > 0$. For all $e \in \mathrm{edge}_1\Gamma$ we impose

the condition that $\mathcal{A}_{\mathcal{X}}(e) = c(e) \in (-1, 1)$. If $e \in \text{edge}_2\Gamma$, say $e = \{(a_1, b, A), (a_2, b, A)\}$ as above, then there are two cases. If $\alpha_1(a_1) = \alpha_1(a_2)$ then the edge is called an *improper*, and we impose the condition $\mathcal{A}_{\mathcal{X}}(e) = z(e) \in S^1$. This has the effect of fixing chirality (among other things). If $\alpha_1(a_1) \neq \alpha_1(a_2)$ then the edge is called an *dihedral*, and it is unrestricted unless the bond b is part of a cycle (ring) in the graph $\mathcal{G}$. If the bond b is part of a five or six membered ring then (as we indicated above) we impose the condition $\mathcal{A}_{\mathcal{X}}(e) = z(e) \in S^1$. Taken together, these constraints determine a submanifold of (nondegenerate) conformation space diffeomorphic to a torus, whose dimension is the number K of unconstrained dihedral angles (torsion angles). If the smallest ring the bond b is part of has seven or more bonds in it then instead of imposing the ring closure constraint by restricting dihedral angles we impose the remaining one bond length constraint and the remaining two bond angle constraints directly on the configuration $\mathcal{X}$. These last three constraints will not have been imposed earlier as part of Γ since the tree nature of Γ requires it to omit one bond and two bond angles (containing that bond) from every ring. (An example of this situation is a disulfide bond in proteins.) It would be nice to prove that these additional constraints determine a submanifold of the above K-dimensional torus.

Each type of chemical element is assigned a *van der Waals radius*, and a sphere of that radius is assumed to be centered at the point of space where the nucleus is located. If two atoms are not covalently bonded to each other then the interiors of their respective spheres cannot intersect. The van der Waals radii are given by a function $r\colon \text{vert}\mathcal{G} \to (0, \infty)$. Let $\binom{\text{vert}\mathcal{G}}{2}$ denote the set of all two element subsets of $\text{vert}\mathcal{G}$. We let $\mathcal{V} \subset \binom{\text{vert}\mathcal{G}}{2} \setminus \text{edge}\mathcal{G}$ be the subset consisting of all pairs of atoms for which we must impose this restriction. For example, suppose A_1 is a hydrogen atom covalently bonded to an electronegative atom like oxygen or nitrogen (called the *donor*) and A_2 is some other electronegative atom (called the *acceptor*). Then we should exclude the pair $\{A_1, A_2\}$ from $\mathcal{V}$ since in hydrogen bonds the van der Waals sphere of the hydrogen atom overlaps significantly with that of the acceptor atom, whereas the van der Waals spheres of the donor and acceptor atoms are usually touching but not significantly overlapping.

We say that a conformation $G\mathcal{X}$ is *sterically allowed* if for all $\{A_1, A_2\} \in \mathcal{V}$ we have $\|\mathcal{X}(A_1) - \mathcal{X}(A_2)\| \geq r(A_1) + r(A_2)$. When we consider the closed subset of all confomations in the above submanifold which are also sterically allowed we obtain all the low energy conformations of the molecule. Call this set $\mathcal{R}$ (in honor of Ramachandran, a famous protein chemist). We would like to prove that the sets $\mathcal{R}$ for all the basic molecules of biochemistry (all of which have the property that for each bond the smallest ring containing that bond has length less than seven) have positive volume with respect to a K-dimensional volume element of the torus.

Rather than attempting an exhaustive case by case analysis we would like to develop a general theory which will enable us to build up the $\mathcal{R}$ set of a larger molecule from the $\mathcal{R}$ sets of its smaller pieces. The theory of section 1.2 allows us to relate the GZ-trees of the pieces to those of the larger conglomerate. The intermediates of the reactions involve conformations not in the $\mathcal{R}$ sets since bonds break and bond angles are distorted. However the overall result can often (not always) be accomplished by breaking a single bond in each of the smaller molecules, and joining the two "ends", just as one would do with

molecular models. I would like to develop this into an operation on $\mathcal{R}$ sets in a manner analogous to the linking operation between GZ-trees. I would like to prove a theorem giving a lower estimate of the volume of the $\mathcal{R}$ set of the conglomerate from geometric information and the volumes of the $\mathcal{R}$ sets of the pieces. Then starting from a small number of rigid molecules one would attempt to build up all the basic molecules of biochemistry by repeated applications of this theorem.

One goal is to prove that the $\mathcal{R}$ set of a protein formed from an arbitrary amino acid sequence has positive volume. A more refined goal is to provide asymptotic estimates for this volume as the number of amino acids in the chain increases. Of course volume is a rather crude measure of the $\mathcal{R}$ set. It is of great practical interest to have more information about how the $\mathcal{R}$ set sits inside the K-dimensional torus, if that information enables us to more effectively sample low energy conformations.

Packing. $\mathcal{R}$ sets are the mathematical tools which enable us to formulate precisely a great many questions in the geometry of biomolecules. For example, can a given sequence of amino acids form an α-helix? An α-helix is a conformation of a polypeptide chain where a linear hydrogen bond exists between the amide hydrogen of residue $i+4$ and the carbonyl oxygen of residue i. This establishes a ring of length 13 in the molecule. If the backbone dihedrals are assumed to be periodic then solutions to the ring closure problem have been found [Q]. If we forget about side chains is this solution in the $\mathcal{R}$ set? What about if we include side chains? This sort of thing has been studied by chemists, but no rigorous theorems have been proved.

These questions have to do with certain types of packing of linked spheres in space, and so are related to other sphere packing problems, even the simplest of which are notoriously difficult. However we are not trying to find the best way to pack these spheres, but rather if a certain geometrical construction is sterically feasible. Nevertheless, the packing of secondary structural elements (α-helices, β-sheets) and side chains in the interior of proteins is very nearly the density of close packed spheres [Cr].

Another interesting issue related to packing and folded proteins has to do with packing water molecules around the protein and inside a container. Water molecules form an interesting hydrogen bonding network in ice which can be easily described geometrically. But the density of ice is about 10% less than the usual density of liquid water, which is the environment of the folded protein. The network of hydrogen bonds formed between the surrounding water and the folded protein is an important part of the folded structure. Something about the folded structure gives the water more freedom than it would have if the protein were unfolded. It would be very interesting to devise a method of packing water molecules around a given protein conformation and inside a container which would maximize the number of hydrogen bonds formed and achieve the correct density.

2 Protein Folding

2.1 Dynamical System. The fact that when one surrounds an unfolded protein with water molecules at the appropriate density, temperature, and pressure, it spontaneously folds upon itself culminating, in not very much time, in a very specific three dimensional conformation of the protein (the same one every time the experiment is repeated) is surely

one of the big wonders of biochemistry. Physically it is a behavior intrinsic to many body problems, and it is still beyond our computational simulation capabilities. It is also a considerable challenge to model mathematically [N].

We would like to set up a smooth dynamical system whose large-time behavior models the protein folding process. This will not be for simulation purposes but as a foundation for simpler and more refined models. Furthermore we hope to illustrate in a mathematically rigorous manner certain general physical principles at work in protein folding.

We begin with a classical mechanical system of particles. The particles are the nuclei of all the atoms in the system, which consists of a protein molecule in its ionized form which predominates at ph 7 and very many complete water molecules. Suppose there are N nuclei altogether, with postion vectors (in an inertial reference frame) $\mathbf{r}_1, \ldots, \mathbf{r}_N$, masses $m_1, \ldots, m_N$, and momentum vectors $p_1, \ldots, p_N$. Each of these particles is positively charged and interacts with the other particles through electrostatic repulsion. But this repulsive force is compensated by the presence of very many electrons (if the atomic numbers of the nuclei are $Z_1, \ldots, Z_N$ and the net charge on the protein is Qe, where $-e$ is the charge on an electron, then there are $\sum_{j=1}^{N} Z_j - Q$ electrons in the system.) We cannot account for the positions of these electrons classically; this requires quantum mechanics. But because electrons are 10^3 to 10^4 times lighter than the nuclei, they adjust very rapidly to any change in the positions of the nuclei. (Here we are assuming low energies so the nuclei are definitely nonrelativistic.) Thus, using an approximation introduced by Born and Oppenheimer, we may assume the nuclei are fixed in position and find the lowest energy electronic configuration. One must be careful when $Q < 0$ because clearly if $|Q|$ is too large, some electrons may escape to infinity. Let the energy of this configuration (including the electrostatic repulsion of the nuclei) be denoted by $V(\mathbf{r}_1, \ldots, \mathbf{r}_N)$. This function can be rigorously defined in terms of the machinery of quantum mechanics, but it is very complicated, even to compute numerically. For systems of the size we are discussing no method of computing V has any rigorous claim to accuracy. Nevertheless V is a mathematically well-defined function, and certain of its properties can be proved to hold. If $1 \leq i < j \leq N$ and $\|\mathbf{r}_i - \mathbf{r}_j\| \to 0$ then $V(\mathbf{r}_1, \ldots, \mathbf{r}_N) \to \infty$. Also, if for all $(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ in an open set U we have that the electronic ground state is nondegenerate, then V is a smooth function on U. (We may have to prove these if we cannot find the proofs in the literature.)

V has singularities in its first derivatives on a certain subset of nuclear configuration space, called *conical singularities*. For example, in the system of a single water molecule conical singularities exist on the submanifold of configuration space where the three nuclei lie on the same line (not all collinear configurations are singular) [Mu]. However, at a temperature of 300 kelvin the probability density at these singularities is about e^{-190} times the probability density at the ground state configuration of a water molecule. This is because the conical singularity occurs at an energy at least 5eV above the ground state energy. Such singularities have also been found to be a means by which many photoexcited molecules relax to their electronic ground states. But general experience leads one to the conjecture that there is a critical energy E_c such that whenever $V(\mathbf{r}_1, \ldots, \mathbf{r}_N) < E_c$ there is no conical singularity, and the energy E_c is large enough so that all normal biological

processes not involving photoexcitation take place below this energy level. The proof of this conjecture seems to be extremely difficult, since few mathematical methods exist for proving the nondegeneracy of the ground state electronic eigenstate [DL]. Our pragmatic approach is therefore to *assume* this conjecture holds in regard to the phenomenon of protein folding.

Unless our system of particles is artificially confined to a region of space it will probably spread out under the influence of the potential V. (It is possible that for droplets of a certain size that surface tension would mostly hold the system together, but water molecules would occasionally escape, cooling the droplet. The large time behavior would probably not be a droplet in liquid form.) If we confine the system to a fixed region of space, say spherically symmetric, the linear momentum will not be strictly conserved, even if the initial condition has zero total linear momentum. However mostly it will be very close to zero. This complicates the analysis of the large time behavior. If we assume periodic boundary conditions (as is often done in simulations of liquids) then the "angular momentum" will not be strictly conserved, even if it is zero initially. Also these boundary conditions are problematic unless the system has $Q = 0$. We could confine the system to a 3-sphere with small positive curvature, but the analogs of linear and angular momentum do not Poisson commute and this leads again to approximate conservation laws which are not exact. However there is a trick to resolve this problem nicely, which is due to the proposer and Vassiliy Lubchenko. We introduce another artificial particle, the *containment particle*, of position $\mathbf{r}_0$, mass m_0, and momentum $\mathbf{p}_0$ to the system. This particle interacts with the real particles through a smooth potential $V_c(\|\mathbf{r}_j - \mathbf{r}_0\|)$. This potential satisfies $V(l) = 0$ if $0 \leq l \leq a$, $V(l) > 0$ for $a < l < b$, and $V(l) \to \infty$ as $l \to b^-$. Thus a real particle within a distance of a of the containment particle has no interaction with it at all. For larger distances there is a strong confining force. By adjusting N, and $a < b$ one can insure proper density of the system. The total linear and angular momentum of the system (including the containment particle) is exactly conserved, hence we may restrict attention to the case where both of these types of momentum are zero. Using the reduction theory [AM] we obtain a smooth Hamiltonian system on the cotangent space $T^*\mathcal{C}$, where $\mathcal{C}$ is the manifold of noncollinear conformations of the system of $N + 1$ particles. The level sets of the Hamiltonian are compact, and are manifolds for almost every value of the energy. On such compact symplectic manifolds the Hamiltonian vector field is complete, and so we obtain a flow ϕ_t, $t \in \mathbb{R}$. We restrict attention to one such level set Φ_E of energy $E < E_c$. This level set is equipped with an invariant measure, obtained from the symplectic volume form by "dividing by dH". This measure μ is finite.

Depending on the value of E we expect Φ_E to have many connected components. There are a variety of chemical reactions which might take place in our system, each with its own transition state $S^\ddagger$. If $E < V(S^\ddagger)$ then the system does not have enough energy for the reaction to take place. Some examples of these reactions are the following.

(1) Hydrolysis of a peptide bond.
(2) Epimerization at any of the backbone C^α atoms, or at the C^β atoms of isoleucine or threonine.
(3) Formation or breakage of a disulfide bond.

(4) Exchange of a proton between the protein and the water, or between two water molecules.

If E is small enough so that reaction (1) cannot take place, then we are only interested in the connected component(s) of Φ_E where all the peptide bonds remain intact. Likewise if E is small enough so that reaction (2) cannot take place, then we are only interested in the connected component(s) of Φ_E where the chirality is correct at all the stereocenters. Thus we must be careful of discrete constants of the motion. Suppose it is possible to choose E in this manner. Suppose one of the "interesting" connected components contains a phase point corresponding to a low energy sterically allowed conformation of an intact protein packed to the correct density with intact water molecules. Redefine Φ_E to be this connected component and renormalize the measure μ to be a probability measure on it.

Statistically we can model an unfolded protein by a probability measure $\mu_0 = \rho_0\, d\mu$ where ρ_0 is chosen so that the expected radius of gyration of the non-hydrogen protein atoms is too large for the protein to be even partly folded. This can be done using Jaynes' principle of maximum statistical entropy [B]. Then the *statistical mechanical protein folding problem* is to understand the large-time behavior of the transported measure $\mu_t = \mu_0 \circ \phi_{-t}$ resulting from the initial measure μ_0 on Φ_E.

If we apply the *Ergodic Decomposition Theorem* [M] to our dynamical system we obtain a set R, a mapping $\psi\colon \Phi_E \to R$, and a probability measure ν_r on the Borel subsets of $\psi^{-1}(\{r\})$ for each $r \in R$, such that $\psi^{-1}(\{r\}) \subset \Phi_E$ is a $\{\phi_t\}_{t\in\mathbb{R}}$-invariant Borel subset for each $r \in R$, the measure ν_r is $\{\phi_t\}_{t\in\mathbb{R}}$-invariant and ergodic for $\tilde{\mu}$-almost every $r \in R$, where $\tilde{\mu}(E) = \mu(\psi^{-1}(E))$ whenever $E \subset R$ and $\psi^{-1}(E)$ is a Borel subset of Φ_E, and for every integrable function f on Φ_E we have

$$\int_{\Phi_E} f\, d\mu = \int_R \int_{\psi^{-1}(\{r\})} f\, d\nu_r\, d\tilde{\mu}.$$

If we apply the *Birkhoff-Khinchin Ergodic Theorem* [CFS] to the function ρ_0 we obtain the existence for μ-almost every $x \in \Phi_E$ of

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \rho_0(\phi_{-t}(x))\, dt \stackrel{\text{def}}{=} \rho_\infty(x), \tag{1}$$

where ρ_∞ is constant on $\psi^{-1}(\{r\})$ for each $r \in R$ and

$$\rho_\infty(x) = \tilde{\rho}_\infty(r) = \int_{\psi^{-1}(\{r\})} \rho_0\, d\nu_r,$$

for all $x \in \psi^{-1}(\{r\})$. Since $\mu_t = \rho_t\, d\mu$, where $\rho_t(x) = \rho_0(\phi_{-t}(x))$ for all $x \in \Phi_E$, this gives us information about the large-time behavior of μ_t in the *Cesaro mean*. In fact the above limit holds in an L^2-sense on Φ_E (the *von Neumann Ergodic Theorem*) and hence for all Borel sets $A \subset \Phi_E$ we have

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \mu_t(A)\, dt = \int_A \rho_\infty\, d\mu = \int_R \tilde{\rho}_\infty(r) \nu_r(\psi^{-1}(\{r\}) \cap A)\, d\tilde{\mu}. \tag{2}$$

The Cesaro mean is necessary in (1) since for each fixed x the limit of $\rho_t(x) = \rho_0(\phi_{-t}(x))$ as $t \to \infty$ almost surely does not exist. However, if the dynamical system is *mixing* on each ergodic component, then the Cesaro mean is not necessary in (2). These theorems describe the general phenomenon of *thermodynamic relaxation*.

We have taken great pains to try to insure that Φ_E consists of a single ergodic component, i.e. R is a singleton. But this is extremely difficult to prove. However if we assume this *ergodic hypothesis* then we can compute the large-time state ρ_∞ much more explicitly, namely $\rho_\infty \equiv 1$. Thus under the ergodic hypothesis we find that the measure μ should describe the folded protein, or at least its large-time state. Whether or not this state represents a folded or denatured protein depends on the thermodynamic conditions.

The arguments given above depend on a number of assumptions. The fact of thermodynamic relaxation is very generally true, but it is tricky to set things up so that this can be given a specific interpretation in terms of protein folding. We intend to continue the study of the assumptions about choosing E and their thermodynamic meaning (i.e. the corresponding temperature and pressure, etc.), and to write a careful account of the above argument.

The most important theoretical questions concern the *rate* of relaxation (which is usually interpreted as a statement about the transition state of the folding reaction) and the degree of *localization* in conformation space of the large-time state μ. In the first case perhaps a rigorous result could be proved that this rate can be no faster than the time scale required for the diffusion of the protein atoms into their folded positions [HH]. Of course the real issue is not why proteins do not fold faster, but why they are capable of folding as fast as they do. This seems to be a property of their select amino acid sequences, since a typical random sequence protein is expected to fold very slowly, much too slowly to be of any biological use. It is unlikely that this question can be addressed at this level of modelling. In regard to localization, it should be noted that smaller polypeptides (with less than 25 amino acids) do not seem to achieve a highly defined folded conformation. One wonders if this phenomenon can be traced to the fact that such smaller sequences do not contain enough *information* to specify a particular conformation. However, longer sequences do not necessarily lead to localized folded states; again this seems to be a property of selected sequences. However, mathematical issues clarifying the definition of localization still need to be resolved, and we propose to do that first. (See the next section.)

Stochastic Process. Until now we have been treating protein folding as a phenomenon of a composite system of protein and solvent. However it is more or less clear that most of the solvent molecules behave almost as if the protein were not there, and so keeping track of all the degrees of freedom of these molecules is probably unnecessary. We intend to replace the "deterministic" model of the previous section with a probabilistic model for the motions of just the non-hydrogen protein atoms.

Recall the phase space for the complete system of particles is $\Phi_{tot} = (\mathbb{R}^3)^{N+1} \times (\mathbb{R}^3)^{N+1}$, and $G = \mathbb{R}^3 \times \mathrm{SO}(3)$ acts on it by the rule $g(\mathbf{r}_0, \ldots, \mathbf{r}_N, \mathbf{p}_0, \ldots, \mathbf{p}_N) = (\mathbf{b} + R\mathbf{r}_0, \ldots, \mathbf{b} + R\mathbf{r}_N, R\mathbf{p}_0, \ldots, R\mathbf{p}_N)$, where $g = (\mathbf{b}, R) \in G$. If particles 1 through M represent the non-hydrogen atoms of the protein, then we obtain a projection map $\pi\colon \Phi_{tot} \to (\mathbb{R}^3)^M \times (\mathbb{R}^3)^M$, which is a left G-map. Let $\tilde{\pi}\colon G\backslash\Phi_{tot} \to G\backslash[(\mathbb{R}^3)^M \times (\mathbb{R}^3)^M]$ be the associated map on

orbits. Since $\Phi_E \subset G\backslash\Phi_{tot}$, we define $\Phi_{E,p}$ to be the image under $\tilde{\pi}$ of Φ_E. Restricting $\tilde{\pi}$ we get a fibration $\tilde{\pi}\colon \Phi_E \to \Phi_{E,p}$ of the probability space (Φ_E, μ). Defining $\mu_p(A) = \mu(\tilde{\pi}^{-1}(A))$ whenever $A \subset \Phi_{E,p}$ and $\tilde{\pi}^{-1}(A)$ is a Borel subset of Φ_E, we obtain a probability space $(\Phi_{E,p}, \mu_p)$. By a theorem on the fibration of probability measures [M] we also obtain probability measures ν_x on $\tilde{\pi}^{-1}(\{x\})$ for all $x \in \Phi_{E,p}$ such that for every μ-integrable function f on Φ_E we have $\int_{\Phi_E} f\, d\mu = \int_{\Phi_{E,p}} \int_{\tilde{\pi}^{-1}(\{x\})} f\, d\nu_x\, d\mu_p(x)$. If $y \in \Phi_E$ and $\tilde{\pi}(y) = x$ then the trajectory $\phi_t(y)$ gets projected to a path $\tilde{\pi}(\phi_t(y))$ which coincides with x when $t = 0$. This path is not determined by x, but depends on the random element $y \in \tilde{\pi}^{-1}(\{x\})$. If $t_1 < \cdots < t_n$ are real numbers and $A_1, \ldots, A_n \subset \Phi_{E,p}$ are measurable subsets, then the probability that the path $\tilde{\pi}(\phi_t(y))$ will pass through the set A_j at time t_j, $j = 1, \ldots, n$, is $\mu(\bigcap_{j=1}^n \phi_{-t_j}(\tilde{\pi}^{-1}(A_j)))$. One then shows that this defines a probability measure on the set of continuous paths in $\Phi_{E,p}$. This will then define the *stochastic process* of protein dynamics. We propose to study this process in detail. In particular after its existence is rigorously established we are interested in approximating it by stochastic processes of a simpler sort, namely Markov processes or systems of stochastic differential equations. Although we do not know how to compute the Born-Oppenheimer potential very well, perhaps the ability to approximate the process of protein dynamics by simpler processes does not depend on the fine details of the potential. Most practical models of protein folding involve some sort of approximate stochastic process, but very little work has been done on the exact sorts of approximations that are involved. We would like to clarify this aspect.

Besides the removal of the solvent, which we outlined above, there are several other reductions which are important. Usually the momentum variables of the protein atoms are also integrated out, leaving a process in the protein conformational variables. Also, many of the conformational variables are executing vibrations (e.g. bond lengths and bond angles), hence one would like to integrate out these variables, leaving only the free dihedral angles (as were discussed in the section on low energy molecules). It is in these variables that the phenomenon of localization of the folded state is most striking. Even if one starts with a uniform distribution in the free dihedral variables, after sufficient time has elapsed this distribution evolves into one which is sharply peaked about a particular choice of each free dihedral angle, i.e. the folded state. The rate of relaxation and the degree of localization of the long-time state can be studied more productively at the level of stochastic models, and this is something we desire to pursue. However, we first need to study each of the above reductions carefully, so that approximate models can be chosen more rationally.

The fundamental assumption underlying the energy landscape models of protein folding is that a rapid rate of folding and a localized folded state are consequences not so much of special properties of the Born-Oppenheimer potential but of any potential yielding an funnel-like free energy landscape. Of course this raises many mathematical questions. Proof that rapid folding and localization are consequences of hypotheses imposing funnel-like behavior on the potential needs to be given. There is also the question of how the nature of the Born-Oppenheimer potential allows an adequate number of sequences with funnel-like landscapes to exist. But first we need to give a rigorous definition of the free

energy landscape so that these problems can become mathematically well-posed. This is our goal for the present time.