52nd IEEE Conference on Decision and Control
December 10-13, 2013. Florence, Italy

Flexible Refinement of Protein-Ligand Docking on Manifolds*

Hanieh Mirzaei†, Elizabeth Villar◇, Scott Mottarella†, Dmitri Beglov‡, Ioannis Ch. Paschalidis§, Sandor Vajda‡, Dima Kozakov‡, Pirooz Vakili¶

Abstract— **Our work is motivated by energy minimization of biological macromolecules, an essential step in computational docking. By allowing some ligand flexibility, we generalize a recently introduced novel representation of rigid body minimization as an optimization on the $SO(3)\times\mathbb{R}^3$ manifold, rather than on the commonly used Special Euclidean group $SE(3)$. We show that the resulting flexible docking can also be formulated as an optimization on a Lie group that is the direct product of simpler Lie groups for which geodesics and exponential maps can be easily obtained. Our computational results for a local optimization algorithm developed based on this formulation show that it is about an order of magnitude faster than the state-of-the-art local minimization algorithms for computational protein-small molecule docking.**

I. INTRODUCTION

Predictive molecular docking is a fundamental and challenging problem in computational structural biology. Given two component molecules, termed *receptor* and *ligand*, the goal is to determine the most likely structure of a receptor-ligand complex by minimizing an energy-like target function.

The challenge for predictive docking is to start with the coordinates of the unbound component molecules and to obtain computationally a model of the bound complex [1], [2], [3]. All successful state-of-the-art docking methods employ a so-called *multistage approach* that include some type of *local continuous minimization* of the energy function in order to remove steric clashes and obtain more reliable energy values [3]. Local optimization is used to identify, for each initial configuration, a minimum energy configuration in its vicinity; the minimum energy configuration is then used in an exterior loop of a global optimization algorithm in a refinement stage. As a result, the refinement stage involves repeated use of local minimization and deriving more efficient algorithms for this purpose has a direct impact on the overall efficiency of docking protocols.

In this paper, we generalize the manifold (local) optimization approach introduced in [4], [5] by allowing some molecular flexibility while following the same basic optimization approach.

A. Manifold optimization approach

Receptor and ligand molecules consist of atoms held together by covalent chemical bonds. One possible formulation of docking as an optimization problem, to which we will refer to as *full-atomic optimization*, is to assume all atoms can move freely and rely on energy minimization to enforce partial rigidities that exist or are assumed in each molecule. In this case, the conformation space of docking, or, equivalently, the optimization search space, is simply the $3k$-dimensional Euclidean space $\mathbb{R}^{3k}$ where k is the total number of atoms of the complex.

In contrast to the full-atomic optimization approach, one can explicitly take all assumed rigidities into account when defining the conformation/search space. For example, if both receptor and ligand are assumed to be rigid, then the conformation space is the space of movements of the ligand relative to the receptor, namely, the 6-dimensional space of rigid body motion. This search space can be endowed with the structure of a Riemannian manifold and the energy minimization can be formulated as a *manifold optimization*.

The advantage of a full-atomic formulation is that the search space is always a Euclidean space with a well-known geometry for which various efficient and well-understood optimization algorithms are available. On the other hand, its drawback is that the search/conformation space is often of a very high dimension leading to slow convergence of optimization algorithms. By formulating the energy minimization problem as an optimization problem on manifolds we can often arrive at a search space with the smallest possible dimension. However, the geometry of the resulting manifold may present challenges for optimization (see, e.g., [6]) and, more generally, the application of manifold optimization algorithms are more difficult than their Euclidean counterparts (see, e.g., [7], [8]).

In [4], [5] we presented a manifold optimization approach to rigid body minimization with application to rigid molecular docking. Instead of the common Special Euclidean Group $SE(3)$, we introduced an alternative group of rigid body transformations that corresponds to the Lie group $SO(3)\times\mathbb{R}^3$. We showed that the new formulation avoids the difficulties associated with optimization on $SE(3)$, better matches the moves of the optimization with the dynamics of molecular interactions, and provides an additional flexibility, namely choosing the initial center of rotation, that can be used to improve the performance of the optimization algorithm. We showed that the resulting algorithm substantially outperforms state-of-the-art local minimization algorithms

*Research partially supported by the NIH/NIGMS under grants GM093147 and GM061867, by the NSF under grants CNS-1239021 and IIS-1237022, by the ARO under grants W911NF-11-1-0227 and W911NF-12-1-0390, and by the ONR under grant N00014-10-1-0952.

† Division of Systems Eng., Boston University, hanieh@bu.edu

‡ D. Kozakov, D. Beglov, and S. Vajda are with the Dept. of Biomedical Eng., Boston University, {midas, dbeglov, vajda}@bu.edu

§ Dept. of Electrical & Computer Eng., and Division of Systems Eng., Boston University, 8 Mary's St., Boston, MA 02215, yannisp@bu.edu

¶ Corresponding author. Dept. of Mechanical Eng. and Division of Systems Eng., Boston University, vakili@bu.edu

◇ Dept, of Chemistry, Boston University, eavillar@bu.edu

† Dept, of Bioinformatics, Boston University, semottar@bu.edu

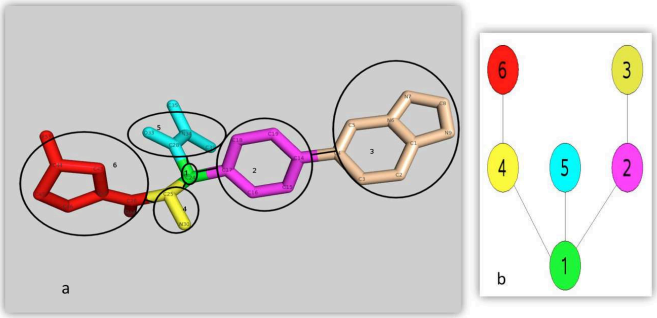

Fig. 1. a) The atoms of the ligand of 2FJP complex decomposed into rigid clusters, each cluster is colored differently and we assign a number to each cluster. b) The tree structure of the corresponding rigid decomposition. Two rigid clusters are connected if there is a covalent bond between the corresponding clusters in the molecule.

used for rigid docking.

In this paper, motivated by protein-small molecule docking problems, we extend our manifold optimization approach by allowing some internal ligand flexibility while assuming the receptor remains rigid. We show that the corresponding conformation space is a Riemannian manifold and a Lie group. As a result, an optimization approach similar to that presented in [4], [5] can be applied. Our computational results show that the algorithm is about an order of magnitude faster than a full-atomic optimization algorithm.

The rest of the paper is organized as follows. We describe ligand flexibility and its representation in Section II. In Section III we show that the conformation space can be considered as a Lie group. Our local optimization algorithm and the experimental results are provided, respectively, in Sections IV and V. We conclude in Section VI.

II. Ligand Flexibility & Representation

We assume that the ligand is composed of a set of rigid clusters, such as methylene group or phenyl ring, held together by covalent bonds (see Figure 1). Within each cluster the relative position of the atoms does not change. As is common in the literature, we refer to the bonds between clusters as *hinges* (see, e.g., [9] [10]).

In general, freely rotating and translating clusters will have six degrees of freedom; however, bond lengths and bond angles of small-molecules do not change significantly upon binding and it is reasonable to fix them during the minimization step of docking (see, e.g., [11], [12]). By fixing bond lengths and bond angles, the only internal flexibilities of the ligand are torsional moves along rotatable bonds. A bond for which only changes in the torsional bond angle are permitted is modeled as a one degree of freedom rotational hinge.

Following [13], [9], [10], we use a *torsion tree* to represent the rigid and rotatable parts of the ligand.

A. Tree Topology Model

We form a topology graph $G=(V,E)$ of the ligand such that each node of the graph corresponds to a rigid cluster of the molecule. Two nodes are connected by an edge if and only if there is a rotatable covalent bond between the corresponding rigid clusters in the molecule. We assume that the resulting graph does not have any cycles and is a connected graph, therefore, a tree. The program AutoDock [13] outputs such graphs for input ligands.

We select one particular node/cluster in the tree as the *root* cluster. Once the root cluster is selected, the parent of each node in the tree is uniquely and completely determined. For example, in Figure 1, cluster 1 is chosen as the root cluster; cluster 1 is the parent of clusters 4 and 5 and 2, cluster 4 is the parent of cluster 6 and cluster 2 is the parent of cluster 3.

Each hinge is between a pair of parent-child clusters and connects an atom in the parent cluster to an atom in the child cluster. We assign a coordinate frame to each hinge which is parallel to the reference coordinate frame. The center of the coordinate frame is the end atom of the hinge in the corresponding child cluster. The motion of the hinges is characterized by the motion of these frames with respect to the reference coordinate frame.

B. General equation for the ligand displacement

We use the following notation:

- R denotes the rotation and t denote the translation of the whole ligand.
- O denotes the center of mass of the ligand.
- We denote atoms in cluster A having m_A atoms as $(A,0),\cdots,(A,m_A-1)$ and their Cartesian coordinates with respect to a fixed reference frame as $q_{A,0},\cdots,q_{A,m_A-1}$.
- Let hinge k be the hinge between parent cluster A and child cluster B. Assume hinge k is between atoms $q_{A,a_k}\in A$ and $q_{B,0}\in B$.
  - u_k is the unit vector in the direction $q_{B,0}-q_{A,a_k}$.
  - θ_k is the torsion angle along hinge k in the direction u_k.
  - R_k is the rotation matrix corresponding to θ_k torsion along hinge k in the direction u_k.
  - $O_k=q_{B,0}$ is the center of reference coordinate frame corresponding to hinge k.

The torsion along hinge k would move the atoms in cluster A, if and only if hinge k appears in the path from cluster A to the root cluster. To find the position of atom $q_{A,i}\in A$ after rotation along hinges, we need to first find the path from cluster A to the root cluster. Let $P=\{h_p,..,h_0\}$ be the hinges on the path from cluster A to the root cluster. We can apply hinge rotations in different order; for example, we can apply them in the order that they appear on path P or in the reverse order of their position on path P. It is more advantageous to apply the hinge rotations in the order that they appear on path P as explained below (also, see, e.g., [9]).

According to axis-angle parametrization, we have

$$R_{h_i}=e^{\theta_{h_i}u_{h_i}}. \qquad (1)$$

Assume $q_{A,i}$ denotes the Euclidean coordinates, with respect to the fixed reference frame, of atom i in cluster A; then after rotation with rotation matrix R_{h_i}, we have

$$q'_{A,i} = R_{h_i}(q_{A,i} - O_{h_i}) + O_{h_i},$$

where $q'_{A,i}$ is the new position of the atom i in cluster A and O_{h_i}, as mentioned before, is the center of the coordinate frame assigned to hinge h_i.

If we apply the torsion rotations on the path from cluster A to the root cluster, then the new position of atom i in cluster A would be:

$$q'_{A,i} = R_{h_0}(...(R_{h_p}(q_{A,i} - O_{h_p}) + O_{h_p}) + ... - O_{h_0}) + O_{h_0}.$$

The closed form formula for $q'_{A,i}$, which can be easily verified by induction on the length of the path, is as follows.

$$q'_{A,i} = \Pi_{j=0}^{p} R_{h_j}(q_{A,i} - O_{h_0}) + \Sigma_{k=0}^{p-1}\Pi_{j=0}^{k} R_{h_j}(O_{h_{k+1}} - O_k) + O_{h_0}$$

After applying the internal motions of the ligand, we also need to rotate and translate the ligand as a rigid body. We use our formulation of rigid body motion introduced in [4], [5], and briefly described in the next section, for this purpose. In this formulation of rigid body motion, we have the freedom of choosing the center of rotation. We select the center of rotation as the center of mass of the ligand. After rigid body displacement, the position of the atom $q_{A,i}$ would be

$$R(q'_{A,i} - O) + O + t.$$

For local minimization of the energy function, we need to calculate the gradient of the Cartesian coordinates of atoms with respect to torsional rotation parameters. By selecting to move on the path from the cluster towards the root, the energy functions will have a simpler form and the gradients can be more easily computed.

III. THE CONFORMATION SPACE: $SO(3) \times \mathbb{R}^3 \times T^d$

We begin by a brief review of our proposed representation of rigid-body conformation space. For more details, see [5].

A. *Rigid motion space:* $SO(3) \times \mathbb{R}^3$

Let

$$SO(3) = \{R \in \mathbb{R}^{3\times3}; R^T R = I; det(R) = 1\}$$

denote the group of orientation-preserving rotations on $\mathbb{R}^3$.

The direct product "multiplication" on $SO(3) \times \mathbb{R}^3$ is naturally defined by

$$g' * g = (R'R, t' + t).$$

We use $*$ to denote the direct product multiplication. The novel element of our representation, compared to $SE(3)$ is the action we associate with this group. We define this action on $\mathbb{R}^3 \times \mathbb{R}^3$ as follows. For $g \in SO(3) \times \mathbb{R}^3$, let $g : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3 \times \mathbb{R}^3$ be defined by

$$g(q,p) = (R(q-p) + p + t, p + t),$$

where $q,p,t \in \mathbb{R}^3$, a d $R \in SO(3)$.

In words, the action of g on the first component $q \in \mathbb{R}^3$ is to rotate q according to the rotation matrix R but with the center of rotation p and translate it by t. The action of g on the second component simply translates the point p by t. Equivalently, we can think that the action on the second component is of the same type as the action on the first component since $R(p-p) + p + t = p + t$. The following is an immediate result.

Proposition 1: The above transformation defines an action of the group $SO(3) \times \mathbb{R}^3$ on $\mathbb{R}^3 \times \mathbb{R}^3$.

For any center of rotation $p \in \mathbb{R}^3$, the action of $SO(3) \times \mathbb{R}^3$ on $\mathbb{R}^3 \times \mathbb{R}^3$ is a rigid body transformation of the first component $\mathbb{R}^3$.

Let $\pi_i : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ $(i = 1,2)$ be projections on the first and second coordinate ($\pi_1(q,p) = q$, $\pi_2(q,p) = p$). For any fixed $p \in \mathbb{R}^3$, let

$$g_p : \mathbb{R}^3 \to \mathbb{R}^3 \times \mathbb{R}^3$$

be defined by $g_p(q) = g(q,p)$. Then, we have the following proposition.

Proposition 2: For any p,

$$\pi_1 \circ g_p : \mathbb{R}^3 \to \mathbb{R}^3$$

is a rigid body transformation of $\mathbb{R}^3$.

The rigid body motion that we associate with $SO(3) \times \mathbb{R}^3$ is a natural motion in the context of the molecular docking problem that we consider next. Furthermore, since the group $SO(3) \times \mathbb{R}^3$ is a direct product of $SO(3)$ and $\mathbb{R}^3$ both as groups and as Riemannian manifolds, there is no mismatch between the group and the natural Riemannian structures and we do not face the complications that are associated with $SE(3)$ rigid body transformations.

Furthermore, as we mentioned earlier, in the $SO(3) \times \mathbb{R}^3$ formulation the user can choose the initial center of rotation. This gives a valuable flexibility to the user to better match the moves of the optimization with the dynamics of molecular interactions.

B. *Internal ligand motion space:* T^d

Assume the ligand has d rotatable bonds/hinges and let $\theta_i \in \mathbb{R}$ be the rotational parameter associated with the ith hinge. Then $\theta = (\theta_1, \cdots, \theta_d) \in \mathbb{R}^d$ represents an internal ligand motion.

$\mathbb{R}^d$ with the addition operation is a trivial Lie group. Given this group structure, $C = \{2\pi a;\ a \in \mathbb{Z}^d\}$ is a normal subgroup of $\mathbb{R}^d$. Define an equivalence relation on $\mathbb{R}^d$ by letting

$$\theta \sim \theta' \Leftrightarrow \theta - \theta' \in C.$$

The quotient space of $\mathbb{R}^d$ with respect to the equivalence relation $\sim$ (or subgroup C) is a homogeneous space and a Lie group isomorphic to the d-dimensional *torus*

$$T^d = \underbrace{S^1 \times S^1 .. \times S^1}_{\text{d times}}.$$

See, e.g., [14], section 3.2.

C. Conformation space: Lie Group $SO(3) \times \mathbb{R}^3 \times T^d$

Given the above associations, it is easy to see that the conformation space of the ligand movement including rigid and internal motions is given by the direct product of $SO(3) \times \mathbb{R}^3$ and T^d both as Riemannian manifolds and as groups.

The direct product group operation on $SO(3) \times \mathbb{R}^3 \times T^d$ is naturally defined by

$$(R',t',[\theta']) * (R,t,[\theta]) = (R'R, t'+t, [\theta']+[\theta])$$

where $[\theta]$ represents the equivalence class of θ.

Therefore, as in the case of rigid body optimization considered in [5], the search space of the optimization is a Lie group whose geodesics and exponential coordinates are easy to compute. Therefore, the basic local optimization approach of [5] can be adopted in this case as well.

IV. LOCAL OPTIMIZATION ALGORITHM

We begin with defining more formally the energy function and the optimization problem.

A. The optimization problem

Assume the tree structure of the ligand is given and the initial center of rotation for the rigid motion of the ligand p is selected. Let $\tilde{g} = (R,t,[\theta]) \in SO(3) \times \mathbb{R}^3 \times T^d$. Then, transformation of $\mathbb{R}^3$ given by $\pi_1 \circ \tilde{g}$ extends naturally to a transformation of the ligand, namely

$$\tilde{g} : \mathbb{R}^{3\times m_l} \to \mathbb{R}^{3\times m_l}.$$

Let $\mathbf{q} = (q_1, \cdots, q_{m_l}) \in \mathbb{R}^{3\times m_l}$ denote a position of the ligand and $\mathbf{q}' = (q_1', \cdots, q_{m_l}') \in \mathbb{R}^{3\times m_l} = g(\mathbf{q})$. Then, $q_i' \in \mathbb{R}^3$ is defined as

$$q_i' = \pi_1 \circ \tilde{g}(q_i).$$

Therefore, the local optimization problem can be defined as the following optimization on $SO(3) \times \mathbb{R}^3 \times T^d$

$$\min_{\tilde{g}} E(\tilde{g}(\mathbf{q})) \quad \tilde{g} \in SO(3) \times \mathbb{R}^3 \times T^d.$$

B. The optimization approach

Our approach to solving the above optimization problem is driven by practical considerations and follows one of the standard options available for manifold optimization, namely using a local parametrization of the manifold. This parametrization is given by the exponential map on the tangent space of $SO(3) \times \mathbb{R}^3 \times T^d$ at $(I, \mathbf{0}, \mathbf{0})$. It is well-known that using local coordinates introduces nonlinearities that can degrade the performance of optimization algorithms. Hence, there is a tradeoff in using local parameterizations (see, e.g., [6]). Our motivation for using a local parametrization is to be able to take full advantage of well-developed and optimized Euclidean optimization algorithms. Our evaluation of the tradeoff is driven by a very practical concern; namely, we aim to develop an algorithm of equal or higher quality that is more efficient than current algorithms used in protein-small molecule docking. Our experimental results, given in Section V show that our approach has been successful.

One can also justifiably argue that the local parametrization based on the exponential map is particularly well-suited for the local optimization we consider. The fact that our rigid body movement is defined on $SO(3) \times \mathbb{R}^3$ rather than $SE(3)$ is another argument in favor of the suitability of exponential parametrization, as pointed out in [5].

C. Local parametrization of $SO(3) \times \mathbb{R}^3 \times T^N$

The following properties hold for $SO(3) \times \mathbb{R}^3 \times T^d$. (i) The product manifold inherits its Riemannian metric in a natural way from its component factor manifolds; (ii) the geodesics of the product manifold is the product of geodesics of the factor manifolds; and, (iii) the exponential map on the product map is the product of the exponential map on the factor manifolds. (For a brief review of product Riemannian manifolds and optimization on them, see, e.g., [15], Appendix A). $\mathbb{R}^3$ is a trivial manifold and the exponential map on T^d is the product of exponential maps defined for each coordinate as in equation (1). Therefore, we briefly review the $SO(3)$ manifold.

Geodesics of $SO(3)$ are given by $R(u) = R_0 e^{[\omega]u}$, $\omega \in \mathbb{R}^3$ and $u \in \mathbb{R}$, they are one-parameter subgroups of $SO(3)$, and they correspond to the projection by the exponential map of lines going through the origin on the tangent space.

The exponential map of $SO(3) \times \mathbb{R}^3 \times T^d$ can be easily obtained from that of $SO(3)$. Consider the exponential map at the identity of the product Lie group $SO(3) \times \mathbb{R}^3 \times T^d$, i.e., $(I,0,0)$. The tangent space can be identified with $\mathbb{R}^{(6+d)}$. Let $(\omega, \upsilon, \theta) \in \mathbb{R}^{(6+d)}$ be a point in the tangent space. Then,

$$\exp_{(I,\mathbf{0},\mathbf{0})}(\omega, \upsilon, \theta) = (e^{[\omega]}, \upsilon, [\theta]).$$

Therefore,

$$\exp_{(I,\mathbf{0},\mathbf{0})} : \mathbb{R}^{(6+d)} \to SO(3) \times \mathbb{R}^3 \times T^d$$

defines a local parametrization for $SO(3) \times \mathbb{R}^3 \times T^d$ in the neighborhood of $(I, \mathbf{0}, \mathbf{0})$.

D. The Optimization algorithm

Given the exponential map parametrization, the minimization is defined on the $(6+d)$-dimensional Euclidean space $\mathbb{R}^{(6+d)}$. From among the many deterministic algorithms available to solve local minimization problems on a Euclidean space, we have selected the quasi-Newton method of Limited memory BFGS (LBFGS) [16]. In our parametrization, the gradient and the Hessian of the energy function with respect to the parameters of optimization can be explicitly calculated. However, these are costly operations, evaluating the Hessian being significantly more costly than evaluating the gradient. Our choice of LBFGS is based on the fact that it uses only gradient information to obtain second order information about the energy function.

Denoting the elements of $\mathbb{R}^{(6+d)}$ by x, the LBFGS method consists of the following iterations[16]

$$x_{k+1} = x_k + \alpha_k d_k, \tag{2}$$

where

$$d_k = -H_k \nabla E_k, \tag{3}$$

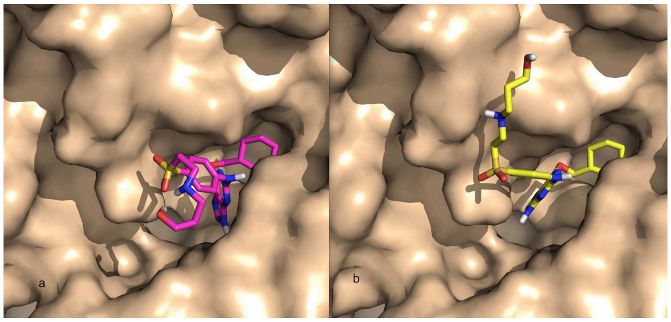

Fig. 2. a) 2G9X complex before minimization b) 2G9X complex after minimization

where ∇E_k is the gradient of the energy function, H_k is the LBFGS approximation of the inverse of the Hessian of the energy function as described in [16], and α_k is an appropriately selected step-length as described in [17].

As pointed out in [16], the choice of H_0 influences the behavior of the algorithm. When the diagonal entries of the Hessian are all positive, it is recommended to let H_0 be a diagonal matrix with the diagonal entries of the inverse of the Hessian. Given that in our problem the diagonal entries of the Hessian are sometimes negative, we use the identity matrix as the initial H_0.

Figure 2 shows the position of a ligand for the receptor-ligand complex 2G9X before and after minimization.

V. EXPERIMENTAL RESULTS

In this section we describe the experimental setup and results from the application of the proposed manifold optimization algorithm to protein-small molecule docking. We compare the performance of the algorithm with full-atomic optimization, the commonly used algorithm for protein-small molecule docking. The results show that the quality of solutions produced by the manifold optimization algorithm is better than full-atomic optimization and its computational efficiency is significantly superior to it.

A. *Application to Protein Mapping*

We apply the proposed manifold optimization algorithm to protein-small molecule docking as a complement to our protein mapping program FTMap [18]. Mapping places molecular probes, small organic molecules that vary in size and shape, on a dense grid around the protein to identify potentially favorable binding positions. For each probe type, the first step of FTMap is a global sampling of the 6D space using the FFT correlation approach. In the current version, the docked structures generated by this calculation are minimized off-grid using the CHARMM [19] potential and full-atomic minimization.

Out of 16 probes considered in FTMAP, 10 probes have no rotatable bonds and are fully rigid, whereas the other 6 are flexible. The result of the comparison of our rigid-body manifold optimization algorithm and *full-atomic optimization* for the rigid probes is reported in [4]. Here, we compare our flexible manifold optimization algorithm with full-atomic minimization based on the 6 flexible probes.

We compare the two minimization algorithms based on the quality of their solutions and their computational efficiency. The cases where the local minima found by the two algorithms are within 0.05 Å RMSD distance of each other or their energy differences are less than 0.01 $\frac{kcal}{mol}$ are considered ties. In all other cases, the quality of the solution of one algorithm relative to the other is considered superior if it has a lower energy. As for the measure of computational efficiency of each algorithm, we have selected the number of energy function evaluations needed to converge to a local minimum. Given that energy function evaluations are the most costly operations, the number of energy evaluations is used as a measure of run time efficiency of the algorithm. Furthermore, since the same energy function is used for both algorithms, the number of energy function evaluations is a fair comparison between runtime of the two algorithms.

As the convergence rate of full-atomic minimization is low, we stop the algorithm after 500 energy function evaluations. Also, to obtain a more reliable energy value for flexible manifold optimization, we relax the bond lengths and bond angles of the conformations after the manifold minimization by performing 20 steps of full-atomic minimization.

To compare the two algorithms 14 protein structures, shown in I, were selected from the Protein Data Bank (PDB) [20]. All ligand and bound water molecules were removed prior to mapping. For each target FTMap performs a grid search using the Fast Fourier Transform (FFT) correlation approach to find the low energy docked positions of the flexible probes. Each complex is evaluated using an energy expression that includes van der Waals and electrostatic interaction energy terms as well as solvation effects. In the current version of FTMap, the 2000 most favorable docked positions of each probe are then energy-minimized using the CHARMM force field and a full-atomic minimization. During this minimization the probe molecules are considered fully flexible, but the atoms of the receptor protein are taken as fixed.

The results for comparison of the two algorithms based on six flexible probes are reported in Table I. Complexes are identified by their 4-letter PDB code[20] in the first column of the table. The second column gives the number of ties between the two algorithm. The third column is the number of conformations in which manifold optimization (denoted by MO) converged to a local minimum with lower energy. The forth column is the number of conformations in which full atomic minimization (denoted by FA) produced better result. The fifth column is the average number of energy function evaluation in manifold optimization and the last column is the average number of energy function evaluations by the full atomic minimization. As can be seen, the quality of the solutions of the manifold optimization is better than full-atomic optimization, while the manifold optimization algorithm is 5.1 times faster than the full-atomic minimization algorithm.

TABLE I

COMPARISON OF THE QUALITY OF SOLUTIONS & COMPUTATIONAL EFFICIENCY OF MANIFOLD OPTIMIZATION (MO) WITH FULL-ATOMIC MINIMIZATION (FA) FOR FLEXIBLE PROBES

	Quality of solutions: Which performs better			Computational efficiency: Average no. of steps	
Complex	*FA* = *MO*	*FA* < *MO*	*MO* < *FA*	MO	FA
2CAB	5683	3691	1874	81.6	451.3
1IVG	5752	3319	2135	86.3	423.0
1BBC	4929	4033	1855	67.1	453.5
1O8A	5428	2103	2088	93.0	421.3
1F5L	5926	3195	2066	78.5	439.8
1S3E	5392	2013	1932	93.6	424.8
2O8T	5866	2951	2239	77.5	433.8
1W50	6003	2880	2685	94.3	420.3
1J2E	4833	1775	2001	93.3	388.5
1YES	5281	3797	1914	82.0	438.1
1HCL	5594	2955	1711	83.5	428.6
1THS	5810	3009	1896	86.6	420.3
1BN5	6295	2952	1849	77.3	418.5
1PUD	6371	3080	2137	71.5	427.3
	53.02%	**27.96%**	**19.02%**	**5.1**	**1**

VI. CONCLUSIONS

In this paper, by allowing some ligand flexibility, we generalize a new formulation for rigid body minimization that is based on a new group of rigid body transformations. We show that the resulting flexible docking can also be formulated as an optimization on a Lie group that is the direct product of simpler Lie groups for which geodesics and exponential maps can be easily obtained. Therefore, a local manifold optimization algorithm similar to one developed for rigid docking is proposed. Experimental results provided in the paper show that our algorithm is substantially more efficient than full-atomic minimization used for protein-small molecule docking.

An optimization algorithm that evolves directly on the manifold, rather than a local parametrization of it, may have advantages to the one presented in this paper and exploring whether it does is a subject of our future research.

REFERENCES

[1] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov, "Principles of docking: An overview of search algorithms and a guide to scoring functions." *Proteins*, vol. 47, pp. 409–443, 2002.

[2] G. Smith and M. Sternberg, "Prediction of protein-protein interactions by docking methods." *Curr Opin Struct Biol*, vol. 12, pp. 28–35, 2002.

[3] S. Vajda and D. Kozakov, "Convergence and combination of methods in protein-protein docking," *Curr. Opin. Struct. Biol.*, vol. 19, pp. 164–170, Apr 2009.

[4] H. Mirzaei, D. Beglov, I. C. Paschalidis, S. Vajda, P. Vakili, and D. Kozakov, "Rigid body energy minimization on manifolds for molecular docking," *Journal of Chemical Theory and Computation*, vol. 8, no. 11, pp. 4374–4380, 2012.

[5] H. Mirzaei, D. Kozakov, D. Beglov, I. C. Paschalidis, S. Vajda, and P. Vakili, "A new approach to rigid body minimization with application to molecular docking," in *51th IEEE Conference on Decision and Control*, 2012, pp. 2983–2988.

[6] S. Gwak, J. Kim, and F. C. Park, "Numerical optimization on the Euclidean group with applications to camera calibration," *IEEE Trans. on Robotics and Automation*, vol. 19, no. 1, pp. 65–74, Feb. 2003.

[7] S. T. Smith, "Optimization techniques on Riemannian manifolds," in *Proc. Fields Inst. Workshop on Hamiltonian and Gradient Flows, Algorithms, and Control*, 1994.

[8] P. A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ: Princeton University Press, 2008.

[9] A. Jain, N. Vaidehi, and G. Rodriguez, "A fast recursive algorithm for molecular dynamics simulation," *Journal of Computational Physics*, vol. 106, no. 2, pp. 258–268, 1993.

[10] C. D. Schwieters and G. Clore, "Internal coordinates for molecular dynamics and minimization in structure determination and refinement," *Journal of Magnetic Resonance*, vol. 152, no. 2, pp. 288–302, 2001.

[11] A. MacKerel Jr., C. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus, *CHARMM: The Energy Function and Its Parameterization with an Overview of the Program*, ser. The Encyclopedia of Computational Chemistry.

[12] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker, "Rosetta Ligand: Protein small molecule docking with full side-chain flexibility," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, pp. 538–548, 2006.

[13] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, and D. S. G. A. J. Olson, "Autodock4 and autodocktools4: Automated docking with selective receptor flexiblity," *J. Computational Chemistry*, vol. 16, p. 2785, 2009.

[14] J. M. Selig, *Geometric Fundamentals of Robotics*. Springer, 2005.

[15] Y. Ma, J. Koseck, and S. Sastry, "Optimization criteria and geometric algorithms for motion and structure estimation," *International Journal of Computer Vision*, vol. 44, pp. 219–249, 2001.

[16] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, pp. 503–528, 1989.

[17] D. Xie and T. Schlick, "A more lenient stopping rule for line search algorithms," *Optimization Methods and Software*, vol. 17, pp. 683–700, 2002.

[18] R. Brenke, D. Kozakov, G. Y. Chuang, D. Beglov, D. Hall, M. R. Landon, C. Mattos, and S. Vajda, "Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques," *Bioinformatics*, vol. 25, pp. 621–627, Mar 2009.

[19] B. R. Brooks, R. E. Bruccoleri, D. J. Olafson, D. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, pp. 187–217, 1983.

[20] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Research*, vol. 28, p. 235242, 2000.