Abstract—The determination of whether a molecule can bind or "dock" a protein to a certain site depends on the orientation of the molecules, the charge of the atoms that comprise the molecules, and the electrical potential of the proposed area. It is said that a fundamental problem with molecular docking is that the orientation space itself is very large and grows in a combinatorial manner with the number of degrees of freedom of the interacting molecules. We tried to cleverly solve this problem by using shape definitions and fuzzy logic to help reduce the search size of possible docking locations and predict which locations and orientations are the most likely docking locations.

Index Terms—Molecular Protein Docking Algorithm; Molecular Binding; Computational Prediction; Computational Probability and Analysis; Data Mining; Relational Database; Fuzzy Decisioning

I. INTRODUCTION

Molecular docking is a computational procedure that attempts to predict noncovalent binding of macromolecules with other ligands; this binding can affect the behavior of the molecules are around the protein [1] as well as, in organisms, lead to various different effects in that organism whether it is the production of inhibitors or the reproduction of virus. For this reason, the determination of the binding of small molecules to proteins has a very practical application in the field of drug and toxin design [2]. There are many methods for molecular docking and the algorithms to predict binding that is being developed are steadily increasing [3]. Many of these algorithms share common methods with some of the most popular being:

- Molecular Dynamics Methods [4] [5]
- Genetic Algorithms and evolutionary programming [6]
- Fragment-based methods [7] [8]
- Point Complementary Methods [9]
- Quantitative Structure-Activity Relationship (QSAR) Models [10]

Our novel method performs the classification of shapes along each protein and molecule to significantly reduce the search space of likely docking locations by eliminating pairings of unlikely shapes. The cornerstone of our algorithm was the integration of a relational database to hold the atom and protein data. We use the energies of the atoms from the protein and molecule entries to create "surfaces" around each object. Geometric shapes are approximated from the imposed surface structure and the shape of the space around each object. The coordinates of the discovered shapes are saved into the database, then compared to see possible docking locations based on the geometry that is created. This reduces the actual problem space required to find potential docking locations of molecules to proteins. As surfaces are created on both objects, structures and pockets arise much like in a puzzle: a surface of one object may be able to fit into a pocket of the other.

During the initial Discovery step, molecule border definition takes place. We take each atom that comprises the ligand that we are working with and create a surface for each using its

<table>
<thead>
<tr>
<th>Atomic Number</th>
<th>Element Symbol</th>
<th>Van Der Waals Radius</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>1.2</td>
</tr>
<tr>
<td>6</td>
<td>C</td>
<td>1.7</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>1.55</td>
</tr>
<tr>
<td>8</td>
<td>O</td>
<td>1.52</td>
</tr>
<tr>
<td>9</td>
<td>F</td>
<td>1.47</td>
</tr>
<tr>
<td>15</td>
<td>P</td>
<td>1.8</td>
</tr>
<tr>
<td>16</td>
<td>S</td>
<td>1.8</td>
</tr>
<tr>
<td>–</td>
<td>–</td>
<td>2</td>
</tr>
</tbody>
</table>

II. ALGORITHM

1) Import the atoms into the database.
2) Add the shape and charge definitions for all proteins.
3) Determine the shape and charge definitions for the small molecule.
4) Cross-reference the two different lists of shapes to generate possible interactions and rank these interactions according to electrostatic favorability.

III. SHAPE DETERMINATION

To begin, we extracted key information from standard protein database files (PDB files) [11] into a relational database that we created to hold atom and protein data. We use the energies of the atoms from the protein and molecule entries to create “surfaces” around each object. Geometric shapes are approximated from the imposed surface structure and the shape of the space around each object. The coordinates of the discovered shapes are saved into the database, then compared to see possible docking locations based on the geometry that is created. This reduces the actual problem space required to find potential docking locations of molecules to proteins. As surfaces are created on both objects, structures and pockets arise much like in a puzzle: a surface of one object may be able to fit into a pocket of the other.
Van-Der-Waals radius, as referenced in Table I. Using the radii of the respective atom and a shell spacing constant $s$ we represent the surface of each atom by using a system of $x$, $y$, $z$ coordinates, as show in Equation (1).

$$\theta = \arccos \left( -s^2 - 2r^2 \right) \over 2r^2 \tag{1}$$

$$A = 2\pi, \quad B = \pi + 1$$

atom$_{surface} = (x + r \cos (\theta \ast a) \sin (\theta \ast b), \quad y + r \sin (\theta \ast a) \sin (\theta \ast b), \quad z + r \cos (\theta \ast b))$$

$\exists a \in [0, A], \exists b \in [0, B]$ 

To model the surface, a shell is built around each atom; the radius of the shell is determined by the radius of the element. Next, all of the points on this sphere that overlap a neighboring atom’s sphere are absorbed and removed, clumping the atoms together, and the protein is defined as a molecule. The union of all of the atom surfaces represents the complete surface of the ligand as a whole.

A sphere is now constructed around this molecule, and probes are extended inwards towards until they touch the ligand surface (Figure 2). These probes are stored in the database as the definition of the shape of the molecule.

Individual surface Shape determination is generated by taking collections of adjacent probes. Each collection has its own individual shape, which is determined by the total number of probes in each shape, as well as the maximum difference along each axis. In order to divide the ligand surface into individual shapes suitable for matching as potential docking sites, each probe is examined and determined whether it is a local minimum (for a Valley) or maximum (for a Peak).

If it is not a Peak or a Valley, then the probe is added to a queue, and the next point down is examined in sequence until a known Shape is encountered or a minimum or maximum is reached, defining a new Shape. This process is repeated for all probes until every part of the surface has been explored and assigned to a Shape.

To calculate if a binding between the two shapes is electrochemically favorable, we measure the electrical potential and the electric dipole potential. Both are used to measure the possible interaction between the respective shape and the others stored in the database.

Electric potential is applied to each molecule as a whole as well as the shapes that are created by them. The calculation is done by taking a ratio of a point charge $Q$ and a distance from the atoms, $r$, with $\epsilon_0$ representing the Coulomb potential, which is an electrical constant measuring the electric potential of free space.

The calculation of the electric dipole moment in Equation (4) is taken to determine the strength of the charge as well as overall polarity of the shape. The potential of an electrical dipole is found by superposing the point charge potentials of the charges. The calculation we used assumes that the system has an overall neutral charge.

$$V_E = \frac{1}{4\pi\epsilon_0} \frac{Q}{r} \tag{2}$$

$$\nabla V_E = -\frac{Q}{4\pi\epsilon_0 r^3} (x \hat{i} + y \hat{j} + z \hat{k}) \tag{3}$$

$$D_E = Qx\hat{i} + Qy\hat{j} + Qz\hat{k} \tag{4}$$

V. FUZZY LOGIC

The fuzzy logic is applied to both the shape and the electrical potential separately; then, using standard fuzzy logic, the membership values are applied to a logical AND operation by multiplying the two values together. For simplicity, we name the membership values for shape fitting and electrical potential the fit value and the affinity value respectively.
The fit value is calculated by comparing the overall volume, size, depth, and width \( w \) of the shapes. From this we calculate a shape similarity by Equation (5). If the size of the docking molecule is ever overly large for a prospective pocket shape, it is automatically given a membership value of zero which effectively cancels the likelihood of membership. A comparison of similarity is then taken and multiplied together to get total fuzzy membership value.

The dot product of the vector of the charges of the molecule \( m \) and protein \( p \) are calculated to determine the fuzzy membership value for binding affinity (Equation (6)). Ideally, the two charges will be equal and opposite, meaning that perpendicular charges would be unfavorable. From this logic we can then calculate the binding affinity membership. When we combine shape similarity and binding affinity calculations we will calculate our overall membership value for binding. The Fuzzy Determination calculation in Equation (7) weights the shape more heavily in the overall membership function.

\[
SS (p, m) = \left( 1 - \frac{w_p - w_m}{w_p} \right) \ast \left( 1 - \frac{d_p - d_m}{d_p} \right)
\]

\[
BA (p, m) = \frac{C_p \cdot C_m}{\text{max} (\|C_m, C_p\|)}^2
\]

\[
f_d (p, m) = \frac{SS_{p,m} (BA (p, m) + 1)}{2}
\]

VI. Conclusions and Future Work

The current version of the algorithm uses a single probe per shape, comparison measures for shape similarity to consider the width, height and depth of a potential docking location, and produces good results in terms of speed of return. Running the Darunavir molecule against the shapes that correlated to the each of the 7 proteins in the database completed in 27,644 seconds, an average of roughly 4 seconds per protein when searching through all feasible docking locations on each protein.

Natural usage and growth in our database will cause it to become a repository of similar data: in the future it would also be desirable to use machine learning techniques to find functional groups of proteins using shared shape information, as geometry plays a large role in protein function and the determination and classification of functional groups of proteins is a large open area in the bioinformatics community.

REFERENCES


