A Patch-by-Patch Shape Matching Procedure for Rigid Body Docking

Vipin K. Tripathi *

Abstract—**Docking simulates molecular interactions. Protein - protein docking, owing to the sizes of molecules, is a very challenging problem. As the number of degrees of freedom in the molecular structure of proteins is huge, docking algorithms consider the molecules as relatively rigid bodies. Shape matching of two protein molecules is computationally difficult because the possibilities of matching the molecules grow exponentially with the size of molecules which gives a large search space. A new scheme is proposed in this paper for shape matching in rigid body docking. The scheme models the molecular surface as a network of locally characterized surface patches connected to each other at their boundaries. For shape matching of molecules, a patch-by-patch shape matching strategy is used. The procedure suggested for shape matching can be easily included in an optimization routine for finding the best binding conformations of two protein molecules.**

Keywords: docking, shape matching, bioinformatics, computer aided design

1 Introduction

Computational molecular docking aims at finding the best binding conformations of two given molecules. The molecules may be two protein molecules or a protein and a ligand molecule. Docking is important in relation to molecular recognition, macromolecular assemblies, cellular pathways, protein folding etc. [1]. It has an industrial application in rational drug design [2]. A comprehensive review of the docking algorithms can be found in [3]. Docking of two molecules is shown in Figure 1. The patches on the surface of outer molecule are indicated by small triangles. Lines indicate the boundaries of the patches.

A computational docking procedure involves three steps: (i) shape representation of molecules (ii) shape matching and (iii) optimization of shape matching. Generally, the methods used for docking are, broadly, classified into two categories: rigid body docking and flexible ligand
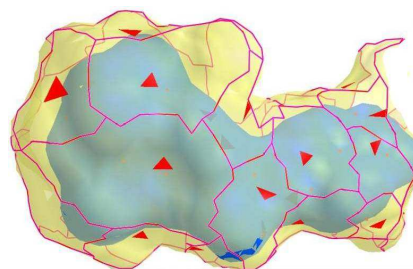

Figure 1: Docking of two molecules

docking. The rigid body docking methods consider both the molecules as rigid bodies, whereas the flexible ligand methods consider one or both the molecules flexible.

Different mathematical model have been used for molecular shape representation. Connolly [4] laid the foundation of representing the molecular surface by its geometric features. Lin et al. [5] describe the surface by sparse critical points. Sanner et al. [6] compute a reduced surface in their MSMS algorithm. Some researchers have used 'shape-implicit' methods which discretize the molecule onto a grid in space and consider only the occupied cells, i.e., voxels, to define the shape of the molecule [7].

The shape matching procedure proposed in this paper, considers molecular surface as a network of locally characterized surface patches connected to each other at their boundaries. Those surface patches of the molecules are modeled by B-spline surface representation technique. In the process, an inverse problem in the B-spline surface fitting is solved. B-spline surface approximation technique offers a lot of flexibility required for fitting a surface through scattered data [8]. It is, therefore, a promising shape representation technique for molecular docking. A number of studies [9] focus on modeling a surface patch from 2-D parameterized data points. The procedure used in this study for modeling the surface patches is described in section 2 of this paper.

Factors like alternate binding locations and orientations in which two molecules can bind with each other, non-exact nature of structure of the molecules, irregular molecular surfaces, conformational changes taking place upon association, etc. make docking a very complex

*Mechanical Engineering Department, Government Engineering College Aurangabad, Maharashtra India 431005. Email: vipinkt_2000@yahoo.com. Author is thankful to Santosh Tiwari, Research Associate at I.I.T. Kanpur, for helping in elegant implementation of the algorithm proposed in this paper. Manuscript submitted on 13th March, 2007

problem. Many approaches have been tried for shape matching in docking. These matching algorithms may be classified into two broad categories: (i) brute force enumeration algorithms and (ii) local shape feature matching algorithms. Local shape feature matching algorithms have been pioneered by Kuntz [10]. These algorithms use complementary features of surfaces to align them. They include algorithms like distance geometry algorithms [10], pose clustering [7], geometric hashing and geometric hashing and pose clustering.

In the present study, a different approach is proposed for shape matching of the molecules. A point on each of the two surfaces and a triad of three mutually perpendicular unit vectors at each of the two points is used as local shape feature for matching the surface patches from two different molecules. An association of two molecules can be rated as correct or incorrect based on different criteria like shape complementarity between molecules, binding energy of a complex, electrostatic and van der Waals interactions etc. Although other criteria are also important, shape complementarity is considered a necessary condition in rigid body docking. Shape complementarity criterion was initially adopted by Kuntz and his colleagues [10] and was further elaborated and rationalized by Connolly [11]. Here also, shape complementarity has been used as the criterion for matching of the surface patches. For shape matching, only the patches with complementary curvatures are aligned with each other. The method proposed here is simple and effective. It can easily be used in an optimization routine with a suitable scoring function to find the best binding conformations.

2 Modeling of Molecular surface

The molecular surface is considered to be made up of surface patches of different forms such as convex, concave , flat etc. connected to each other. B-spline surface representation technique is used in this paper to generate the patches of molecular surface. The B-spline surface is controlled by a characteristic polyhedron as shown in Figure 2. Advantages of this surface are the local control of surface and independence of the degree of surface from number of control points. The surface generation scheme requires molecular data already parameterized on a 2-D grid as input. The rectangular data grid is shown in Figure 3.

A tensor product B-spline surface is formulated as

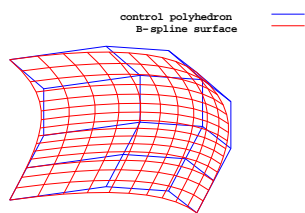$$\mathbf{Q}(u,v) = \sum_{i=0}^{n} \sum_{j=0}^{m} \mathbf{P}_{ij} N_{i,k}(u) N_{j,l}(v) \tag{1}$$
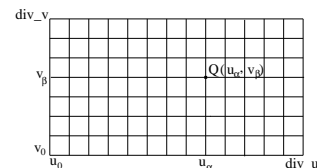

Figure 2: B-spline surface Figure 3: Rectangular 2-D grid of data points

where

$$
\begin{array}{rcl}
\mathbf{Q}(u,v) & = & \text{a point on the surface,} \\
\mathbf{P}_{ij} & = & \text{vertices of the defining} \\
& & \text{polyhedron,} \\
N_{i,k}(u),\ N_{j,l}(v) & = & \text{blending functions in u,} \\
& & \text{v directions, respectively.} \\
k,\ l & = & \text{orders of curves in u,} \\
& & \text{v directions, respectively.}
\end{array}
$$

The B-spline blending functions are computed by the following recursive expressions:

$$N_{i,1}(u) = \begin{cases} 1 & \text{if } t_i \leq u < t_{i+1}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

$$N_{i,k}(u) = \frac{(u - t_i)N_{i,k-1}(u)}{t_{i+k-1} - t_i} + \frac{(t_{i+k} - u)N_{i+1,k-1}(u)}{t_{i+k} - t_{i+1}} \tag{3}$$

The t_i are known as knot values. They relate the parametric variable u to the control points $\mathbf{P}_{ij}$. For an open curve, t_i are given by

$$t_i = \begin{cases} 0 & \text{if } i < k, \\ i - k + 1 & \text{if } k \leq i \leq n, \\ n - k + 2 & \text{if } i > n, \end{cases} \tag{4}$$

with

$$0 \leq i \leq n + k. \tag{5}$$

The range of parametric variable u is

$$0 \leq u \leq n - k + 2. \tag{6}$$

Finding $\mathbf{Q}$, the surface data points, by an input of $(n + 1) \times (m + 1)$ number of control points which are the vertices of control polyhedron $\mathbf{P}_{ij}$, is known as the 'forward problem'.

For shape representation the inverse problem is of interest, i.e., given a known set of data points on a surface, determine the defining polyhedron for B-spline surface that best approximates the data. The control points obtained after solving the inverse problem, being very less in number as compared to the given data points, are used for generation and processing of surface patches. This needs less storage and processing time. Major steps of

the method used here for solving the inverse problem are given below.

Writing out Eq. 1 for a single surface data point, say, point $\mathbf{Q}(u_\alpha, v_\beta)$ as shown in Figure 3 yields

$$
\mathbf{Q}(u_\alpha, v_\beta) = \begin{bmatrix} N_{0,k}(u_\alpha) & \cdots & N_{n,k}(u_\alpha) \end{bmatrix} \begin{bmatrix} \mathbf{P}_{ij} \end{bmatrix} \begin{bmatrix} N_{0,l}(v_\beta) \\ N_{1,l}(v_\beta) \\ \vdots \\ N_{m,l}(v_\beta) \end{bmatrix}. \tag{7}
$$

Writing an equation of this form for each data point on the rectangular grid yields a system of simultaneous equations. In the matrix form the result can be written as

$$
[\mathbf{Q}] = [\mathbf{C^t}][\mathbf{P_{ij}}][\mathbf{D}] \tag{8}
$$

where $\mathbf{Q}$ is the matrix of data points. It has $(div\_u+1) \times (div\_v+1)$ elements in it. $\mathbf{C^T}$ is the matrix of row vectors on the right hand side of Eq. 7. It contains $(div\_u+1) \times (n+1)$ elements. $\mathbf{P}$ is a matrix of control points. It contains $(n+1) \times (m+1)$ elements. $\mathbf{D}$ is the matrix of column vectors on the right hand side of Eq. 7. It contains $(m+1) \times (div\_v+1)$ elements.

Eq. 8 is simplified to the standard form,

$$
q = [A]p \tag{9}
$$

where
q =Column vector of data points of dimension
    (div_u+1)*(div_v+1),1,
p =Column vector of control points of dimension
    (n+1)*(m+1),1,
[A] =Matrix of dimension
    (div_u+1)*(div_v+1),(n+1)*(m+1).

Matrix [A] is formed as,

$$
A_{I,J} = d_{J,I}[C^t] \tag{10}
$$

where $d_{(J,I)} = (J,I)^{th}$ element of D matrix.

As the surface point data available in molecular docking problem will give a highly over-specified matrix $\mathbf{A}$, system $[\mathbf{A^T A}]\mathbf{p} = [\mathbf{A^T}]\mathbf{q}$ is solved instead of the system $\mathbf{Ap} = \mathbf{q}$, using Cholesky decomposition of matrix $\mathbf{A}$ [12]. Matrix $[\mathbf{A^T A}]$ on the left hand side is a square matrix then. If Cholesky decomposition fails due to matrix

$[\mathbf{A^T A}]$ not being positive definite, a method based on singular value decomposition is used which is applicable to a general rectangular $M \times N$ matrix. For detailed discussion related to the procedure used for modeling the molecular surface [13] can be referred.

3   Molecular Shape Matching

A patch by patch matching strategy is used for shape matching of the two molecules. After solving the inverse problem, the control points of the two surface patches, one from each of the two molecules, are available. The problem of surface matching can now be described as:

1. Given the control points of two surface patches, check whether the patches are complementary.

2. If the patches are complementary, (i) compute the translation required to match two given surface patches at a point and (ii) compute the rotation required to match the surface patches in the same orientation.

Before matching the surface patches of two different molecules, they are checked for complemetarity as described in the following section.

3.1   Check for Complementary Surface patches

To check if the surface patches are complementary, the principal curvatures of both the surface patches are calculated by Eq. 11.

$$
\begin{aligned}
\kappa_{max} &= H + \sqrt{H^2 - K} \\
\kappa_{min} &= H - \sqrt{H^2 - K}.
\end{aligned} \tag{11}
$$

In Eq. 11, H is mean curvature and K is Gaussian curvature [8].

If

$$
\begin{aligned}
\kappa_{max1} &= \text{maximum curvature of the first surface patch at the matching point,} \\
\kappa_{min1} &= \text{minimum curvature the first surface patch at the matching point,} \\
\kappa_{max2} &= \text{maximum curvature of the second surface patch at the matching point,} \\
\kappa_{min2} &= \text{minimum curvature the second surface patch at the matching point,}
\end{aligned}
$$

the surface patches are **not** complementary when $\kappa_{max1} \times \kappa_{max2} > 0$ or $\kappa_{min1} \times \kappa_{min2} > 0$.

If the the surface patches under consideration are found complementary, the next task is to determine the translation and rotation required to superimpose one surface over the other in the same orientation. Superimposing the surfaces gives an idea about the quality of match between the surfaces. This is described in the next section.

3.2 Computing the Translation and Rotation for the Surface Patches

3.2.1 Translation

For matching the surface patches at a given point on both the surface patches, the strategy used here is to translate both the surface patches to the origin of the coordinate frame such that the given matching points on both the surface patches coincide with the origin.

The translation for the first surface patch is computed by the following equation

$$\mathbf{P}_{1ijT} = \mathbf{P}_{1ij} - \mathbf{Q}_1 \qquad (12)$$

where

$$\begin{aligned}
\mathbf{P}_{1ij} \quad &= \quad \text{control points of the first} \\
&\quad\quad \text{surface patch,} \\
-\mathbf{Q}_1 \quad &= \quad \text{translation for the first} \\
&\quad\quad \text{surface patch,} \\
\mathbf{P}_{1ijT} \quad &= \quad \text{changed control points of} \\
&\quad\quad \text{the first surface patch after} \\
&\quad\quad \text{translation .}
\end{aligned}$$

New control points will generate the surface patch in translated position, i.e., at the origin. This is due to the property of B-spline basis functions known as 'partition of unity'.

Translation for the second surface Patch is computed by

$$\mathbf{P}_{2ijT} = \mathbf{P}_{2ij} - \mathbf{Q}_2 \qquad (13)$$

where

$$\begin{aligned}
\mathbf{P}_{2ij} \quad &= \quad \text{control points of the second} \\
&\quad\quad \text{surface patch,} \\
-\mathbf{Q}_2 \quad &= \quad \text{translation for the second} \\
&\quad\quad \text{surface patch,} \\
\mathbf{P}_{2ijT} \quad &= \quad \text{changed control points of} \\
&\quad\quad \text{the second surface patch after} \\
&\quad\quad \text{translation.}
\end{aligned}$$

Changed control points will generate the second surface patch in translated position such that the given matching point on the second surface patch will coincide with the origin and thus, with the matching point on the first surface patch. This completes the translation of surface patches.

3.2.2 Rotation

Translation brings the surface patches at the origin where they meet each other at a point. But it does not change their orientation. The next task is to find the required rotation which will re-orient and superimpose one surface patch over the other. This is required to judge as to how well they match with each other.

The strategy used to match the surface patches in the same orientation is to compute a triad of unit vectors at the given point on both the surface patches and then to match these unit vectors of a triad with the coordinate axes. A triad at the matching point on any of the two surface patches consists of three unit vectors: a unit tangent vector in u parametric direction $\mathbf{q}_u$, a unit normal to the surface patch $\mathbf{n}$ and a unit vector $\mathbf{m}$ normal to both of $\mathbf{q}_u$ and $\mathbf{n}$ (refer [8]). A matrix $\mathbf{V}$ is formed by writing vectors of a triad in the rows of the matrix and a matrix $\mathbf{W}$ is formed similarly by the coordinate axes vectors. Since $\mathbf{V}$ and $\mathbf{W}$ are unit vector matrices, it can be written

$$[\mathbf{x} \ \mathbf{y} \ \mathbf{z}]^T = [\mathbf{q}_{u1} \ \mathbf{m}_1 \ \mathbf{n}_1]^T [\mathbf{R}].$$

Then, required rotation matrix for system $\mathbf{V}$ with respect to system $\mathbf{W}$ is simply

$$[\mathbf{R}] = [\mathbf{V}]^{-1}[\mathbf{W}]. \qquad (14)$$

Using Eq. 14 now, changed control points of the surface patches after rotation can be obtained. For the first surface patch,

$$\mathbf{P}_{1ijR} = \mathbf{P}_{1ijT} \times \mathbf{R}_1 \qquad (15)$$

where

$$\begin{aligned}
\mathbf{R}_1 \quad &= \quad \text{rotation matrix for the first} \\
&\quad\quad \text{surface patch,} \\
\mathbf{P}_{1ijR} \quad &= \quad \text{changed control points of the} \\
&\quad\quad \text{first surface patch after} \\
&\quad\quad \text{rotation.}
\end{aligned}$$

Similarly, for the second surface patch,

$$\mathbf{P}_{2ijR} = \mathbf{P}_{2ijT} \times \mathbf{R}_2 \qquad (16)$$

where

$$\begin{aligned}
\mathbf{R}_2 \quad &= \quad \text{rotation matrix for the second} \\
&\quad\quad \text{surface patch,} \\
\mathbf{P}_{2ijR} \quad &= \quad \text{changed control points of the} \\
&\quad\quad \text{second surface patch after} \\
&\quad\quad \text{rotation.}
\end{aligned}$$

Changed control points will generate the surface patches in rotated position. If both the surface patches are generated with the changed control points and plotted on the same plot, they match in the same orientation.

4 Proposed Algorithm

Algorithm proposed for matching two surface patches at a point and in same orientation is as follows:

Step 1 : Input the 2-D parameterized data points of the surface patches and obtain the control points for the patches as described in section 2.

Step 2 : Compute the principal curvatures of both the surface patches and check for complementarity as in section 3.1.

Step 3 : If surface patches are complementary, input the values of parameters u and v for a point and the control points $\mathbf{P}_{1ij}$ for the first surface patch. Find translation $\mathbf{Q}_1$ by Eq. 1.

Step 4 : Obtain changed control points of the first surface patch after translation $\mathbf{P}_{1ijT}$, using Eq. 12.

Step 5 : Compute triad $(\mathbf{q}_{u1}, \mathbf{m}_1, \mathbf{n}_1)$ and form matrix $\mathbf{V}$ as $[\mathbf{q}_{u1}, \mathbf{m}_1, \mathbf{n}_1]^T$.

Step 6 : If principal curvatures of the first surface patch are positive, form matrix $\mathbf{W}$ as $[\mathbf{x}, \mathbf{y}, \mathbf{z}]^T$ and if they are negative, form matrix $\mathbf{W}$ as $[\mathbf{x}, \mathbf{y}, -\mathbf{z}]^T$.

Step 7 : Find rotation matrix $\mathbf{R}_1$, using Eq. 14.

Step 8 : Obtain changed control points of the first surface patch after rotation $\mathbf{P}_{1ijR}$, using Eq. 15.

Step 9 : Repeat steps 3 to 8 for the second surface patch.

Step 10 : Generate both the surface patches using changed control points after rotation $\mathbf{P}_{1ijR}$ and $\mathbf{P}_{2ijR}$ in Eq. 1. Plot both the surface patches in the same plot to see the match.

5   Simulation Results and Discussion

A few results of testing the algorithm given in the previous section on different types of surface patches are shown here. The surface patches are generated by method given in section 2 using already available 2-D parameterized data sets. The algorithm is, yet, to be tested on actual protein data available in protein data bank. For generation of surfaces, 651 data points have been used in a data set. The curves in u and v parametric directions are cubic. The algorithms are coded in C programmming language and the results are displayed in *gnuplot* in *linux* environment.

Two surfaces patches are shown in Figures 4 and 5 from two different molecules which are to be matched at a given point and in the same orientation. The first surface patch is a concave surface patch while the second surface is a convex one. Figure 6 shows both the surface patches in translated position. The given points of both the surfaces coincide with the origin of the coordinate frame in the translated position. Thus, the surfaces meet each other at the origin. To show this aspect, scale is plotted in the figure. Both the surfaces are shown in rotated position in Figure 7. In the rotated position both the surfaces are matching in the same orientation. Another matching of the two different patches shown in Figures 8 and 9 is displayed in Figure 10. A matching of two flat

surface patches from two different molecules is shown in Figure 11. These results demonstrate the effectiveness of the algorithm for shape matching.
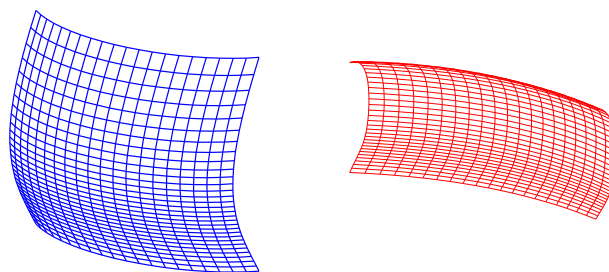

Figure 5:   Surface patch from the second molecule

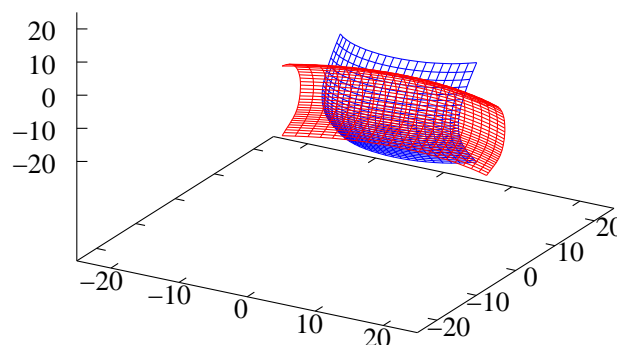Figure 4:   Surface patch from the first molecule

Figure 6: Surface patches translated to the origin (the first step)

6   Conclusions and future work

Shape complementarity based molecular docking criterion depends upon the method to detect the effective match between the molecular surfaces. A simple and effective procedure is proposed in the paper which matches the surface patches after checking their shape complementarity. The B-spline surface representation technique offers the required flexibility for modeling irregular molecular surfaces. Matching strategy uses single point from each of the surface patch. So, problem of non-existence of matching point pairs between the receptor and the ligand does not arise which is the case with some strategies used earlier in the literature. Shape matching based on the matching of triads reduces the number of potential candidate transformations. So, this gives an efficient shape matching algorithm. The algorithm can be easily extended into a complete docking method.
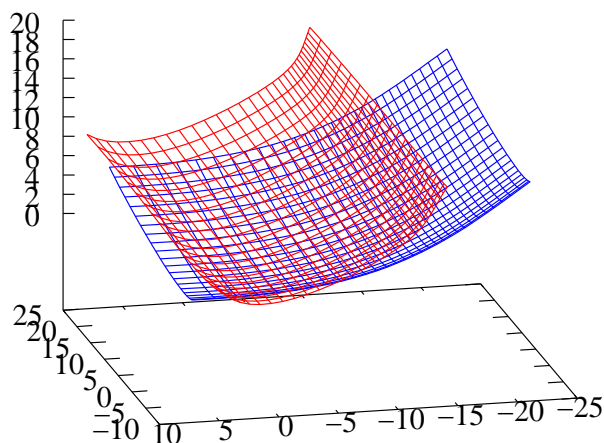
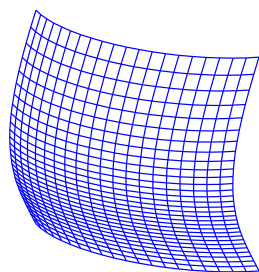Figure 7: Surface patches rotated to match in the same orientation (the second step)

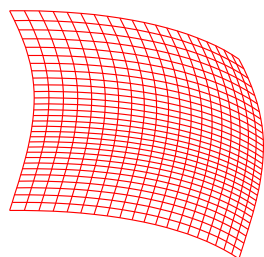Figure 8: Surface patch no. 2 of the first molecule

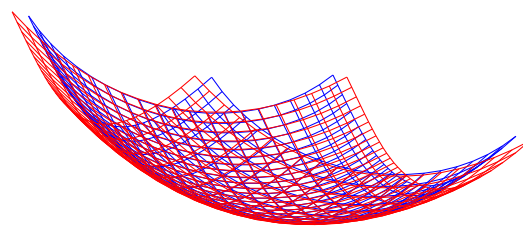Figure 9: Surface patch no. 2 of the second molecule

Figure 10: Surface patch no. 2 of the first molecule matching with patch no. 2 of the second molecule
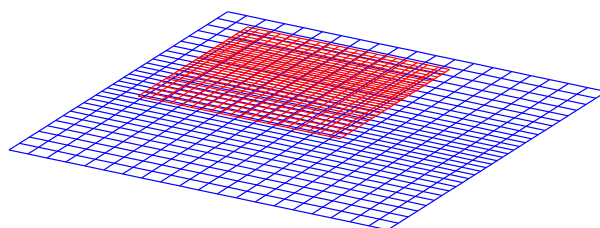
Figure 11: Two flat surface patches from different molecules matching with each other

References

[1] Tsai, C. J., Xu, D., Nussinov, R., "Protein folding via binding and vice versa," *Fold Design* pp. R71–R80 3/98

[2] Abagyan, R., Totrov, M., "High throughput docking for lead generation," *Curr Opin Chem Biol*, pp. 375–382, 5/01

[3] Halperin, I., Ma, B., Wolfson, H., Nussinov, R., "Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions," *Proteins*, pp. 409–433, 47/02

[4] Connolly, M., "Analytical molecular surface calculation," *J. of Appl. Cryst.*, pp. 548–558, 16/83

[5] Lin, S. L., Nussinov, R., Fischer, D., Wolfson, H. J., "Molecular surface representation by sparse critical points," *Proteins*, pp. 94–101, 18/94

[6] Sanner, M. F., Olson, A. J., Spehner, J. C., "Reduced surface: an efficient way to compute molecular surfaces," *Biopolymers*, pp. 305–320, 38/96

[7] Goldman, B. B., Wipke, W. T., "QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock)," *Proteins*, pp. 79–94, 38/2000

[8] Mortenson, M. E., *Geometric Modeling*, John Wiley and Sons (1985)

[9] Tripathi, V. K., Dasgupta, B., Deb, K., "An approach for surface generation from measured data points using B-splines," *Int. Conf. on Systemics, Cybernetics and Informatics*, Hyderabad, India, pp. 578–581, 2005

[10] Kuntz, I., Blaney, J., Oatley, S., Langridge, R., Ferrin, T., "A geometric approach to macromolecule-ligand interactions," *J. of Mol. Biol.*, pp. 269–288, 161/82

[11] Connolly, M., "Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface," *Biopolymers* pp. 1229–1247, 25/86

[12] Kreyszig, E., *Advanced Engineering Mathematics*, 8th Edition, John wiley & sons, inc., New York, 2002

[13] Tripathi, V. K., Dasgupta, B., Deb, K., Voruganti, H. K., "A simple surface representation scheme for rigid body docking," *Nanotech 2006*, Boston, U.S.A. pp. 313–316, 2006