ArtSurf: A Method for Deformable Partial Matching of Protein Small-Molecule Binding Sites

Jeffrey R. Van Voorst
Dept. of Medicinal Chemistry
University of Minnesota–Twin Cities
vanv0059@umn.edu

Yiying Tong
Dept. of Computer Science and Engineering
Michigan State University
ytong@msu.edu

Leslie A. Kuhn
Dept. of Biochemistry and Molecular Biology
Dept. of Computer Science and Engineering
Michigan State University
kuhnl@msu.edu

ABSTRACT

Proteins and many other biologically relevant molecules are flexible, and the flexibility of a given molecule is one of its important characteristics. In particular, the degree of global and local flexibility of proteins is an important characteristic of protein small-molecule binding sites. In this article, the binding site comparison problem [31, 11, 20, 30, 24, 33] is presented as a deformable partial surface matching problem [7, 13]. The problem is labeled as partial matching because we seek the best match for a given smaller query site (e.g. a small molecule fragment binding site) with respect to each larger binding site (in the search dataset). The question we address is "how can a given *part* of a binding site be realistically deformed to obtain the best partial match between that part and a full binding site"? This goal is addressed by optimizing the similarity between the site representations subject to modeling the proteins as kinematic chains. The preliminary results of our implementation (ArtSurf) show that the proposed approach is feasible, and that the implementation gives physically reasonable results which improve the detection of partial binding site similarity.

Categories and Subject Descriptors

J.2 [**PHYSICAL SCIENCES AND ENGINEERING**]: Chemistry
; J.3 [**LIFE AND MEDICAL SCIENCES**]: Biology and genetics
; I.2.9 [**ARTIFICIAL INTELLIGENCE**]: Robotics—*Kinematics and dynamics*

Keywords

protein structures, binding sites, deformable surface matching, partial matching, inverse kinematics, biophysics, computational biology

ACM-BCB '12, October 7-10, 2012, Orlando, FL, USA

1. INTRODUCTION

A common issue for object recognition methods is that techniques based on rigid body alignments generally fail to recognize as similar the same object before and after that object undergoes a significant deformation. As an example, the limbs of the human body can move large distances relative to the scale of the body. A specific example is that many of the point correspondences found by rigid matching will be incorrect when comparing a person touching her toes to a person with her arms raised overhead. However, if the human body is modeled as a shell (surface) over a stick figure, the joints and connectivity of the human body can be exploited as part of the matching algorithm [25, 10, 12]. Such algorithms can be used to locate and track humans via sensors (stereo cameras, active sensing depth cameras, laser range finders, etc.). These algorithms allow for a system to place avatars in a virtual environment in the same poses as the sensed humans (e.g. Microsoft Kinect and Xbox) and allow for interaction between robots and humans [12].

We seek to assess how deformable partial surface matching can be tuned and applied to the comparison of protein small-molecule binding sites, and use the matching results to provide a more accurate indication of sites' similarities and differences. It is understandable that questions posed for deformable matching of binding sites have details that differ significantly from human face recognition or human pose recognition. Unlike the case of limbs, head, and torso in human pose recognition, most of the amino acids in a binding site are not fully exposed. As a result, it would be very challenging to accurately determine the underlying protein structure based solely on a molecular surface of a binding site. Conversely, given that a binding site is determined by discontinuous parts of the protein's amino acid sequence coming together to form a local interface for molecular interaction, the optimal match of two binding site surfaces would be very difficult to determine by directly comparing side chains when the two proteins have dissimilar amino acid sequences. The primary goal of our work is to address the problem of recognizing binding sites that bind the same small molecule, but are from otherwise unrelated proteins. This implies that in the general case the underlying structure of the deformable site will have no direct atomic correspondences to the structure of the sites to which it is compared. That is, for two small-molecule binding sites to be considered as similar, we only require that the binding site surfaces and chemistry are similar.

Before presenting the algorithm applied to deformable partial binding site matching, some motivation and background on the binding site comparison problem is presented.

2. MOTIVATION AND BACKGROUND

Structural biologists study the interaction of protein structures with other molecules, and they compare and contrast pairs of protein structures using a variety of tools and techniques. At the present, the preferred method is to experimentally determine the relative positions (coordinates) of the molecules' atoms via x-ray crystallography. Given such atomic coordinates, molecular graphics can be used, in conjunction with domain knowledge and experience, to visually compare and contrast aligned pairs of molecular structures. After such analysis, one generally proposes a hypothesis (e.g., that the binding site binds small molecule Z, based on a significant match to another site that is known to bind Z) and tests it using experimental techniques such as biochemical assays, x-ray crystallography, and nuclear magnetic resonance.

Given that there are now more than 80,000 structures in the Protein Data Bank (PDB) [5], a number which increases by about 10 percent a year [26], efficient and accurate computational methods are needed to effectively leverage the information encoded in PDB structures. An area of research that would immediately benefit from more advanced tools to search for local protein similarities is determining and understanding which small molecules partner with a given site on a protein. Because such processes are of vital importance in understanding the effects of pharmaceuticals, it is very beneficial to develop methods that can recognize local similarities between binding sites and bring them to the view of the domain scientists.

Many computational methods have been published that can search part of the structural information in the PDB. Some of the classes of problems addressed are protein structure comparisons [18], protein small-molecule binding site comparisons, and protein-protein binding site comparisons [17]. In the category of protein small-molecule binding site comparison tools, the tools generally compare sets of amino acids directly [21, 4] or features derived from the amino acids based on the categories and relative positions of the proteins' atoms [31, 11, 20, 30, 24, 33].

Many of the previously published binding site comparison tools assume is that it reasonable to compare binding sites as rigid objects of approximately the same size. The rigid binding site assumption reduces the complexity of binding site models and the computational complexity of comparing such models (hereafter termed "comparing binding sites"). However, proteins are not rigid molecules, and there are many examples of proteins requiring the correct flexibility to perform their function(s).

The partial matching problem of protein small-molecule binding sites is interesting for a number of reasons. Pocket mining is an idea that is similar to small-molecule fragment screening in that the goal is to find small-molecule scaffolds that bind to a pocket target protein. However, with pocket mining, the assumption is that by looking for significant partial matches between the target pocket and a structural dataset of small-molecule binding sites (with small molecules bound), one has a greater chance that the small molecule fragment hits actually bind to the target pocket. It is plausible that methods developed for comparing binding sites can be generalized and extended to address partial matching problems in other fields. Also, given the nature and assumptions of partial matching, we expect that our method is better able to address more applications of binding site matching and searching than methods that assume that for two binding sites to be similar they must have approximately the same size and shape (e.g. principal components based alignment methods).

At the present, the authors are unaware of any other computational methods that address binding site deformations for the purpose of comparing binding sites at the atomic or even protein side-chain level of detail. There are some authors and tools that do address protein and/or small-molecule flexibility by decomposing the small molecules into rigid fragments and matching/comparing each small molecule fragment binding site as separate rigid objects [24, 34]. Sael and Kihara have decomposed binding sites into surface patches to account for small molecule and protein flexibility [28]. However, such methods do not directly address protein side-chain flexibility. Also, the problem of the placement and orientation of amino acid side chains has been studied extensively in protein homology modeling [9] and protein-ligand docking [2, 36] However, that work addresses how to best model the structure of a given protein rather than enhance the detection of binding site surface and chemical similarity. In this article we propose, implement, and test a geometric model of binding site deformations at the atomic level, starting from a reasonable alignment between a pair of binding sites, provided by a tool such as SimSite3D [33].

3. ALGORITHM

The deformable partial binding site comparison method is formulated as an optimization problem, and a number of steps are presented to reduce the problem to a form that can be solved by an iterative linear solver. The general form of a constrained optimization problem is an objective function to maximize or minimize subject to zero or more constraints. In this article, the goal is to maximize the similarity of two binding sites, starting with some reasonable initial alignment of the sites, by allowing the model of one site to flex or deform to match the representation of the other site as closely as possible. Any surface or set of features can be deformed to look like another [7], but deformations must be physically reasonable for many applications (e.g. biological systems and human face recognition). Thus, we propose physically reasonable deformations by adding structural constraints based on known favorable geometries in protein structures.

3.1 Rationale

Given two aligned binding sites, our goal is to deform the features of one binding site to better match the features of the other site subject to the physical constraints of the proteins. Our algorithm permits a wide variety of binding sites features provided the features depend on the relative positions and orientations of the binding site atoms. Protein structures are represented as a system of **kinematic chains**, that is, as a set of joints and a set of links between pairs of joints. A **link** is a rigid connection between two joints, and it can be used to represent a covalent bond between two atoms. Each atom is represented by a joint at its atomic center, and a given **joint** models the relative motions (bond-rotational dihedral angle changes) of the protein with

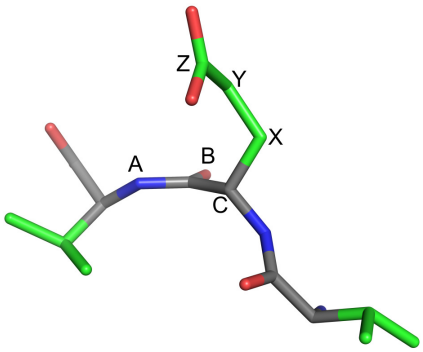

Figure 1: A tube representation of three consecutive amino acids. A, B, and C are backbone nitrogen, oxygen, and alpha carbon atoms. The bond dihedral angle $\angle$ Y is determined by the positions of C, X, Y, and Z. The convention is to hold CXY fixed and to vary the position of Z. In particular, $\angle$ Y is the angle between the vectors CX and YZ when they are projected in a plane whose normal is parallel to the vector XY. In the kinematics model, each atom is a joint and each bond (tube) is a link. An example of a kinematic chain is C-X-Y-Z, and C is a Cα atom. An end effector is a end point of a kinematic chain that is to be moved to a goal. Examples of end effectors are the oxygen (red) atoms at the top of the figure.

respect to that atom and the relative angles of its covalent bonds.

Given a finite set of atomic position-dependent features of a binding site and a kinematic model for protein atom motions, the features (such as atomic charges) can be rigidly linked to the atoms that gave rise to the features. A linked feature can be used to exert a force on its kinematic chain, and it will move with its linked atom. The kinematic model describes how changes in joint (or bond dihedral) angles can be affected by changes in the positions of the site features and changes in the positions of site features are directly determined by changes in joint angles. The beauty of this idea is that it does not depend on a specific site representation, and it leverages optimization techniques based on kinematic models that are well studied in robotics and computer graphics [35, 3].

In this section, the objective function is intentionally left arbitrary, but would typically be a combination of surface shape and surface features (e.g., chemical type of the atoms contributing to that surface patch for a biochemistry problem, or color of eyes and skin if the matching involved face recognition). Here, the matching problem is posed in the context of searching for significant partial matches of **dataset** small molecule binding sites to a user provided **query** site. In this section, the query and dataset site features that are position and orientation dependent are represented by concatenating the positions and orientations into the long vectors $\mathbf{X}$ and $\mathbf{Y}$, respectively. Given an objective function depending on $\mathbf{X}$ and $\mathbf{Y}$, numerical optimization methods can be used to estimate the direction in which to move the system.

3.2 Inverse Kinematics

To better understand inverse kinematics [3], it is helpful to present the forward kinematics problem and kinematics terminology. Given a set of joints, a set of links, and the associated joint angles, **forward kinematics** computes the position of the end effectors (see Figure 1). One way to solve this problem is to start at the root joint (in our case, the Cα atom, which does not move), and for each successive joint in the chain compute the effect of its joint angle on the positions of the end effectors and the joints between the current joint and the end effectors.

The **inverse kinematics** problem is: given the desired positions of end effectors, find a set of angles that will move the end effectors to the desired positions. This problem is difficult because the inverse of a chain of rigid transformations determined by angles is nonlinear, and there could be anywhere from an infinite number of solutions to none. One technique is to use a linear approximation to the forward kinematics problem for the current joint and end effector configuration and solve for the joint angles [3]. Let $\mathbf{Q} = (Q_0, Q_1, \cdots, Q_m)$ be the set of joint angles of the deformable protein (this set is known as the **joint configuration space**). Let $\mathbf{Q}_0$ and $\mathbf{X}_0$ be the initial joint configuration and features' positions, respectively. Then, given a small change in the joint configuration space $\mathbf{Q}_0 + \Delta\mathbf{Q}$, what is the corresponding change in the end effector configuration space $\mathbf{X}_0 + \Delta\mathbf{X}$? This forwards kinematic problem may be represented by an operator

$$f(\mathbf{Q}_0 + \Delta\mathbf{Q}) = \mathbf{X_0} + \Delta\mathbf{X} \tag{1}$$

Since we seek to solve the inverse kinematics problem, one method is to assume that there is an inverse operator f^{-1}. Then applying f^{-1} to both sides of Equation 1 would yield

$$\mathbf{Q}_0 + \Delta\mathbf{Q} = f^{-1}(\mathbf{X_0} + \Delta\mathbf{X}) \tag{2}$$

One of the better ways to estimate f^{-1} is to compute the Jacobian $\mathbf{J}$ (i.e. partial derivatives) of the end effectors' positions $\mathbf{X}$ with respect to the joint angles $\mathbf{Q}$ at the current positions of the end effectors $\mathbf{X}_0$ [3, 8]. By letting $\mathbf{J} = [\partial X_i / \partial Q_j]$ a linear approximation to the forward kinematics problem is

$$\mathbf{J}\Delta\mathbf{Q} \approx \Delta\mathbf{X} \tag{3}$$

Supposing one can estimate $\Delta\mathbf{X}$ by the gradient of the objective function or another method, the idea is to find the inverse of the Jacobian and multiply both sides of Equation 3 from the left to yield the following

$$\Delta\mathbf{Q} \approx \mathbf{J}^{-1}\Delta\mathbf{X}$$

Since the Jacobian is rarely a square matrix, a generalized or pseudoinverse of the Jacobian must be used in the place of the inverse of the Jacobian. One can use the Moore-Penrose pseudoinverse of the Jacobian $\mathbf{J}^\dagger$ to obtain

$$\Delta\mathbf{Q} \approx \mathbf{J}^\dagger\Delta\mathbf{X} \tag{4}$$

In the remainder of this article, it is assumed that the pseudoinverse of the Jacobian $\mathbf{J}^\dagger$ refers specifically to the Moore-Penrose pseudoinverse.

A linear iterative method can be used to numerically solve the inverse kinematics problem. It is assumed that a reasonable binding site alignment is used to initialize the method. The initial joint and end effector configurations are given by the molecular structure (e.g. PDB coordinates) and the corresponding feature-labeled molecular surface representation as aligned to another protein structure and its binding

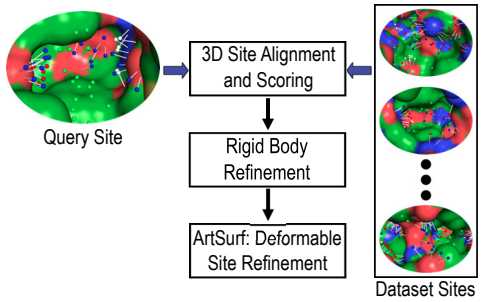

Figure 2: A high-level flowchart showing ArtSurf in the context of partial binding site matching.

site surface. At the ith iteration, compute or estimate how to move each end effector to optimize the objective (that is, determine a suitable estimate for $\Delta\mathbf{X}_i$). Next, compute the pseudoinverse of the Jacobian $\mathbf{J}^\dagger$ for the current joint and end effector configurations $\mathbf{Q}_i$ and $\mathbf{X}_i$, respectively. Then, given the desired change in the position(s) of the end effector(s) $\Delta\mathbf{X}_i$, use Equation 4 to estimate the change in the joint configuration $\Delta\mathbf{Q}_i$ that will approximately give the desired change in the position(s) of the end effector(s)

$$\mathbf{Q}_{i+1} = \mathbf{Q}_i + \Delta\mathbf{Q}_i = \mathbf{Q}_i + \alpha\mathbf{J}^\dagger\Delta\mathbf{X}_i \quad (5)$$

Next, apply the updated joint configuration $\mathbf{Q}_{i+1}$ using forward kinematics (Equation 1) to update the position(s) of the end effectors $\mathbf{X}_{i+1}$. Finally, repeat the steps till the process converges or a maximum number of iterations is reached.

4. METHODS

The deformable partial matching site comparison method as outlined in the preceding section is abstract. Depending on the level of detail used to represent binding sites and protein motions, implementing and using such a search method could require significant scientific, software development, and computational resources. To implement a deformable binding site comparison method, one must specify how to evaluate site similarity by providing a site representation and an objective function to optimize. Some of the more common features used to describe and compare binding sites are:

- Alpha and/or beta carbon atoms of amino acids [14]
- All non-hydrogen atoms that are near the surface of the protein [16]
- Chemical points that represent locally important chemical features [31, 11, 20, 30, 24, 33]
- A molecular surface of the binding site [27, 31, 30, 33]
- Electrostatic potential due to the separation between charged protein atoms [22]

The choice of features generally depends on the questions to be addressed and the resources that are available.

In order to narrow the scope and implement a prototype, several simplifying assumptions are used here:

- The relative positions of amino acid backbone atoms are held constant
- The relative positions of the atoms in amino acids outside the binding site are held constant

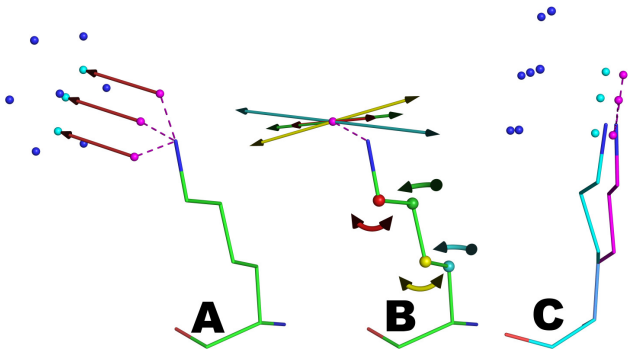

Figure 3: An illustration of steps in our deformable matching implementation. Some vectors and distances have been scaled to make the images more clear. A) The magenta and blue points represent query and dataset site molecular surface points, respectively. The cyan points are the closest points on the dataset molecular surface with respect to the three query points. The dashed lines imply that the query points are rigidly connected to the terminal nitrogen atom of the lysine side chain. The arrows denote the point correspondences. B) The four balls represent four joints, and the straight arrows correspond to the linear approximation of joint rotations. C) The magenta and cyan points are the positions of some site points before and after one iteration of our site matching method, respectively.

- Each binding site is represented by a SimSite3D binding site model [33]:
  - molecular surface patches represented by triangle meshes
  - chemically labelled points
- The length of the covalent bonds are constant
- Revolute joints (i.e. side-chain bond, or equivalently, dihedral angles rotations) are sufficient to model significant side chain motions

Here, locally deformable partial matching of binding sites is tested for the ability to improve chemical and molecular shape matching between a query site and a database site to which it has been aligned.

The following subsections highlight implementation details for the deformable partial matching binding of binding sites. These details include: representing binding sites, estimating point correspondences for two aligned binding sites, rigid body refinement of aligned sites using the Iterative Closest Point method (ICP) [6], and deformable refinement of aligned sites.

4.1 Binding Site Representation

For the results in this article, each binding site was represented by a SimSite3D model [33]. This model represents a binding site by its shape and chemistry. For a given binding site, the **binding site volume** was estimated as the union of spheres with radius r and centered at each non-hydrogen atom of the ligand that was bound to the protein.

The shape component of a binding site is represented by the intersection of that site's volume and a solvent accessible molecular surface of the protein. The surfaces were com-

puted using MSMS [29]. Because it is difficult to compare and render analytical representations of protein molecular surfaces, each molecular surface is represented by a triangle mesh. The default MSMS parameters (1 vertex per Å^2, 1.4 Å) were used for the vertex density and probe radius. In the interest of keeping contiguous triangle mesh patches, each triangle mesh, was pruned to its binding site by removing all triangles that did not have at least one vertex in the binding site volume (r is 4.0 Å). Therefore, the shape of a binding site is represented by one or more triangle meshes.

The chemical nature of a binding site is represented by local features computed from the protein's unpaired polar hydrogen atoms and lone pairs of electrons. Each chemical feature represents the interaction type, location, and favored orientation at which an atom in a small molecule could be placed in the protein's binding site and form a hydrogen bond with the corresponding protein atom. To allow for efficient and accurate matching, a spherical cap (section of a spherical shell) is used to approximate favored volumes for matching hydrogen-bonding atoms. Points on the cap that are within 2.5 Å of any protein atom are removed as ligand atoms at those points would severely overlap with one or more protein atoms. The relative orientation at any point on a spherical cap is approximated by a unit vector that is normal to the cap at that point. The details of defining the spherical caps, accounting for partial occlusion from protein atoms, and how to determine the closest point on a spherical cap can be found in [33].

4.2 Estimating Site Correspondences

One straightforward method to estimate the degree of correspondence between two sites is to compute the site complementarity at a number of sample points and feed the samples into a site similarity function [6]. The query site is represented by the vertices of the mesh representing the site's shape and by points sampled relatively uniformly on the site's polar spherical caps. The shape point correspondences are determined by finding the closest point on the dataset site's mesh surface for each query shape point. The polar point correspondences are determined by finding the closest complementary point on the dataset site's spherical caps for each query polar point (respecting the chemical types) [33]. Each query point has at most one corresponding point, and if there is no corresponding point within 1.5 Å of a given query point, that point does not have a correspondence.

By defining point correspondences for the query (partial site) sample points and using a maximum correspondence threshold, we can minimize the distances between corresponding points to optimize a partial match between the two sites. Alternating between computing point correspondences and minimizing an objective function of the correspondences is an iterative process used in many contexts to numerically optimize difficult problems. A well known rigid body refinement method, ICP, uses such iterations [6], and ICP is directly applicable to our partial matching problem.

4.3 Objective Function

Since the goal is to maximize the degree of similarity between two SimSite3D site representations, the objective function reduces to minimizing the sum of squared distances between corresponding points. Note that the points can move relative to each other and in different directions. Therefore, suppose we have a query site with n points, then the n query points x_i (including the surface shape and chemical points) can be concatenated to form a $\mathbf{X}$ vector of length $3n$. A corresponding vector $\mathbf{Y}$ of the length $3n$ is computed by finding the closest corresponding dataset point y_i for each x_i. If there is not a corresponding y_i with in the cutoff distance of a given x_i, then that y_i and x_i have their three coordinates set to zero. The objective function $\mathbf{E}$ is half of the square of the Euclidean distance between $\mathbf{X}$ and $\mathbf{Y}$. The gradient of $\mathbf{E}$ is elementary:

$$\nabla_{\mathbf{X}}\mathbf{E} = (X_0 - Y_0, X_1 - Y_1, \cdots, X_{3n} - Y_{3n}) \qquad (6)$$

Because the gradient gives the direction of the greatest increase, one can minimize the difference of the point correspondences by moving in the direction of the negative gradient.

	X_{3i}	X_{3i+1}	X_{3i+2}
Lys CA-CB bond angle($\chi_{j,1}$)	$J_{A,3i}$	$J_{A,3i+1}$	$J_{A,3i+2}$
Lys CB-CG bond angle($\chi_{j,2}$)	$J_{B,3i}$	$J_{B,3i+1}$	$J_{B,3i+2}$
Lys CG-CD bond angle($\chi_{j,3}$)	$J_{C,3i}$	$J_{C,3i+1}$	$J_{C,3i+2}$
Lys CD-CE bond angle($\chi_{j,4}$)	$J_{D,3i}$	$J_{D,3i+1}$	$J_{D,3i+2}$

Table 1: Example of the part of a Jacobian block corresponding to a lysine side chain and the position of a chemical feature (Figure 3). The values of row 1, 2, 3, and 4 are multiples of the components of the cyan, yellow, green, and red tangent vectors, respectively, (Figure 3).

4.4 Data Structures

A computationally efficient set of data structures is needed to implement an efficient variant of the deformable matching as presented in this section. Besides the typical data structures used for matching point clouds and triangle meshes, the dependencies of the site features on the protein's side-chain dihedral angles need to be represented with respect to the forward and inverse kinematics problems. Because the protein's backbone dihedral angles are kept constant and each site point is assigned to exactly one kinematic chain, the Jacobian in the linear approximation to the forward kinematics problem (Equation 3) is a block matrix. Each block matrix and the memory addresses (pointers) of the positions and orientations of the associated site features are stored in a class (or object). Since some positions may not depend on all of the joint angles, each feature has a count of the number of joint angles on which it depends. By storing the Jacobian (and its pseudoinverse) as a block matrix and associating the joint dependent components of the features with each block, block matrix operations can be used. This helps the implementation to be efficient with respect to memory space and computational time.

4.5 Deformable Site Matching

Putting the ideas together, an iterative method can be applied to deform one binding site to be more similar to another. Since the point correspondences between the sites are local and the deformations are local, a good site alignment is needed before deformations are attempted. In our implementation, initial site alignments are refined by using ICP on the sites' surfaces and spherical caps.

The deformable matching loop is as follows:

- Given an initial alignment or an alignment from a previous iteration, compute the current polar and molecular surface site correspondences (see Figure 3 A).
- Compute the Jacobian $\mathbf{J}$ from the query protein's current joint configuration space $\mathbf{Q}$ and site point positions $\mathbf{X}$ (see Figure 3 B).
- Determine the pseudoinverse of the Jacobian $\mathbf{J}^\dagger$.
- Use the negative gradient of the objective function (Equation 6) to estimate the desired change in position of the query's site points.
- Multiply a fraction α of the negative gradient by the pseudoinverse of the Jacobian to determine small changes in the joint angles (Equation 5) that will reduce the current value of the objective function.
- Rotate the query's joints with respect to the computed changes in joint angles to update the positions of the query's atoms and site features (see Figure 3 C).
- Terminate the loop when the last change in the joint angles is sufficiently small or after a maximum number of iterations is reached.

4.6 Implementation Details

The numerical portion of the deformable matching consists of numerically solving the optimization and inverse kinematics problems. The Jacobian is a block matrix and can be computed relatively easily. The key idea is the portion of the Jacobian that corresponds to a linear approximation of a dihedral bond rotation's effect on the position of a site point reduces to the cross product between the axis of rotation (as a unit vector) and the vector from the joint center to the center of the site point.

Some care is needed in computing the pseudoinverse of the Jacobian $\mathbf{J}^\dagger$. One common method to compute the pseudoinverse is

$$\mathbf{J}^\dagger = (\mathbf{J}^t\mathbf{J})^{-1}\mathbf{J}^t \tag{7}$$

However, numerically computing the inverse of $\mathbf{J}^t\mathbf{J}$ will be problematic if $\mathbf{J}^t\mathbf{J}$ is singular or almost singular. Damped least squares regularization can be used to keep the computation relatively stable by adding a small positive constant to the diagonal of $\mathbf{J}^t\mathbf{J}$ that results in the approximation

$$\mathbf{J}^\dagger \approx (\mathbf{J}^t\mathbf{J} + \lambda^2\mathbf{I})^{-1}\mathbf{J}^t \tag{8}$$

where $\mathbf{I}$ is the identity matrix of the same size as $\mathbf{J}^t\mathbf{J}$ and λ is a small positive constant [8]. Note that if $\mathbf{J}$ is an $m{\times}n$ matrix and $m \ll n$ then the order of operations can be reversed so that the size of the matrix to invert is $m \times m$ rather than $n{\times}n$. LAPACK [1] methods can be used to compute $(\mathbf{J}^t\mathbf{J})^{-1}$ by first computing an LU decomposition of $\mathbf{J}^t\mathbf{J}$ and then determining the inverse.

After $\mathbf{J}^\dagger$ is computed, (Equation 5) can be used to update the joint angles, and forward kinematics can be used to update the positions of the site points. Empirically, it was found that α values of 0.4 and 0.1 for the chemical point and surface point correspondences, respectively, provide a reasonable trade off between step size and the number of iterations to converge. Additionally, since the pseudoinverse Jacobian method for inverse kinematics is a linear method, a maximum rotation of five degrees per iteration appears to be within reason.

One final item to consider is that severe overlaps of pairs of atoms within a protein are not physically realistic. The overlap of any two protein atoms (in the same protein) is allowed to be at most five percent of the sum of the atoms' Van der Waals radii (as determined by Li and Nussinov, Table XI [23]). The two exceptions to the overlap constraint are atoms that could participate in a hydrogen bond are allowed to be within 2.5 Å and atoms that had a severe overlap (five percent or greater) in the provided structure. For cases of severe overlap (more than five percent), dihedral angle rotations that would result in an increase in overlap are clamped (at five percent overlap).

5. RESULTS

A suitable binding site dataset is needed to gauge the applicability and performance of our method and implementation. Before applying the method to non-homologous binding sites, it is a good practice to first assess if the method can adequately handle different conformations of the same binding site. Thus, two sets of molecular dynamics (MD) trajectories were used to show performance on the same binding site that has undergone increasing displacements in mainchain atomic positions (conformational change). The two MD trajectories (one conventional and one Hamiltonian replica exchange (HREM)) were provided by Su and Cukier [32], and these simulations show the pterin binding site of *Y. pestis* 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase (HPPK) as it undergoes (simulated) low-energy conformational changes over time. Each MD trajectory has about 3,000 snaphots (atomic coordinate files) that correspond to $1ps$ timesteps of $3ns$ of MD simulation. Both trajectories have had the coordinates of the all snapshots brought into the same reference frame (as aligned by Su [32]). Applying our implementation to selected snapshots provides an example of deforming a binding site representation via directed sidechain refinement in the case of sites with the same sequence that have undergone increasing main chain conformational changes.

The HPPK MD binding site dataset was constructed as follows. The binding site amino acids (of HPPK) were selected using molecular graphics. The selected amino acids correspond to the residue numbers 7, 9, 45, 51, 54, 83, 88, 89, 91, 93, 96, 98, 116-118, 122-125, and 156 from the PDB structure 2qx0. The first snapshot of the conventional MD trajectory was selected as the reference coordinates, and the root mean squared deviations (RMSD) of the main chain atoms of the binding site amino acids was computed for each snapshot with respect to the reference. The snapshots were partitioned by RMSD into non-overlapping bins with width 0.25 Å for the interval $[0.0, 4.0]$ Å RMSD. The conventional snapshots were placed in the first eight bins, and any snapshot with an RMSD greater than 2.0 Å was discarded. The HREM snapshots with RMSD in the interval $[2.0, 4.0]$ Å RMSD were placed in the next eight bins. The only snapshot in the interval $[0.0, 0.5]$ Å RMSD was the reference snapshot. For each bin, the representative coordinates of that bin are given by the snapshot with the lowest main chain RMSD in that bin. This resulted in an HPPK MD binding site dataset with 15 sets of coordinates (snapshots) with increasing main chain binding site RMSD (with respect to the reference coordinates).

Because the goal was to evaluate the performance of the deformable site matching implementation, each of the snapshots was aligned to the reference snapshot by minimizing the binding site main chain atomic RMSD between each

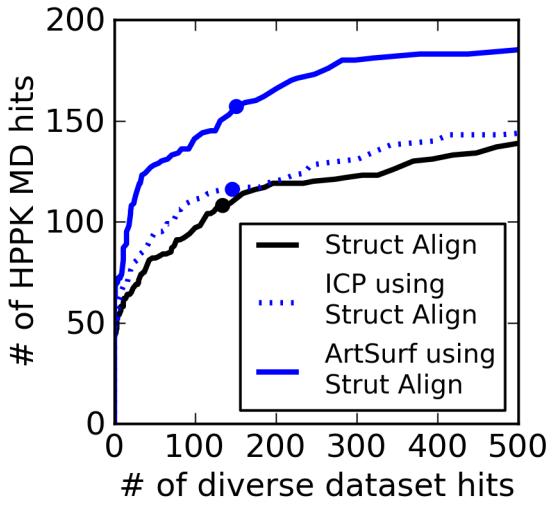

Figure 4: Performance of ICP and Artsurf starting with the initial alignments of the *Yp* HPPK dataset. Each curve shows the number of true positive versus false positive matches as the score thresholds are decreased from most (lower left corner) to least (upper right corner) stringent. The number of false positive matches is estimated by the number of high scoring matches between the 15 *Yp* HPPK query sites and 140 diverse sites [33]. The plotted scores are z-scores (normalized per query site), and the dots denote a standard deviation of 1.5 better than the mean. Struct Align is the normalized SimSite3D similarity scores for the initial (coarse) alignments. The ICP and ArtSurf lines are after applying ICP and ArtSurf respectively.

snapshot and the reference snapshot. In keeping with the partial matching theme of the article, 15 somewhat larger dataset site and 15 query site representations were constructed using SimSite3D [33]. The PDB structure of *Yp* HPPK (2qx0) was aligned to the reference coordinates, and the pterin ligand (PDB small molecule code PH2) from 2qx0 was used to define the volume of the query site for each of the fifteen snapshots. The dataset site for each of the snapshots was defined by a 8.0 Å sphere centered at the center of the pterin ring system. Each query site was compared with each dataset site for a total of 225 site comparisons.

To ignore protein small-molecule binding site comparision tool effects, the initial alignment for each query, dataset pair of sites was given by the provided main chain alignments to the reference structure [32]. Reference scores for each site pair were computed by applying ICP to each initial alignment and then, computing the SimSite3D site similarity score. Because the SimSite3D site similarity score is directly dependent on the method used to compute the corresponding points, each of the 225 site similarity scores were more favorable after deforming the query sites. In addition, we saw that applying the deformable matching method to each query site versus a set of 140 diverse protein small-molecule binding sites, on average, did not result in as much of a score increase as a query site versus an HPPK dataset site. The end result is deformable partial matching resulted in an increase of about 50 true positive matches at virtually no increase in the number of false positive matches.

6. DISCUSSION AND FUTURE WORK

A relatively straightforward and general method using the pseudoinverse Jacobian method of inverse kinematics was presented to adjust protein dihedral angles to achieve the desired changes in positions of binding site features. The deformable binding site matching implementation is novel in that one can direct protein side-chain conformational change via optimization of binding site features. At no point is our method directly dependent on the tool used to provide an initial alignment of the binding sites; applicable methods include Dali [19], substructure matching [15], and SimSite3D [33]. Also, to the authors' knowledge, the problem of realistically deforming the site representation of one site to maximize its similarity to a second site has not been addressed before. This article illustrates that deformable partial matching of protein small-molecule binding sites can be performed in a reasonable amount of time, and can help increase the number of true positive matches.

Areas of improvement include: prioritization of rotations at each iteration, allowing more atoms to move (e.g. beta carbon atoms), tuning and validation on protein families, and better molecular surface and chemical group approximations. Also, because the method uses many matrix and vector operations, it would greatly benefit from modifying the implementation to use multiple CPU cores or graphics cards for most of the computation.

7. ACKNOWLEDGEMENTS

The authors take this opportunity to thank Barry C. Finzel for his helpful comments and insight with regards to preparing this manuscript. The work was funded by grants from Pfizer and the NSF (IIS 0953096) and (CCF 0811313) and by a Dissertation Completion Fellowship from Michigan State University.

8. REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, PA, third edition, 1999.

[2] N. Andrusier, R. Nussinov, and H. J. Wolfson. FireDock: fast interaction refinement in molecular docking. *Proteins: Structure, Function, and Bioinformatics*, 69(1):139–159, 2007.

[3] P. Baerlocher. *Inverse Kinematics Techniques for the Interactive Posture Control of Articulated Figures*. PhD thesis, Ecole Polytechnique Federale de Lausanne, 2001.

[4] J. A. Barker and J. M. Thornton. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, 19(13):1644–1649, Sept. 2003.

[5] H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide protein data bank. *Nat. Struct. Mol. Biol.*, 10(12):980, Dec. 2003.

[6] P. Besl and H. McKay. A method for registration of 3-D shapes. *IEEE TPAMI*, 14(2):239–256, 1992.

[7] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Generalized multidimensional scaling: A framework for isometry-invariant partial surface matching. *PNAS*, 103(5):1168–1172, 2006.

[8] S. R. Buss and J. Kim. Selectively damped least squares for inverse kinematics. *Journal of Graphics Tools*, 10(3):37–49, 2005.

[9] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Prot. Sci.*, 12(9):2001–2014, Sept. 2003.

[10] S. Corazza, L. Mundermann, E. Gambaretto, G. Ferrigno, and T. Andriacchi. Markerless motion capture through visual hull, articulated ICP and subject specific model generation. *IJCV*, 87(1):156–169, Mar. 2010.

[11] Z. Deng, C. Chuaqui, and J. Singh. Structural interaction fingerprint (SIFt): a novel method for analyzing Three-Dimensional Protein-Ligand binding interactions. *J. Med. Chem.*, 47(2):337–344, Jan. 2004.

[12] D. Droeschel and S. Behnke. 3D body pose estimation using an adaptive person model for articulated ICP. volume 2, pages 157–167, Aachen, Germany, 2011. Springer.

[13] I. Eckstein, A. A. Joshi, C. J. Kuo, R. Leahy, and M. Desbrun. Generalized surface flows for deformable registration and cortical matching. *MICCAI*, 10(1):692–700, 2007.

[14] H. J. Feldman and P. Labute. Pocket similarity: are alpha carbons enough? *J. Chem. Inf. Model.*, 50(8):1466–1475, Aug. 2010.

[15] B. C. Finzel, R. Akavaram, A. Ragipindi, J. R. Van Voorst, M. Cahn, M. E. Davis, M. E. Pokross, S. Sheriff, and E. T. Baldwin. Conserved core substructures in the overlay of Protein-Ligand complexes. *J. Chem. Inf. Model.*, 51(8):1931–1941, Aug. 2011.

[16] N. D. Gold and R. M. Jackson. A searchable database for comparing Protein-Ligand binding sites for the analysis of Structure-Function relationships. *J. Chem. Inf. Model.*, 46(2):736–742, Mar. 2006.

[17] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker. Protein-Protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, 331(1):281–299, Aug. 2003.

[18] H. Hasegawa and L. Holm. Advances and pitfalls of protein structural alignment. *Current Opinion in Structural Biology*, 19(3):341–348, June 2009.

[19] L. Holm and C. Sander. Dali: a network tool for protein structure comparison. *Trends in Biochemical Sciences*, 20(11):478–480, Nov. 1995.

[20] M. Jambon, A. Imberty, G. Delage, and C. Geourjon. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins: Structure, Function, and Genetics*, 52(2):137–145, 2003.

[21] O. Kalinina, M. Gelfand, and R. Russell. Combining specificity determining and conserved residues improves functional site prediction. *BMC Bioinformatics*, 10(1):174, 2009.

[22] K. Kinoshita and H. Nakamura. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Prot. Sci.*, 12(8):1589–1595, Aug. 2003.

[23] A. Li and R. Nussinov. A set of van der waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins: Structure, Function, and Genetics*, 32(1):111–127, 1998.

[24] F. Moriaud, S. A. Adcock, A. Vorotyntsev, O. Doppelt-Azeroual, S. B. Richard, and F. Delfaud. A computational fragment approach by mining the protein data bank: Library design and bioisosterism. In *Library Design, Search Methods, and Applications of Fragment-Based Drug Design*, volume 1076 of *ACS Symposium Series*, pages 71–88. ACS, 2011.

[25] S. Pellegrini, K. Schindler, and D. Nardi. A generalization of the ICP algorithm for articulated bodies. In *British Machine Vision Conference*, 2008.

[26] RCSB Protein Data Bank. Yearly growth of total structures. *http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100*, 2012.

[27] M. Rosen, S. L. Lin, H. Wolfson, and R. Nussinov. Molecular shape comparisons in searches for active sites and functional similarity. *Protein Engineering*, 11(4):263–277, Apr. 1998.

[28] L. Sael and D. Kihara. Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. *Proteins: Structure, Function, and Bioinformatics*, 80(4):1177–1195, 2012.

[29] M. F. Sanner, A. J. Olson, and J. Spehner. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, 1996.

[30] S. Schmitt, D. Kuhn, and G. Klebe. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, 323(2):387–406, Oct. 2002.

[31] A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson. Recognition of functional sites in protein structures. *J. Mol. Biol.*, 339(3):607–633, June 2004.

[32] L. Su and R. I. Cukier. Hamiltonian replica exchange method study of escherichia coli and yersinia pestis HPPK. *J. Phys. Chem. B*, 113(50):16197–16208, Dec. 2009.

[33] J. R. Van Voorst. *Surface Matching and Chemical Scoring to Detect Unrelated Proteins Binding Similar Small Molecules*. PhD thesis, Michigan State University, 2011.

[34] I. Wallach and R. H. Lilien. Prediction of sub-cavity binding preferences using an adaptive physicochemical structure representation. *Bioinformatics*, 25(12):i296–i304, June 2009.

[35] C. Welman. Inverse kinematics and geometric constraints for articulated figure manipulation. Master's thesis, Simon Fraser University, 1993.

[36] M. I. Zavodszky and L. A. Kuhn. Side-chain flexibility in protein-ligand binding: The minimal rotation hypothesis. *Prot. Sci.*, 14(4):1104–1114, Apr. 2005.