BetaSuperposer: superposition of protein surfaces using beta-shapes

Jae-Kwan Kim and Deok-Soo Kim*

Department of Industrial Engineering and Voronoi Diagram Research Center, Hanyang University, 17 Haengdang-dong, Seongdong-gu, Seoul 133-791, South Korea

Communicated by Ramaswamy H. Sarma

(Received 13 December 2011; final version received 14 March 2012)

The comparison between two protein structures is important for understanding a molecular function. In particular, the comparison of protein surfaces to measure their similarity provides another challenge useful for studying molecular evolution, docking, and drug design. This paper presents an algorithm, called the BetaSuperposer, which evaluates the similarity between the surfaces of two structures using the beta-shape which is a geometric structure derived from the Voronoi diagram of molecule. The algorithm performs iterations of mix-and-match between the beta-shapes of two structures for the optimal superposition from which a similarity measure is computed, where each mix-and-match step attempts to solve an NP-hard problem. The devised heuristic algorithm based on the assignment problem formulation quickly produces a good superposition and an assessment of similarity. The BetaSuperposer was fully implemented and benchmarked against popular programs, the Dali and the Click, using the SCOP models. The BetaSuperposer is freely available to the public from the Voronoi Diagram Research Center (http://voronoi.hanyang.ac.kr).

Keywords: molecule structure; molecule superposition; molecule structure alignment; Voronoi diagram; beta-complex; beta-shape; PDB

1. Introduction

Similar molecules are evolutionarily related. If a new molecule is evolutionarily related with a known molecule, the function of the new molecule can be inferred from that of the known molecule with a degree of accuracy depending on the evolutionary distance between them (Holm & Sander, 1996). In general, there are different ways to define a similarity between two objects. As a simple example, two points in the plane may have different measures of similarity defined by the distance between them. Different metrics (e.g. $l_1$-, $l_2$-, and $l_\infty$-metrics) produce different distances from the two points. Likewise, there are different ways to define similarity between two molecular structures even if the measure of similarity is mathematically well defined. A most immediate similarity is the distance between the sequences of DNAs, RNAs, or proteins. Ever since Needleman and Wunsch used dynamic programming (DP) to measure the sequence similarity between two proteins (Needleman & Wunsch, 1970), a sequence similarity has been very popular for comparing proteins.

It is generally agreed that a structure is more informative in determining a molecular function, because it is more conserved than a sequence during evolution. Hence, a structural similarity can facilitate better understanding of a new molecule by referring to a known molecule whose function has been well studied and it may detect evolutionary links between them that cannot be identified by a sequence analysis only. Therefore, measuring a structural similarity between two molecules is important in understanding molecules. A structural similarity between two molecules in this paper is the similarity between their geometric shapes of three-dimensional molecular structures. There are two different points of view about the geometric similarity. The first is about an entire molecular structure including both the interior and the boundary of a molecule and the second is only about the molecular boundary. In both views, it is necessary to have two proteins appropriately “aligned” and “superposed” before the distance between them is measured.

While a protein structure may experience a considerable topology change during evolution due to the sequence change caused by genetic mutations and crossovers, residues constituting a catalytically active site
tends to be conserved (Chothia & Lesk, 1986; Fischer, Bachar, Nussinov, & Wolfson, 1991; Fowle & Stillman, 1997; Holm & Sander, 1996) and a protein function occurs predominantly on or near a protein surface (Holm & Sander, 1996; Via, Ferré, Brannetta, & Helmer-Citterich, 2000). For example, Fischer et al. showed the well-known similarity between the active sites of trypsin-like and subtilisin-like serine protease families even though they did not fold into similar structures in their entirety (Fischer, Wolfson, Lin, & Nussinov, 1994). Kauvar and Villar reported cases of nonhomologous proteins leading to an identical active site and proteins with a similar structure leading to different active sites with different functions (Kauvar & Villar, 1998). Rost reported the existence of the twilight zone of the sequence alignment and the structure alignment for proteins (Rost, 1999). Hence, the strength of evolutionary persistence is in the order of surface, structure, and sequence and thus measuring the surface similarity can be of priority.

This paper presents an effective and efficient algorithm, called the BetaSuperposer, to measure the similarity between two protein surfaces. We define a match (usually called the “equivalence”) between two protein boundaries and perform a mix (usually called the “transformation”) between their boundaries by rigid-body motions (i.e. translations and rotations) of one with respect to the other. Because a pair of these two steps iterates, we call it a mix-and-match. Each mix-and-match step attempts to solve an NP-hard problem. This implies that there should be two important considerations in the design of an algorithm: (i) How to reduce the problem size and (ii) how to devise a good heuristic. This paper contributes to both. In the previous studies, the first issue was handled by selecting a subset of molecule such as alpha-carbons, backbone atoms, fragments of residues, or secondary structure elements. This paper proposes a more effective and efficient approach of using protein boundary extracted by the beta-shape, which is a derivative structure of the Voronoi diagram. Regarding on the second issue, traditional approaches frequently used DP to compute a new match. This paper introduces a formulation using an assignment problem (AP), a special case of integer linear programming (ILP) problem. The BetaSuperposer algorithm also plays a core role in the recently developed docking software BetaDock by quickly finding a shape-complementarity between a ligand and a pocket on a protein boundary (Kim et al., 2011). The BetaSuperposer was fully implemented and benchmarked against the popular programs Dali (Holm & Park, 2000) and Click (Nguyen, Tan, & Madhusudhan, 2011) using a set of 24 structures from the SCOP database. Even if we describe the algorithm in terms of proteins, we emphasize that the BetaSuperposer can also be used for other macromolecules such as DNAs and RNAs. The BetaSuperposer program, for both 32-bit and 64-bit Linux platforms, is freely available from the Voronoi Diagram Research Center website (http://voronoi. hanyang.ac.kr).

2. Structure comparison vs. sequence comparison

From a computational point of view, comparing protein structures is more complicated than comparing their sequences (Via, Ferré, Brannetta, & Helmer-Citterich, 2000). Sequence comparison problem is one-dimensional (Frenkel, 2008) whereas structure comparison problem is three-dimensional with increased algorithmic complication and computational burden. It is not a surprise in that the alignment and superposition problems in a structure comparison are NP-hard, because many problems related to sequences are already NP-hard. Structure comparison involves both geometry and topology; whereas sequence comparison involves symbols and their relative orders in computational process. Structure comparison requires a systematic representation of topology among objects constituting proteins. Sequence comparison does not have any floating point arithmetic and thus there is no room for numerical error. Structure comparison, on the other hand, almost always involves floating point arithmetic, and therefore the accumulation of round-off errors always matters. Sequence comparison has relatively well-defined measures of similarity and its current main research concern lies in computational efficiency. In fact, a sequence comparison has a theoretical justification based on the extreme value distribution of the maxima of finite sequences of random variables (Coles, 2001). However, we are not aware of its counterpart for structure comparison.

While an amino acid is represented by a symbol in a sequence, it usually consists of dozens of atoms in a molecular structure. Hence, in structure comparison, different methods are used to reduce computational burden depending on an abstraction method used (Kar, Sherman, & Johnson, 1994; Kini & Evans, 1992; Lee, Suh, Kim, & Kim, 1999; Zhang & Wang, 2010). For example, atom groups, residues, secondary structures, tertiary structures, quaternary structures, domains, alpha-carbons, or backbone atoms are different units of abstraction.

Sequence comparison uses sequence data available from UniPROT that can be safely assumed complete and accurate without any noise. On the other hand, structure comparison uses structure data available from PDB, but the data are almost always incomplete and contains noise. The noise in the structure data can cause a significant twist in the topology due to its sensitivity to noise. The main source of the noise is as follows: (i) instances of a protein usually have different conformations, (ii) errors are involved in a measurement, and (iii) a subset of structure is usually missing. Carugo showed that there is a (linear) relationship between the RMSD and the
resolution of PDB files in comparing domains (Carugo, 2003). Figure 1 is a screen capture of PDB website that shows the sequence of Antithrombin (one of the serpin proteins, PDB accession code: 1ant), which has 432 residues. The bar over the sequence denotes the residues existing in the PDB file. In other words, the three intervals marked by the ellipses denote the missing residues (40 residues in total) in the PDB file.

3. Surface comparison vs. entire structure comparison

Early studies of structure comparison were done mostly for entire molecular structure by aligning sequences based on structural criteria primarily based on inter-residue distances. Usually assuming the collinearity of two sequences, DP was used for the minimization of distance between two-aligned sequences based on the inter-residue Euclidean distances. It is important to note that DP preserves the sequence collinearity both before and after the optimization process. Comparing entire structure is usually applied to measure the similarity between proteins that are believed to be already somewhat similar.

Holm and Sander reported the Dali for pairwise alignments of structures based on a residue–residue distance matrix, where each residue was represented by alpha-carbons (Holm & Sander, 1993). Then, similar contact patterns in the matrix were compared by Monte Carlo method. Gerstein and Levitt reported a structural alignment algorithm using the backbones of two structures with iterations of DP and least-squares fitting (Gerstein & Levitt, 1998). Shindyalov and Bourne reported an incremental combinatorial extension algorithm for the alignment between two structures using DP and Monte Carlo simulation (Shindyalov & Bourne, 1998). Wohlers et al. reported an algorithm for the alignment of inter-residue matrices by building upon a maximum contact map overlap (Wohlers, Domingues, & Klau, 2010). Taylor and Orengo proposed a double DP which has two levels of DP running $O(n^3)$ in the worst and average cases where $n$ is the number of residues (Taylor & Orengo, 1989).

A surface comparison is different from the approaches above in that a sequence collinearity should not be a constraint in the minimization of the distance between two structures. This is because surface comparison ignores the correspondence among atoms internal of the boundaries. Hence, a surface comparison problem can be formulated as a combinatorial optimization problem using an ILP minimizing the interdistance for the equivalence set from all possible pairs of atoms in two protein structures. It turns out that determining the
equivalence is a particular case of ILP called the AP that can be solved in \( O(n^3) \) time in the worst case using the Hungarian method (Kuhn, 1955; Edmonds & Karp, 1972), where \( n \) is the number of atoms in the equivalence. Note that an ILP takes an exponential time to solve despite that there is a reliable software like CPLEX (IBM, 2012a). Once the equivalence is determined, the superposition can be done by employing a coordinate transformation. We emphasize the process of determining the equivalence and call it equivalencing.

The reduction of the problem size in a formulation is important. Traditionally, the comparison of an entire structure was done using an abstraction such as alpha-carbons or backbone atoms. However, in the surface comparison, such an abstraction may not be very effective, because it can significantly twist protein surface. Therefore, it is desirable, if not necessary, to use the entire atom set. On the other hand, it is also desirable to use as small a dataset as possible for computational efficiency, and therefore methods to abstract a protein boundary is also necessary. The theory of the beta-shape provides a rigorous and efficient framework for these needs.

4. Beta-shape

Suppose that \( A = \{a_1, a_2, \ldots, a_n\} \) is a protein, where \( a_i = (p_i, r_i) \) is a three-dimensional spherical atom with the center \( p_i \) and radius \( r_i \). Two atoms may intersect, but one cannot contain another. Let \( \text{VC}(a_i) \) denote a Voronoi cell for \( a_i \) defined as \( \text{VC}(a_i) = \{ x \in \mathbb{R}^3 \mid d(x, p_i) - r_i \leq d(x, \ p_j) - r_j, \ i \neq j \} \) where \( d(x, y) \) denotes the Euclidean distance between two points \( x \) and \( y \). Then, the Voronoi diagram \( \text{VD}(A) \) is defined as \( \text{VD}(A) = \{ \text{VC}(a_1), \text{VC}(a_2), \ldots, \text{VC}(a_n) \} \). Note that \( \text{VD}(A) \) is the Voronoi diagram of spheres and is different from the ordinary Voronoi diagram of points. \( \text{VD}(A) \) is represented as a quadruplet \( \text{VD}(A) = (V^A, E^A, F^A, C^A) \) where \( V^A = \{ v_1^A, v_2^A, \ldots, v_k^A \} \), \( E^A = \{ e_1^A, e_2^A, \ldots \} \), \( F^A = \{ f_1^A, f_2^A, \ldots \} \), and \( C^A = \{ c_1^A, c_2^A, \ldots \} \) are the set of vertices, edges, faces, and cells in \( \text{VD}(A) \), respectively. The topology among these entities is also stored in a data structure. A Voronoi vertex is the center of an empty sphere tangent to the boundaries of four nearby atoms; a Voronoi edge is a locus of points equidistant from the boundaries of three nearby atoms; a Voronoi face is the mid-surface between the boundaries of two nearby atoms. Readers are referred to (Kim, Cho, & Kim, 2005; Kim, Cho et al., 2005; Kim, Kim, Cho, & Sugihara, 2006) for more details on the properties and the algorithms.

It is well known that the dual of the Voronoi diagram of points is the Delaunay triangulation and a simplicial complex which can be stored in a compact data structure and the topology traversal among the simplices in the complex is quite easy and efficient (Munkres, 1984; Boissonnat & Yvinec, 1998). However, the dual structure of the Voronoi diagram of atoms, called a quasi-triangulation, is not always a valid triangulation. However, the quasi-triangulation is mostly triangulation and can be stored in Intervoxel data structure after the anomalies are identified. The conversion between the Voronoi diagram and the quasi-triangulation takes \( O(m) \) time in the worst case, where \( m \) is the number of geometric entities in either structure. In the molecular world, \( m = O(n) \) for \( n \) atoms. Hence, the topologies of the Voronoi diagram and the quasi-triangulation are equivalent from both a computational and an informational point of view. For the details of quasi-triangulation and the duality between the Voronoi diagram and the quasi-triangulation, readers are referred to (Kim, Kim, et al., 2006).

The constructs called the beta-shape and beta-complex have been recently proposed, and their properties and computational algorithms are reported in (Kim, Cho, Kim, & Cho, 2006; Kim, Cho, Sugihara, Ryu & Kim, 2010). The beta-hull is defined in a way similar to the alpha-hull (Edelsbrunner & Mücke, 1994). Consider a three-dimensional space filled with a soft matter with some spherical atoms of varying radii scattered within the matter. Carving out the matter with an omnipresent spherical eraser whose radius is \( \beta \) will result in a shape which we call a beta-hull. Since the eraser is omnipresent, there can be interior voids as well. If the spherical eraser is the probe for a water molecule, the boundary of the beta-hull is indeed the molecular surface, often called the Connolly surface (Connolly, 1983a, 1983b). The spherical eraser is called a probe.

Suppose that we have a beta-hull of \( A \). We can straighten the surface of the beta-hull by substituting straight edges for the circular arcs and planar triangles for the spherical triangles, where the vertices are the centers of the atoms contributing to the boundary of the beta-hull. The straightened object bounded by planar facets is the beta-shape of \( A \). In this paper, we consider that the beta-shape is connected to form a single component. Otherwise, each connected component of the beta-shape may be handled separately. The beta-shape is nonmanifold and may have dangling edges, dangling triangles, and even dangling tetrahedra.

We note the following: (i) the boundary of the beta-shape is the exterior boundary of the beta-complex and (ii) the beta-complex is a subset of the quasi-triangulation which lies within the boundary of the corresponding beta-shape. The beta-shape represents the proximity among the atoms on the boundary of a molecule and the beta-complex represents the proximity among all atoms in a molecule.

Above, we described the concepts of a beta-shape and a beta-complex using an infinite number of probes. However, we note here that their computation is not done in this way, but can be done very efficiently via the analysis of shapes of quasi-triangulation simplices with
respect to the size of atoms and the probe. The algorithm to compute the beta-complex (and therefore the beta-shape as well) consists of three steps and is summarized as follows: (i) the computation of the Voronoi diagram of a molecule (taking \(O(n^3)\) time for general spheres and \(O(n^2)\) time for molecules both in the worst case), (ii) the conversion from the Voronoi diagram to the quasi-triangulation (taking \(O(n)\) time in the worst case), and (iii) the search for simplices for the beta-complex (taking \(O(\log n + k)\) time in the worst case, where there are \(k\) simplices in the beta-complex).

We note here that the first step, the computation of the Voronoi diagram for all proteins we tested in PDB, shows empirically \(O(n)\) time for \(n\) atoms. We want to note that the Voronoi diagram and the quasi-triangulation are used for various applications including one reported in this paper. Hence, the first two steps, the computation of the Voronoi diagram and the corresponding quasi-triangulation, can be done offline as pre-processing and stored in the database. Therefore, the computation of the beta-complex corresponding to a particular value of \(\beta\) requires the third step only.

Figure 2(a) shows a two-dimensional molecule \(A\) consisting of nine two-dimensional atoms, and Figure 2(b) shows the vdw-boundary of \(A\). Note that the vdw-boundary contains two internal voids, a few shallow depressions, and a deep depression on the exterior boundary of the model. Figure 2(c) is the molecular surface of \(A\) corresponding to a small probe (shown in the black circle). Note that only one interior void is left in Figure 2(c), which corresponds to the larger void in the vdw-boundary of Figure 2(b). The void corresponding to the smaller one in the vdw-boundary has disappeared. Note also that the shape of the void in the molecular surface is obtained from the beta-complex.

![Figure 2](image.png)

Figure 2. Protein, the boundaries of the protein, the beta-shape, and the beta-complex. (a) a set of nine atoms in \(R^2\), (b) the vdw-boundary, (c) the molecular surface corresponding to a small probe, (d) the corresponding beta-shape, (e) the corresponding beta-complex, (f) a beta-shape corresponding to a larger probe, (g) the corresponding beta-complex with simplices removed from the quasi-triangulation, (h) the protein hormone receptor (1nav, 1984 atoms), and (i) the beta-shape of 1nav corresponding to \(\beta = 1.4\ \text{Å}\).
surface is different from that of the vdw-boundary. Hence, the molecular surface removes some local shape characteristics which are usually unnecessarily detailed for applications such as finding drug candidates. Regions where a water molecule cannot fit may not usually interact with other meaningful chemicals. Figure 2(d) shows the beta-shape corresponding to the molecular surface of Figure 2(c) where the shaded region is the interior of the beta-shape. The beta-shape in Figure 2(d) has an interior void corresponding to the void of the molecular surface and a dangling edge corresponding to a pair of atoms that are exposed to or touched by the probe. The dangling edge corresponds to the deep depression in the external boundary of the molecular surface. The boundary of the beta-shape has nine vertices and nine edges (six on the exterior boundary and three on the interior void). Figure 2(e) shows the beta-complex corresponding to the beta-shape in Figure 2(d). Note that each vertex of the beta-shape and beta-complex corresponds to an atom. Figure 2(f) and (g) show the beta-shape and beta-complex corresponding to a larger probe, respectively. Note that both the dangling edge and the internal void in the beta-shape of Figure 2(d) have now disappeared. The dotted line segments in Figure 2(g) form the quasi-triangulation of the molecule A together with the beta-complex in Figure 2(g). Figure 2(h) shows a three-dimensional protein model, hormone receptor (PDB id: 1nav), consisting of 1984 atoms, and Figure 2(i) shows the beta-shape of 1nav corresponding to a probe of the radius 1.4 Å, i.e. a water molecule. The white spot in the middle of the beta-shape in fact denotes a tunnel passing through the molecular surface of the protein. Note that the three-dimensional beta-shape is not necessarily manifold, either. Details on the beta-shape and beta-complex can be found in (Kim et al., 2010; Kim, Seo, Kim, Ryu, & Cho, 2006).

5. Abstraction of protein boundary via beta-shape

Being an abstraction of protein boundary corresponding to a β-probe, the beta-shape is a good method to extract a smaller set of atoms that can represent protein boundary. The number of the vertices on the beta-shape can be significantly smaller than the number of atoms in the entire protein if the β-value is properly chosen. Let \( n \) be the number of protein atoms and \( n_β \) be the number of vertices on the beta-shape boundary corresponding to \( β \). In other words, \( n_β \) is the number of boundary atoms touchable by a probe of radius \( β \). Let \( r_β = \frac{n}{2π} \). We have selected a test set of 100 PDB files (see Table 1 in the Supplementary material for details). For a given β-value, we computed \( r_β \) for each of the 100 test proteins, where the β-value is incrementally given in the range of \( 0 \leq β \leq 1000 \text{Å} \). Let \( μ(r_β) \) be the average of the 100 \( r_β \)'s at each β-value. Black circles in Figure 3 denote the curve of \( μ(r_β) \) vs. \( β \) for \( 0 \leq β \leq 15 \text{Å} \). Note that \( μ(r_β) \) shows a very smooth, nonincreasing function.

The nonlinear regression analysis using the statistical package R (The R Project, 2012) yields the regression curve of \( μ(r_β) \) as

\[
\hat{μ}(r_β) = \frac{2.1710}{(β + 1.1843)^{13.89e} + 0.519}
\]

which fits the \( μ(r_β) \) curve with the residual sum of squares \( 0.4614 \). We employ the following

\[
\hat{μ}(r_β) = \frac{1}{(β + 1.0)^2} + 0.05
\]

to get a simpler and more intuitive formula approximating Equation (1). Further analysis finds that \( α = 1 \) fits the dataset quite well. Figure 3 shows curves of Equation (2) with different values of \( α \). Hence, a good approximation of \( N_β \) for a protein with \( n \) atoms can be quickly obtained by

\[
N_β ≈ \hat{μ}(r_β) \times n.
\]

For example, for water molecule that is approximated by a spherical probe with the radius 1.4 Å, \( N_{1.4} ≈ 4.7n \). Similarly, \( N_{2.0} ≈ 38n \), \( N_{3.0} ≈ 30m \), etc. We also conducted the analysis of variance of the 100 \( r_β \)'s for each value of \( β \) between 0 and 1000 Å. Figure 4 shows the standard deviation \( σ \) of the 100 \( r_β \)'s vs. \( β \) where \( 0 \leq β \leq 15 \text{Å} \). The maximum of \( σ = 0.081 \) occurs at \( β = 1.9 \text{Å} \) and monotonically converges to an asymptotic value of .027.

6. Mix-and-match

6.1. Mix: a rigid-body transformation

Definition 1

Let \( A = \{a_1, a_2, ..., a_n\} \) and \( B = \{b_1, b_2, ..., b_n\} \) be ordered sets. An equivalence is \( E(A, B) = \{(a_1, b_1), (a_2, b_2), ..., (a_n, b_n)\} \). The process to determine \( E(A, B) \) is called the equivalencing. The equivalence is an alignment between \( A \) and \( B \).

Definition 2

Let \( \tilde{A} = \{a_{i1}, a_{i2}, ..., a_{ir}\} \) and \( \tilde{B} = \{b_{j1}, b_{j2}, ..., b_{jr}\} \) where \( \tilde{A} \subseteq A \) and \( \tilde{B} \subseteq B \). If \( i_1 < i_2 < ... < i_r \) and \( j_1 < j_2 < ... < j_r \), \( \tilde{A} \) and \( \tilde{B} \) are called collinear.

Definition 3

Given \( E(A, B) \), the transformation of \( \tilde{A} \) towards a fixed \( B \) such that the distance defined on \( E(A, B) \) is minimized is called the superposition.

Hence, an equivalence is the precondition of superposition. The problem to find the optimal alignment of a protein sequence to a known protein structure is called the
Table 1. The test dataset consisting of 24 structures from SCOP.

| Class (A)         | Fold (B)          | Superfamily (C)     | Family (D) | PDB (E) | Codein Test (F) | Chain (E2) | Resolution (E3) | #Residue (E4) | #Atom (E5) |
|-------------------|-------------------|---------------------|------------|---------|-----------------|------------|-----------------|--------------|-----------|
| Alpha             | Globin-like       | Globin-like         | S1         | Globins | 1B0B            | A          | 1.43            | 141          | 1032      | 1          |
|                   |                   |                     |            |         | 1CG5            | A          | 1.60            | 141          | 1116      | 2          |
|                   |                   |                     |            |         | 1FSL            | A          | 2.30            | 143          | 2173      | 3          |
|                   |                   |                     |            |         |                 |            |                 |             |           |            |
|                   |                   | Truncated hemoglobin|            |         | 1DLW            | A          | 1.54            | 116          | 837       | 4          |
|                   |                   |                     |            |         | 1DLY            | A          | 1.80            | 121          | 935       | 5          |
|                   |                   |                     |            |         | 1HDR            | A          | 1.90            | 127          | 955       | 6          |
| Beta              | Immuno-globulin-like_betasandwich | Immuno-globulin | S2         | C1 set domains | 1DN0 | B | 2.28   | 216 | 1657 | 7 |
|                   |                   |                     |            |         | 1ED3            | A          | 2.55            | 275          | 2260      | 8          |
|                   |                   |                     |            |         | 1G84            | A          | –               | 105          | 1587      | 9          |
|                   |                   |                     |            |         |                 |            |                 |             |           |            |
|                   |                   | V set domains       |            |         | 1A3R            | H          | 2.10            | 207          | 1621      | 10         |
|                   |                   |                     |            |         | 1ADQ            | L          | 3.15            | 210          | 1587      | 11         |
|                   |                   |                     |            |         | 1AXT            | H          | 2.15            | 212          | 1675      | 12         |
| Alpha/beta        | TIM betalpha-barrel | Ribulose-phosphate binding barrel | S3         | D-ribulose-5-phosphate epimerase | 1H1Y | A | 1.87   | 220 | 1641 | 13 |
|                   |                   |                     |            |         | 1RPX            | A          | 2.30            | 230          | 1731      | 14         |
|                   |                   |                     |            |         | 1TQJ            | A          | 1.60            | 217          | 1640      | 15         |
|                   |                   | Histidine biosynthesis enzymes |            |         | 1H5Y            | A          | 2.00            | 252          | 1875      | 16         |
|                   |                   |                     |            |         | 1KA9            | F          | 2.30            | 251          | 1882      | 17         |
|                   |                   |                     |            |         | 1QQ2            | A          | 1.85            | 240          | 1890      | 18         |
| Alpha + beta      | Microbial ribonucleases | Microbial ribonucleases | S4         | Bacterial ribonucleases | 1A2P | A | 1.50   | 108 | 863  | 19 |
|                   |                   |                     |            |         | 1GOU            | A          | 1.65            | 109          | 863       | 20         |
|                   |                   |                     |            |         | 1PYL            | A          | 1.51            | 96           | 1476      | 21         |
|                   |                   | Fungal ribonucleases |            |         | 1FUS            | A          | 1.30            | 105          | 764       | 22         |
|                   |                   |                     |            |         | 1H0V            | A          | 1.23            | 104          | 770       | 23         |
|                   |                   |                     |            |         | 1RMS            | A          | 1.90            | 105          | 802       | 24         |

Note: The unit of the resolution is Å.
threading and was shown NP-hard (Lathrop, 1994; Lathrop & Smith, 1996). In structure, even the alignment problem is NP-hard (Eidhammer & Jonassen, & Taylor, 2004, p. 187). However, a surface comparison problem assumes no prior knowledge of what amino acids are equivalent and thus this problem is even harder. Suppose that two proteins \(A\) and \(B\) are given and \(E(A, B)\) is known. Let \(E(A, B) = \{(\alpha_1, \beta_1), (\alpha_2, \beta_2), \ldots, (\alpha_r, \beta_r)\}\). Suppose that \(\alpha_i\) and \(\beta_i\) represent the abstractions of substructures of \(A\) and \(B\), respectively. For notational simplicity, suppose that \(\alpha_i\) and \(\beta_i\) also denote their center coordinates. Assuming \(B\) fixed, we move \(A\) via a rigid-body transformation. Structure superposition finds the optimal transformation so that the distance \(\Delta\) between \(A\) and \(B\) is minimized, where \(\Delta = \sum_k \delta_k, \delta = ||\alpha_k - \beta_k||, 1 \leq k \leq r\). The formalization of the superposition problem follows.

**Problem 1.** Find the transformation \(T\) such that

\[
\text{Minimize RMSD} \quad (4)
\]

where

\[
\text{RMSD} = \sqrt{\sum_{i=1}^{r} \sum_{j=1}^{r} w_i (T\alpha_i - \beta_j)^2, \alpha_i \in A, \beta_j \in B} \quad (5)
\]

The best result can be obtained when \(r = m \leq n\) in Equation (5). However, Equation (4) is NP-hard and is not tractable even for moderate-sized proteins. Suppose that

\[
\sum_{i=1}^{r} (T\alpha_i - \beta_j)^2 = \sum_{i=1}^{r} (R\alpha_i + t - \beta_j)^2 \quad (6)
\]

where \(R\) and \(t\) are a rotation matrix and a translation vector, respectively. Hence, Equation (5) minimizes when Equation (6) minimizes. If \(t\) is appropriately chosen and fixed, the problem can be reduced to a problem to find an optimal rotation \(R\). The most common, reasonable practice for \(t\) is to use the difference vector between the centers of masses of atom sets in \(E(A, B)\). Then, Equation (6) reduces to find the rotation \(R\) such that
Minimize \[ \sum_{i=1}^{r} (\alpha_i - \beta_i)^2 \] \hspace{1cm} (7)

where \( \alpha_i \) is \( \alpha_i \) after the translation by \( t \). Equation (7) solves Equation (5) provided that \( t \) is sufficiently good and \( w_j = 1 \). Therefore, the problem now reduces to find a rotation matrix \( R \) in Equation (7). Because the choice of \( t \) is heuristic, an iterative process (between the equivalence set selection and superposition) is usually employed to obtain a satisfactory solution of Equation (6). There are different ways to to find \( R \): Diamond determined the best transformation between two sets of coordinates and factorized it into a symmetric and an orthogonal matrices (Diamond, 1976); Kabsch determined the best rotation of one point set into a second point set by minimizing the weighted sum of squared deviations (Kabsch, 1976); Hendrickson determined the linear orthogonal transformation to superpose two point sets (Hendrickson, 1979); Kearsley solved the problem by minimizing the sum of squared distances between corresponding points for two point sets using a constrained least-squares method solved as an eigenvalue problem in quaternion (Kearsley, 1989).

6.2. Match: an equivalencing
Suppose \( A \) and \( B \) are proteins consisting of different number of atoms and positioned in a common coordinate system. In a surface comparison, there can be two ways to find an equivalence \( E(A,B) \): the DP approach and the AP formulation approach. Let the \( d_{ij} \) be the distance between \( a_i \in A \) and \( b_j \in B \). Consider the DP formulation of the problem

\[ H_{ij} = \text{Maximize} \{ H_{i-1,j}, H_{i,j-1}, H_{i-1,j-1} - d_{ij} \} \] \hspace{1cm} (8)

where \( H_{i,j} \) is the score function which can be achieved by aligning \( A \) and \( B \). The time complexity of DP approach is \( O(n^2) \) in both average and worst cases. However, the equivalence \( E(A,B) \) thus found is constrained so that the subsets \( A \) and \( B \) are collinear.

The problem can also be formulated as an ILP problem as follows.

Minimize \[ \sum_i \sum_j d_{ij}x_{ij} \]

Subject to \[ \sum_j x_{ij} = 1, \quad \text{for all } j, 1 \leq j \leq r \leq n \]

\[ \sum_i x_{ij} = 1, \quad \text{for all } i, 1 \leq i \leq r \leq m \] \hspace{1cm} (9)

\[ x_{ij} = \begin{cases} 1, & \text{if } a_i \text{ and } b_j \text{ are aligned;} \\ 0, & \text{otherwise.} \end{cases} \]

\( E(A,B) \) found by Equation (9) is globally minimal without collinearity. While solving Equation (9), using an optimization library such as CPLEX takes an exponential time, this problem is the AP that can be solved by Hungarian method in \( O(n^3) \) time in the worst case.

7. Experiments and discussion
The proposed algorithm was validated through an extensive experiment, was implemented into a program called the BetaSuperposer. Then, the performance of the BetaSuperposer was compared with two other methods: the Dali (Holm & Park, 2000) and the Click (Nguyen, Tan, & Madhusudan, 2011). Three experiments were conducted for the following purposes: (i) to validate the superiority of an AP formulation to a DP; (ii) to validate the superiority of the beta-shape abstraction of structures; (iii) the benchmark test. For the first two experiments, we created a test set consisting of 100 arbitrary PDB structures. For the benchmark test, we created a test set consisting of 24 PDB structures from four different superfamilies in the SCOP database so that each of the four classes (i.e. alpha, beta, alpha/beta, and alpha + beta) had a representative superfamily in the test set (Andreeva et al., 2008; Hubbard, Ailey, Brenner, Murzin, & Chothia, 1999). All metals and co-factors were removed from the structures before conducting the experiments. We used the cluster computer of the Voronoi Diagram Research Center consisting of 118 nodes; each node has two CPU’s with 2 GB RAM where each CPU has AMD Opteron Dual Core 2.2 GHz.

7.1. Why an AP formulation?
To compare the AP and DP approach, we performed an experiment with a test set of arbitrarily chosen 100 structures from PDB. Given the test structures in the order of their sizes, we first chose the 50th in the set (1QXH) as the reference structure. Then, we compared all the structures in the test set against the reference structure. Figure 5 shows the comparison between DP and AP approaches for \( E(A,B) \) after 100 iterations (for the experimental purpose) of the mix-and-match. Figure 5(a) and (b) show the computation time and the corresponding RMSD, respectively, where circles denote the AP approach and the triangles denote the DP approach. In both figures, the horizontal axes denote proteins in the order of molecular size. Figure 5(a) shows that AP takes more computation than DP and Figure 5(b) shows that the RMSD of AP is significantly smaller than that of DP. This is because the worst-case time complexity of AP is \( O(n^3) \), whereas one for DP is \( O(n^2) \). Note that AP finds the global optimal solution. We note the performance of the AP approach around where the reference structure is located (This means that the surface comparison is done between similar sized proteins). Note that the AP approach takes only slightly more computation while its
result are significantly better in RMSD than those of the DP approach.

7.2. Why the beta-shape abstraction?

Comparison of different abstraction methods is in order. Abstraction method influences the size of model and so does on solution quality and computation time. We selected three abstraction methods to compare two proteins: all atoms, alpha-carbons, and the beta-shape. Figure 6(a) shows the computation time of the three different abstraction methods using the AP approach taken by 100 iterations of the mix-and-match. The rectangles (topmost) denote using all atoms; the circles (in the middle) denote the beta-shape abstraction where $\beta = 1.4$ Å; the triangles (bottommost) denote using alpha-carbons. Note that the beta-shape abstraction takes significantly smaller computation time than using all atoms while it still takes slightly more computation than using alpha-carbons. Figure 6(b) shows the solution quality of the three abstraction methods (measured in RMSD that is defined by the distance between the pair of atoms in the equivalence set at the final iteration). The symbols used in Figure 6(a) apply in Figure 6(b), too. The quality of solution using all atoms is obviously the best and the alpha-carbon atoms approach shows the worst quality. Figure 6(c) shows a more clear comparison among solution quality of the three abstraction methods. Let $\text{RMSD}(\beta - \text{all})$ denote the RMSD of the beta-shape method minus the RMSD of the all atoms method and $\text{RMSD}(\alpha - \text{all})$ denote the RMSD of the alpha-carbon method minus the RMSD of the all atoms method. In Figure 6(c), the circle denotes $\text{RMSD}(\beta - \text{all})$ (which has the mean value .508 and the standard deviation .827) and the triangle denotes $\text{RMSD}(\alpha - \text{all})$ (which has the mean value 1.233 and the standard deviation 2.315). Note that the circles are much more distributed around the horizontal axis than the triangles. To summarize, the quality of solution using the beta-shape shows the solution quality close to the quality of all atoms despite that this approach requires much less computational requirement. For the details of data, see Figure 1 in the Supplementary material.

Figure 7 shows the surface superposition computed by the AP algorithm using the beta-shapes of different $\beta$-values. Figure 7(a) shows the computation time for the 100 structures after 100 iterations. The topmost, rectangles denote when $\beta = 0$ Å where the number of vertices on the beta-shape boundary is almost always identical to the number of entire atoms in proteins. As the value of $\beta$ increases, the number of vertices on the beta-shape boundary decreases and so does the computation time. The circles correspond to $\beta = 1.4$ Å; the triangles correspond to $\beta = 3.0$ Å, etc. Figure 7(b) shows the computation results in RMSD. Note that the RMSD for larger $\beta$-value is greater than that of a smaller $\beta$-value.

Figure 7(c) shows a more clear comparison among solution quality among different $\beta$ values. In the figure, there are four curves: the circle denotes $\text{RMSD}(\beta_{1.4} - \beta_{0})$ (which has the mean value .508 and the standard deviation .827), the triangle denotes $\text{RMSD}(\beta_{3.0} - \beta_{0})$ (which has the mean value .230 and the standard deviation .438), the cross denotes $\text{RMSD}(\beta_{5.0} - \beta_{0})$ (which has the mean value .285 and the standard deviation .744), and the diamond denotes $\text{RMSD}(\beta_{10.0} - \beta_{0})$ (which has the mean value .509 and the standard deviation .843). Hence, $\beta = 3.0$ seems a good choice of the probe radius for the surface comparison using the beta-shape. See Figure 2 in the Supplementary material for details.

This experiment verifies the following two points: (i) AP approach is much better suited than the DP approach and (ii) the beta-shape abstraction is very well suited for the surface comparison between proteins when the value of $\beta$ is appropriately chosen.
7.3. Benchmark test

The test set for the benchmark test consisted of 24 PDB structures from four different superfamilies in the SCOP database so that each of the four classes (i.e. alpha, beta, alpha/beta, and alpha + beta) had a representative superfamily in the test set. We chose two different families from each superfamily and three different structures from each family, where each structure was a chain in the corresponding protein. The details of the selected structures (e.g. the name of the superfamilies, the PDB IDs, the resolution, and the numbers of residues and atoms) are given in the Table 1. The structure 1G84 (Test Code: 9) does not have the resolution value, because it was determined from an NMR data. With this dataset, we conducted all pairwise comparisons using the three programs: the BetaSuperposer, the Dali, and the Click. Note that both Dali and Click use alpha-carbons only for the structure comparison whereas the BetaSuperposer uses boundary atoms. The comparison between two structures belonging to one superfamily is called an “intra-superfamily” comparison. Otherwise, it is called an “inter-superfamily” comparison. We note that the BetaSuperposer has two system parameters which are based on the previous two experiments. First, the $\beta$-value, corresponding to the probe radius, controls the problem size in that it is related with the number of boundary atoms. Second, the number of mix-and-match iterations controls the computation time taken for the superposition. In this experiment, we used $\beta = 1.4 \text{Å}$ and 50 mix-and-match iterations for each pair of structures.

Figure 8 shows the solution quality (measured in RMSD) after the superposition of all structure pairs. Figure 8(a)–(c) shows a bird’s eye view and Figure 8(d)–(f) shows a top view. In the figures, the numbers in the column and row correspond to the test codes in the Column F of the Table 1. Hence, each distinct structure pair corresponds to a pixel in the top view. The RMSD of each structure pair is depicted by the height in the bird’s eye view and by the intensity in the top view. The darker a pixel is in the top view, the smaller the corresponding RMSD value is. Figure 8(a) and (d) are for the BetaSuperposer; Figure 8(b) and (e) are for the Click; Figure 8(c) and (f) are for the Dali. Note that each figure is symmetric around the diagonal.
The Click (Figure 8(b) and (e)) had an extremely narrow range of the RMSD (i.e. 0–2.510 Å) regardless that a pair consists of structures from same or different superfamily. Hence, the Click could not distinguish whether it was an inter- or intra-superfamily comparison. On the other hand, the Dali (Figure 8(c) and (f)) had a reasonable range of the RMSD (i.e. 0–11.4 Å) for intra-superfamily comparisons. However, the Dali did not produce any result for inter-superfamily comparisons. The BetaSuperposer (Figure 8(a) and (d)) showed the following properties: (i) it produced a reasonable range of the RMSD (i.e. 0–14.816 Å) similar to the Dali for intra-superfamily comparisons and ii) it discriminated inter-superfamily comparisons from intra-superfamily ones with a quantitative measure of structural similarity.

Figure 9 provides visual examples of the observation above. Figure 9(a)–(c) are 1DLY (Test Code: 5, Superfamily Code: S1, Class: Alpha) in Globin-like superfamily, 1DN0 (Test Code: 7, Superfamily Code: S2, Class: Beta) in Immunoglobulin superfamily, and 1TQJ (Test Code: 15, Superfamily Code: S3, Class: Alpha/Beta) in Ribulose-phosphate binding barrel superfamily. In each figure, the left shows an entire structure whereas the right shows its backbone. All three structures belong to different superfamilies. However, the two structures in Figure 9(a) and (c) have relatively similar shapes whereas those in Figure 9(b) and (c) have significantly different shapes.

From the structure comparison between Figure 9(a) and (c), the BetaSuperposer and the Click produced 2.580 and 2.180 Å, respectively. Note that both structures in the Figure 9(a) and (c) have globular shapes despite that they have some difference in the number of atoms. From the structure comparison between Figure 9(b) and (c), the BetaSuperposer and the Click produced 12.368 and 2.310 Å, respectively. Note that both structures in the Figure 9(b) and (c) have significantly different shapes whereas they have similar number of atoms. The Dali did not produce any result for both comparisons because two structures belong to different superfamilies. Hence, this example verifies the computation result in Figure 8.

Recall that the RMSD values from the Click and Dali were computed from pairs of alpha-carbons, whereas the RMSD from the BetaSuperposer was computed from pairs of atoms. It is desirable to produce the RMSDs in a unified measure. Hence, we recomputed the
RMSDs of the Click and Dali by all atom pairs by solving the AP. Figure 10 shows the results of the re-computation and Table 2 reports the detailed statistics.

Figure 8. Similarities between two structures measured in RMSD after the superposition using the three softwares. (a) and (d) from the BetaSuperposer, (b) and (e) from the Click, and (c) and (f) from the Dali.

Figure 9. Structures in different superfamilies. (a) 1DLY (the test code: 5) in Globin-like superfamily, (b) 1DN0 (the test code: 7) in Immunoglobulin superfamily, and (c) 1TQJ (the test code: 15) in Ribulose-phosphate binding barrel superfamily.

RMSDs of the Click and Dali by all atom pairs by solving the AP. Figure 10 shows the results of the re-computation and Table 2 reports the detailed statistics.

Figure 10(a) and (d) are identical to those in Figure 8(a) and (d), respectively. Figure 10(b) and (e) applies to the Click, and Figure 10(c) and (f) applies to the Dali.
The observations are as follows: (i) the range of the RMSD values of the BetaSuperposer (i.e. 0–14.816 Å) is narrower than those of the Click (i.e. 0–34.889 Å) and the Dali (i.e. 0–25.465 Å); (ii) within each superfamily, the BetaSuperposer has significantly lower RMSD values (for both the average and the standard deviation) than the other two programs, except the superfamily S3.

Refer to the Table 2. We further compared the performance of the BetaSuperposer with that of the Dali, because it had been most frequently used. For the BetaSuperposer, the averages of each superfamily were 2.558, 5.031, 3.281, and 2.644 Å, respectively, and thus the grand average was 3.379 Å. For the Dali, the averages of each superfamily were 2.835, 10.939, 2.696, and 3.275, respectively, and the grand average was 4.936 Å. Hence, we claim that the BetaSuperposer produced lower RMSDs than the Dali did for the intra-superfamily comparison. We verified this claim through the Wilcoxon signed rank test (Wilcoxon, 1945), which was a nonparametric statistical hypothesis test, for comparing two RMSDs (i.e. one from the BetaSuperposer and the other from the Dali) for each structure pair. This test assesses whether the population mean ranks of the BetaSuperposer and those of the Dali differ.

Each superfamily consisted of six structures and thus there were 15 structure pairs to be compared by each software for the intra-superfamily comparison. Therefore, there were 60 structure pairs in total to be compared by each software. We used the statistical package SPSS (IBM, 2012b) with the 60 RMSDs for each of the BetaSuperposer and the Dali for the test and obtained the p-value of .001. This p-value implied that the distributions of RMSDs for the two softwares were significantly different from the statistical point of view. Hence, we claim that the BetaSuperposer is superior to the Dali.

For each of the four types of the intra-superfamily comparisons, we show two comparison examples with graphical visualization in the Figures 3–10 of the Supplementary material. Figure 3 of the Supplementary material shows the superposition of 1FSL (Test Code: 3) and 1DLW (Test Code: 4) in the superfamily S1. Figure 3 (a) and (b) are for the BetaSuperposer and the Dali, respectively. In each figure, the left, the middle, and the right columns denote three views from X-, Y-, and Z-directions, respectively; the upper row shows the Connolly surface of the two superposed structures (with respect to water probe); the lower row shows the ribbon diagram of the two superposed structures. All the other figures are given with the same convention. The Figures
3 and 4 are for the superfamily S1; The Figures 5 and 6 are for the superfamily S2; The Figures 7 and 8 are for the superfamily S3; and The Figures 9 and 10 are for the superfamily S4; In addition to its superiority for comparing the boundaries, we make the following observations regarding on the secondary structure:

- In general, the BetaSuperposer results in a better fit for the boundary, whereas the Dali results in a better fit for the secondary structure. This is shown in the superfamily S1 (Figures 3 and 4).
- The BetaSuperposer results in a better fit for the secondary structure than the Dali when the structures are far from globular. This is shown in the superfamily S2 (Figures 5 and 6).
- Suppose that the structures are globular and have one or more flapping loop on the structure boundary. If the size of the structures is sufficiently big, the Dali results in a better fit for the secondary structure than the BetaSuperposer does. This is because of the following reason: the influence of the flapping loop is significant for the BetaSuperposer because it considers the boundary only. On the other hand, this influence for the Dali is less significant if the structure size is sufficiently big. This is shown in the superfamily S3 (Figures 7 and 8).
- Suppose that the structures are globular and have one or more flapping loop. If the structure size is sufficiently small, the BetaSuperposer results in a better fit for the secondary structure than the Dali does. This is because the influence of the loop has relatively is larger for the Dali when the structure size becomes smaller.

### Conclusions

A molecular structure is important for determining its function. Because protein functions occur predominantly on or near protein surfaces, the geometric appearance or shape is important for many applications. Hence, the comparison of protein surfaces is important for both understanding molecular evolution and function and for drug design (Chang et al., 2011; Huang et al., 2010).

In this paper, we proposed an effective and efficient algorithm, called the BetaSuperposer, to measure the similarity between two protein surfaces based on the beta-shape which is a derivative structure from the Voronoi diagram of atoms in molecules. The idea of the BetaSuperposer was the mix-and-match among atoms in two protein structures where each mix-and-match step attempted to solve an NP-hard problem. Therefore, we used the beta-shape as a good abstraction of protein boundary thus significantly reducing problem size. This is because the number of vertices on the beta-shape
boundary is significantly smaller than other abstraction such as all atoms or alpha-carbons. We have also proposed the AP formulation to find an equivalence, because the collinearity is not a constraint in surface comparison of proteins.

The BetaSuperposer was fully implemented and benchmarked against popular programs, the Dali and the Click. The core algorithms were extensively tested with a set of 100 PDB structures. The AP formulation was superior to the ordinary DP approach for the match step. The beta-shape was superior to the other abstraction methods such as alpha-carbons or backbone atoms. The benchmark test revealed that the BetaSuperposer outperformed the Dali and the Click from the experiment using a set of 24 structures (six from each of the four superfamilies in the SCOP database). The BetaSuperposer is freely available to the public from the Voronoi Diagram Research Center (http://voronoi.hanyang.ac.kr).

Acknowledgment
This work was supported by the National Research Lab grant funded by the Ministry of Education, Science, and Technology, Korea [No. 2011-0020410].

Supplementary material
The supplementary material for this paper is available online at http://dx.doi.10.1080/07391102.2012.689700.

References
Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S.E., Hubbard, T.J.P., Chothia, C., & Murzin, A.G. (2008). Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Research*, 36, D419–D425.

Boissonnat, J.-D., & Yvinec, M. (1998). Algorithmic geometry. Cambridge: Cambridge University Press.

Carugo, O. (2003). How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. *Journal of Applied Crystallography*, 36, 125–128.

Chang, T.-T., Sun, M.-F., Chen, H.-Y., Tsai, F.-J., Fisher, M., Lin, J.-G., & Chen, C.Y.-C. (2011). Screening from the world’s largest tcm database against h1n1 virus. *Journal of Biomolecular Structure & Dynamics*, 28, 773–786.

Chothia, C., & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5, 823–826.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. London: Springer-Verlag.

Connolly, M.L. (1983a). Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16, 548–558.

Connolly, M.L. (1983b). Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221, 709–713.

Diamond, R. (1976). On the comparison of conformations using linear and quadratic transformations. *Acta Crystallographica Section A*, A32, 1–10.

Edelsbrunner, H., & Mücke, E.P. (1994). Three-dimensional alpha shapes. *ACM Transactions on Graphics*, 13, 43–72.

Edmonds, J., & Karp, R.M. (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the Association for Computing Machinery*, 19, 248–264.

Eidhammer, I., Jonassen, I., & Taylor, W.R. (2004). Protein bioinformatics—an algorithmic approach to sequence and structure analysis. Chichester: Wiley.

Fischer, D., Bachar, O., Nussinov, R., & Wolfson, H.J. (1991). An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *Journal of Biomolecular Structure & Dynamics*, 9, 769–789.

Fischer, D., Wolfson, H., Lin, S.L., & Nussinov, R. (1994). Three-dimensional, sequence order – independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: Potential implications to evolution and to protein folding. *Protein Science*, 3, 769–778.

Fowle, D.A., & Stillman, M.J. (1997). Comparison of the structures of the metal-thiolate binding site in zn(ii)-, cd(ii)-, and hg(ii)-metallothioneins using molecular modeling techniques. *Journal of Biomolecular Structure & Dynamics*, 14, 393–406.

Frenkel, Z.M. (2008). Does protein relatedness require sequence matching? Alignment via networks in sequence space. *Journal of Biomolecular Structure & Dynamics*, 26, 215–222.

Gerstein, M., & Levitt, M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Science*, 7, 445–456.

Hendrickson, W.A. (1979). Transformations to optimize the superposition of similar structures. *Acta Crystallographica Section A*, A35, 158–163.

Holm, L., & Park, J. (2000). Dalilite workbench for protein structure comparison. *Bioinformatics*, 16, 566–567.

Holm, L., & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233, 123–138.

Holm, L., & Sander, C. (1996). Mapping the protein universe. *Science*, 273, 595–602.

Huang, H.-J., Lee, K.-J., Yu, H.W., Chen, C.-Y., Hsu, C.-H., Chen, H.-Y., Tsai, F.-J., & Chen, C.Y.-C. (2010). Structure-based and ligand-based drug design for her 2 receptor. *Journal of Biomolecular Structure & Dynamics*, 28, 23–37.

Hubbard, T.J.P., Alley, B., Brenner, S.E., Murzin, A.G., & Chothia, C. (1999). SCOP: A structural classification of proteins database. *Nucleic Acids Research*, 27, 254–256.

IBM (2012a). IBM ILOG CPLEX optimization studio. Retrieved from http://www.ibm.com/software/integration/optimization/cplex-optimization-studio

IBM (2012b). IBM SPSS statistics. Retrieved from http://www.ibm.com/software/analytics/spss/products/statistics/

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, A32, 922–923.

Kar, L., Sherman, S.A., & Johnson, M.E. (1994). Comparison of protein structures in solution using local conformations derived from nmr data: Application to cytochrome c. *Journal of Biomolecular Structure & Dynamics*, 12, 527–558.

Kauvar, L.M., & Villar, H.O. (1998). Deciphering cryptic similarities in protein binding sites. *Current Opinion in Biotechnology*, 9, 390–394.

Kearsley, S.K. (1989). Structural comparisons using restrained inhomogeneous transformations. *Acta Crystallographica Section A*, A45, 623–635.

Kim, D.-S., Cho, Y., & Kim, D. (2005). Euclidean Voronoi diagram of 3D balls and its computation via tracing edges. *Computer-Aided Design*, 37, 1412–1424.
Kim, D.-S., Cho, C.-H., Kim, D., & Cho, Y. (2006). Recognition of docking sites on a protein using β-shape based on Voronoi diagram of atoms. *Computer-Aided Design, 38*, 431–443.

Kim, D.-S., Cho, Y., Kim, D., Kim, S., Bhak, J., & Lee, S.-H. (2005). Euclidean Voronoi diagrams of 3D spheres and applications to protein structure analysis. *Japan Journal of Industrial and Applied Mathematics, 22*, 251–265.

Kim, D.-S., Cho, Y., Sugihara, K., Ryu, J., & Kim, D. (2010). Three-dimensional beta-shapes and beta-complexes via quasi-triangulation. *Computer-Aided Design, 42*, 911–929.

Kim, D.-S., Kim, D., Cho, Y., & Sugihara, K. (2006). Quasi-triangulation and interworld data structure in three dimensions. *Computer-Aided Design, 38*, 808–819.

Kim, D.-S., Kim, C.-M., Won, C.-I., Kim, J.-K., Ryu, J., Cho, Y., … Bhak, J. (2011). BetaDock: Shape-priority docking method based on beta-complex. *Journal of Biomolecular Structure & Dynamics, 29*, 219–242.

Kim, D.-S., Seo, J., Kim, D., Ryu, J., & Cho, C.-H. (2006). Three-dimensional beta shapes. *Computer – Aided Design, 38*, 1179–1191.

Kini, R.M., & Evans, H.J. (1992). Comparison of protein models minimized by the all-atom and united-atom models in the amber force field: Correlation of rms deviation with the crystallographic r factor and size. *Journal of Biomolecular Structure & Dynamics, 10*, 265–279.

Kuhn, H.W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly, 2*, 83–97.

Lathrop, R.H. (1994). The protein threading problem with sequence amino acid interaction preferences is np-complete. *Protein Engineering, 7*, 1059–1068.

Lathrop, R.H., & Smith, T.F. (1996). Global optimum protein threading with gapped alignment and empirical pair score functions. *Journal of Molecular Biology, 255*, 641–665.

Lee, S.W., Suh, Y., Kim, S., & Kim, Y. (1999). Comparison of the structures of β amyloid peptide (25–35) and substance p in trifluoroethanol/water solution. *Journal of Biomolecular Structure & Dynamics, 17*, 381–391.

Munkres, J.R. (1984). *Elements of algebraic topology*. New York, NY: Perseus Press.

Needleman, S., & Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology, 48*, 443–454.

Nguyen, M.N., Tan, K.P., & Madhusudhan, M.S. (2011). CLICK-topology-independent comparison of biomolecular 3d structures. *Nucleic Acids Research, 39*, W24–W28.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, 12*, 85–94.

Shindyalov, I.N., & Bourne, P.E. (1998). Protein structure alignment by incremental combinatorial extention (CE) of the optimal path. *Protein Engineering, 11*, 739–747.

Taylor, W.R., & Orengo, C.A. (1989). Protein structure alignment. *Journal of Molecular Biology, 208*, 1–22.

The R Project (2012). The R Project for statistical computing. Retrieved from http://www.r-project.org/

Via, A., Ferré, F., Brannetti, B., & Helmer-Citterich, M. (2000). Protein surface similarities: A survey of methods to describe and compare protein surfaces. *Cellular and Molecular Life Sciences, 57*, 1970–1977.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*, 80–83.

Wohlers, I., Domingues, F.S., & Klau, G.W. (2010). Towards optimal alignment of protein structure distance matrices. *Bioinformatics, 26*, 2273–2280.

Zhang, S., & Wang, T. (2010). A complexity-based method to compare RNA secondary structures and its application. *Journal of Biomolecular Structure & Dynamics, 28*, 247–258.