# Aligning Ligand Binding Cavities by Optimizing Superposed Volume

Ruobing Chen and Katya Scheinberg[1]
*Dept. of Industrial and Systems Engineering*
*Lehigh University*
*Bethlehem, PA, USA*
*ruc310@lehigh.edu, katyas@lehigh.edu*

Brian Y. Chen[1]
*Dept. of Computer Science and Engineering*
*Lehigh University*
*Bethlehem, PA, USA*
*chen@cse.lehigh.edu*

## Abstract

*We describe an optimization-based method that seeks the superposition of ligand binding cavities that maximizes their overlapping volume. Our method, called DFO-VASP, iteratively uses Boolean set operations to evaluate overlapping volume in intermediate superpositions while searching for the maximal one. Our results verify that the superpositions identified are biologically relevant, and demonstrate that DFO-VASP generally discovers cavity superpositions with similar or occasionally larger overlapping volume than those of superpositions generated with existing means.*

## 1. Introduction

Algorithms that analyze protein structures often seek to identify structural similarities that point to functional or evolutionary relationships. Corresponding backbone atoms in similar positions can point to shared evolutionary histories, even in the absence of sequence similarity [1]–[3]. Corresponding atoms around similar ligand binding sites can be markers of a similar catalytic site [4]–[6]. In these and other applications, accurately finding structural markers relies on identifying corresponding atoms between two protein structures and superposing them.

Superpositions produced in this way can reveal structural markers relating to binding specificity: Cavity regions that overlap may be essential for accommodating the same molecular fragment, while regions that do not overlap may cause differences in binding specificity [7]–[9]. Unfortunately, aligning corresponding atoms does not necessarily optimize the overlapping volume between ligand binding cavities. Also, proteins that lack similarities in atomic structure may not be alignable. In such cases, a superposition of cavities based on maximizing the overlapping volume between cavities, rather than on aligning atoms, is required.

This paper presents the first algorithm for finding superpositions of protein-ligand binding cavities that maximize overlapping volume, a specific geometric superposition between two or more cavities that we refer to as the *optimal superposition*. Within the space of all possible rotations and translations that superpose one cavity onto another, we search for the optimal superposition using a version of the trust-region based derivative free optimization (DFO) framework [10]. DFO targets problems where the derivatives of the objective function are either unavailable, or unreliable. As a result, traditional nonlinear optimization techniques cannot apply. In this paper, the objective function is overlapping volume. We evaluate overlapping volume using VASP [7], which uses Boolean set operations to estimate overlapping volume, but not its derivatives with regard to superposition parameters.

Combined, these approaches enable us to detect cavity superpositions with a large overlapping volume without depending on atomic alignments. We refer to the combined approach as DFO-VASP. DFO-VASP can be initiated at any starting superposition, however, if we assume that an atomic superposition provides a good approximation of the optimal superposition we can also use is to *warm-start* our optimization.

DFO-VASP provides unique capabilities not achievable with existing superposition methods: If the volume of the optimal superposition of two cavities is smaller than the volume of some ligand, it is impossible for the ligand to bind in both cavities in the same conformation. This deduction is based on the fact that the largest superposition of the two cavities cannot be larger than the ligand itself, so the ligand must change conformations if it is to be accommodated at all. These functionalities suggest novel applications in protein engineering and in the characterization of the determinants of ligand binding specificity. We demonstrate these capabilities below using the serine proteases and the enolase superfamily.

[1] Co-corresponding Author

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

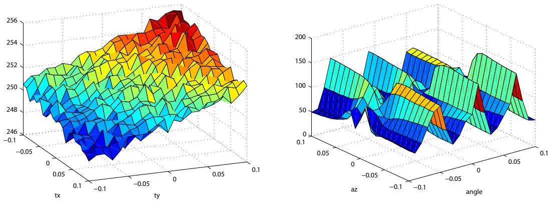

Figure 1. Noisy objective function.

## 2. Related Work

Subalgorithms that compute the least-squares superposition of corresponding points in space [11] are essential in nearly every algorithm for protein structure comparison. Their significance derives from the capability to rapidly determine, for two sets of corresponding points, the geometric alignment that minimizes Root Mean Squared Distance (RMSD). This capability enables structure comparison algorithms to efficiently test hundreds of correspondences between chemically similar atoms or amino acids, in a search for atoms with maximum geometric and chemical similarity. This approach appears frequently among algorithms for comparing whole protein structures [12]–[16].

A second type of structure comparison algorithm focuses on the comparison of binding sites in a similar manner. Here, correspondences are constructed between the amino acids of an input protein structure and “motifs” of amino acids that represent catalytic sites [6], [17]–[19]. Least-squares superpositions are also crucial in algorithms that focus on identifying similar points or patches on the molecular surface of a protein structure [20], [21].

The work presented here explores an entirely distinct strategy for superposing protein structures based on the shape of empty cavities in binding sites that accommodate partner molecules. Binding cavities of this nature can be highly nonconvex and have nonzero genus, making them different from representations based on spherical harmonics [22], [23], but effectively comparable with Boolean Set operations computed with VASP [7]–[9]. Here, we apply DFO to search for the optimal superposition.

## 3. Methods

### 3.1. Derivative-Free Optimization

Derivative-Free Optimization methods target unconstrained minimization problem $\{\min f(x) : x \in R^n\}$ where the first derivatives of the objective function $f(x)$ are unavailable and cannot be approximated by traditional methods because the function value computations are comparatively costly and prone to noise. The problem of finding *optimal superposition* by maximizing the overlapping volume is precisely the problem of this type. VASP approximately computes the volume of the intersection of two or more protein structures given their relative positions. To help visualize the non-smooth aspects of the noisy function computed by VASP, we show in Figure 1 the surface of the noisy function with respect to two of the parameters, with the others fixed.

The model-based DFO algorithm that we use in this paper is based on a trust-region framework described in [10]. At each iteration, one constructs a model that sufficiently approximates the objective function within a “trust region”. The model function is then (approximately) minimized in the “trust region” to define a trial step. This iterative framework produces a sequence of points that lead us to a local optimum. The essential mechanism of the above algorithm lies in the checking for the sufficient reduction, which is corrupted when the underlying function $f$ is computed with noise. Accordingly, we develop a dynamic accuracy increment strategy: at each iteration, given current VASP “resolution” $r$, and the estimate of noise level $\xi_r$; if the model reduction is comparable to $\xi_r$, reduce the level of noise by decreasing the “resolution” $r$, and compute the new model in the “trust region”.

This algorithmic framework has been shown to converge to a local optimal solution in the absence of noise. In practice this method tends to find ”good” local optimal solutions, however, no guarantee of global solution can be provided. Hence different starting points may produce different final result if the optimization problem has multiple optima. Since the atomic and the maximum volume superpositions may be closely related, it is natural to use the atomic superposition as initial point for the optimization. We refer to results of this setting as the *warm-started* alignments. However, considering the limitations of atomic superpositions as discussed earlier, we disregard the dependence by finding *randomly-started* alignments. In this case, ten starting points are chosen by using Latin Hypercube Sampling (LHS) techniques [24]. DFO is thus independently initiated from these starting points, the solution with the largest intersection is returned.

### 3.2. Statistical Modeling

In earlier work, we developed two ways to measure similarity in superposed cavities. First, we measure *volumetric similarity* $d(a, b)$ between two superposed cavities $a$ and $b$, as:

$$d(a, b) = \frac{v(a \cap b)}{v(a \cup b)},$$

where $\cap$ is the Boolean intersection of two cavities, $\cup$ is the Boolean union of two cavities, and $v(r)$ is the volume within a region $r$ generated by a Boolean

**Serine Protease Superfamily:**
**Trypsins:** 2f91, 1fn8, 2eek, 1h4w, 1bzx, 1aq7, 1ane, 1aks, 1trn, 1a0j **Chymotrypsins:** 1eq9, 8gch **Elastases:** 1elt, 1b0e
**Enolase Superfamily:**
**Enolases:** 1e9i, 1iyx, 1pdy, 2pa6, 3otr, 1te6 **Mandelate Racemase:** 1mdr, 2ox4 **Muconate Lactonizing Enzyme:** 2pgw, 2zad

Figure 2. **PDB codes of structures used.**

operation. The geometric interpretation of a set of aligned cavities with high volumetric similarity is that they overlap closely, and thus have very similar shape. In contrast, cavities with low volumetric similarity overlap poorly. A statistical model trained on volumetric similarity between superposed cavities with identical binding preferences can reveal cavities that are too dissimilar to have similar specificity [9].

We can also measure dissimilarity in two superposed cavities based on the volume of *fragments*, which we define as regions in one cavity that do not intersect with the other cavity. A fragment can accommodate parts of a ligand molecule that cannot be accommodated by the other cavity. A statistical model trained on the volumes of fragments that occur between cavities with identical specificity can reveal, between other cavities, fragments too large to suggest identical specificity [8].

In both models, we establish 0.05 as the probability threshold for statistical significance. We thus refer to degrees of volumetric similarity so low as to be observed less than 5% of the time, or fragments so large as to be observed less than 5% of the time as *statistically significant*.

## 3.3. Data Set Construction

**3.3.1. Protein Families.** The serine protease and enolase superfamilies were selected for demonstrating our method because each superfamily contained at least three subfamilies with distinct binding preferences and at least two sequentially nonredundant structural representatives in each subfamily. Figure 2 lists these structures by Protein Data Bank (PDB) [25] code, in groups classified by similar binding preference.

The PDB (6.21.2011) contains 676 Serine proteases from chymotrypsin, trypsin, and elastase subfamilies and 66 enolase superfamily structures from enolase, mandelate racemase, and muconate cycloisomerase subfamiles. From each set, we removed mutant, partially ordered, and structures in “closed” or otherwise inactive conformations. Structures with greater than 90% sequence identity were removed, with preference for those associated with publications, resulting in 14 serine protease and 10 enolase structures. Within these structures, ions, waters, and other non-protein atoms were removed. Hydrogens, unavailable in all structures, were removed for uniformity. Atypical amino acids (e.g. selenomethionines) were not removed. Solid geometric representations of binding cavities were generated with a method described earlier [8].

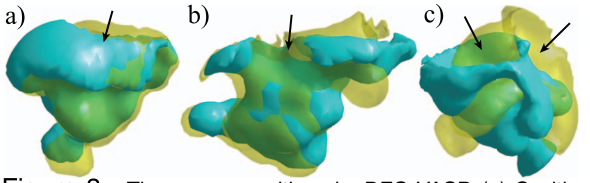

Figure 3. Three superpositions by DFO-VASP. (a) Cavities from 1e9i (teal) and 1te6 (yellow, transparent). (b) Cavities from 1ane (teal) and 1a0j (yellow, transparent). (c) Cavities from 1e9i (teal) and 2pa6 (yellow, transparent). Black arrows indicate the entrance and direction of the cavity.

As a comparison to DFO-VASP, Ska [13], an algorithm for whole-protein structure alignment, was used to superpose all pairs of serine protease structures and all pairs of enolase structures, generating a *backbone superposition* of all cavities. Superpositions were also generated with Dali [14] and CE [12], but since proteins in these datasets have identical folds, there were no significant differences.

# 4. Experimental Results

## 4.1. Validating DFO-VASP Superpositions

To test how effectively DFO-VASP identifies biochemically relevant superpositions, we generated superpositions of all pairs of serine protease and all pairs of enolase cavities. Visually examining all 91 pairs of superposed serine protease cavities, we observed that all 91 cases, superposed cavities were logically oriented: Entrances to each cavity were oriented in exactly the same direction, and conserved cavity shapes were strongly superposed. An example of a superposition like this is Figure 3b. 33 of the 45 pairs of superposed enolase cavities were also superposed in logical orientations, with cavity entrances oriented in nearly identical directions, (e.g. Figure 3a). From the remaining 12, six pairs of enolase cavities were superposed with entrances at an angle of approximately 45 degrees, at an angle where ligand access to both cavities would have been difficult, and six more cavities were superposed at an angle of approximately 90 degrees, where ligand access to both cavities is impossible. Figure 3c is an example of this kind of erroneous superposition. In total, 124 out of the 136 superpositions produced cavities superposed in biochemically consistent orientations.

All superpositions observed here, however, differed in some respects from backbone superpositions. S1 cavities in serine proteases have different lengths, causing DFO-VASP to “center” smaller cavities along

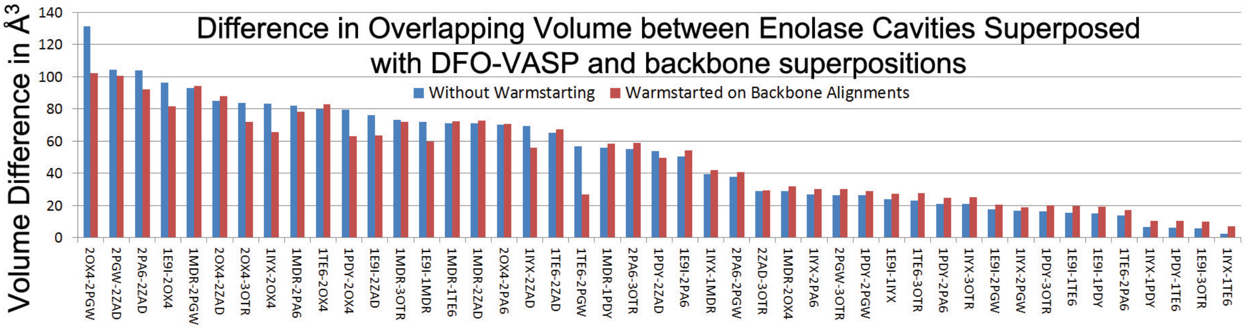

Figure 4.

longer cavities. The entrance to these cavities is defined in part by backbone shape, and as a result, backbone superpositions generally superposed the cavity entrances more closely than the whole volume. Enolase cavities generally had similar depth, and the same effect did not occur.

## 4.2. Comparison to Backbone Superpositions

To further evaluate DFO-VASP, we computed superpositions of each pair of cavities in both data sets. Optimal superpositions were computed using random starting positions, and also by warm-starting with backbone superpositions. The volumes of intersection generated by these two methods were compared to the volume generated by backbone superposition.

In all cases, DFO-VASP superpositions exhibited greater volumes of intersection than backbone superpositions. This larger degree of superposition is apparent in Figure 4, which plot the difference of intersection volumes from DFO-VASP superpositions minus the volumes of intersection from backbone superpositions, among enolase cavities. Serine protease superpositions exhibited a similar trend.

In general, warm-started superpositions exhibited final superpositions that had slightly greater volume than superpositions starting from arbitrary starting points. This effect is to be expected, because the warm starting point enables DFO-VASP to spend more time exploring the neighborhood of the optimal superpositions.

## 4.3. Large-Scale Validation

To verify the predictive accuracy of our method, we constructed statistical models of volumetric similarity [8] and fragment volume [9] based on trypsin and enolase cavities aligned with ska. Using DFO-VASP, we computed all against all superpositions of serine protease cavities and enolase cavities with the latin hypercube starting strategy. For each cavity in each dataset, this computation resulted in a multiple cavity superposition of the rest of the dataset onto the cavity. For each multiple cavity superposition, we computed volumetric similarity and fragment volume between all pairs and evaluated their statistical significance.

Among superpositions of enolase cavities, 3236 out of the 3567 fragments generated between enolase cavities were statistically insignificant. Among the 210 superpositions of an enolase and a non-enolase cavity, 100% of the largest fragments were statistically significant. Volumetric similarity between enolase cavities superposed by DFO-VASP was statistically insignificant in 188 out of 210 superpositions, while volumetric similarity between an enolase and a non-enolase cavity was statistically significant 210 out of 210 times. When aligned with DFO-VASP, volumetric differences and similarities between enolase cavities with different and similar binding preferences were almost always large enough to be detected automatically with statistical models.

Among superpositions of serine protease cavities, 97%, or 20424 out of the 20919 fragments generated between trypsin cavities were statistically insignificant. Among the 462 superpositions of a trypsin and a non-trypsin cavity, the largest fragment was statistically significant in 405 superpositions. Volumetric similarity between trypsin cavities superposed by DFO-VASP were statistically insignificant in 706 out of 770 superpositions, while volumetric similarity between trypsin and non-trypsin cavities was statistically significant in 452 out of 462 superpositions. Like the enolase cavities, when aligned with DFO-VASP, volumetric differences and similarities, between serine protease cavities with different and similar binding preferences, were almost always large enough to be detected automatically with statistical models.

## 5. Conclusions

We have presented an adaptation of DFO and VASP for generating superpositions of ligand binding cavities by maximizing overlapping volume. Our method includes techniques that compensate for noisy, variable-time volume evaluations and methods for warm-starting the search for the optimum superposition.

DFO-VASP generally provided biologically correct superpositions. Cavity entryways, for example, generally overlapped. In all but two cases, cavities aligned by DFO-VASP had similar or greater volumes of superposition than superpositions generated with existing methods. Our results suggest that DFO-VASP can be a viable approach for binding site superposition, and that it exhibits novel capabilities. We also assessed, at a large scale, the potential of DFO-VASP for generating superpositions of binding cavities that can be used to detect influences on binding specificity. On both serine protease and enolase datasets, volumetric similarity and fragment volume were almost always statistically significant between cavities with different binding preferences, and almost always statistically insignificant between cavities with similar binding preferences. DFO-VASP can be effective for detecting influences on specificity.

These results suggest that it is possible to align and compare ligand binding cavities when atomic similarities do not exist. These novel capabilities point to applications in discovering influences on ligand binding specificity.

## References

[1] P. Koehl *et al.*, "Protein structure similarities," *Current Opinion in Structural Biology*, vol. 11, no. 3, pp. 348–353, 2001.

[2] D. Petrey, M. Fischer, and B. Honig, "Structural relationships among proteins with different global topologies and their implications for function annotation strategies," *Proceedings of the National Academy of Sciences*, vol. 106, no. 41, pp. 17 377–17 382, 2009.

[3] B. Rost, "Twilight zone of protein sequence alignments," *Protein engineering*, vol. 12, no. 2, pp. 85–94, 1999.

[4] A. Stark, S. Sunyaev, and R. B. Russell, "A Model for Statistical Significance of Local Similarities in Structure," *J Mol Biol*, vol. 326, pp. 1307–1316, 2003.

[5] T. A. Binkowski, P. Freeman, and J. Liang, "pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins." *Nucleic Acids Res*, vol. 32, web server issue, pp. W555–8, 2004.

[6] B. Y. Chen, V. Y. Fofanov, D. H. Bryant, B. D. Dodson, D. M. Kristensen, A. M. Lisewski, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs." *J Comp Biol*, vol. 14, no. 6, pp. 791–816, 2007.

[7] B. Y. Chen and B. Honig, "VASP: A Volumetric Analysis of Surface Properties Yields Insights into Protein-Ligand Binding Specificity," *PLoS Comput Biol*, vol. 6, no. 8, p. 11, 2010.

[8] B. Chen and S. Bandyopadhyay, "VASP-S: A Volumetric Analysis and Statistical Model for Predicting Steric Influences on Protein-Ligand Binding Specificity," in *Proceedings of 2011 IEEE International Conference on Bioinformatics and Biomedicine*, 2011, pp. 22–9.

[9] ——, "A Statistical Model of Overlapping Volume in Ligand Binding Cavities," in *Proceedings of the Computational Structural Bioinformatics Workshop (CSBW 2011)*, 2011, pp. 424–31.

[10] A. Conn, K. Scheinberg, and L. Vicente, *Introduction to derivative-free optimization*. Society for Industrial Mathematics, 2009, vol. 8.

[11] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica A*, vol. 34, pp. 827–828, 1978.

[12] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path." *Protein Eng*, vol. 11, no. 9, pp. 739–47, Sep. 1998.

[13] A.-S. Yang and B. Honig, "An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance." *J Mol Biol*, vol. 301, no. 3, pp. 665–78, Aug. 2000.

[14] L. Holm and C. Sander, "Mapping the protein universe." *Science*, vol. 273, no. 5275, pp. 595–603, Aug. 1996.

[15] L. Xie and P. E. Bourne, "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments." *Proc Natl Acad Sci U S A*, vol. 105, no. 14, pp. 5441–6, Apr. 2008.

[16] Y. Ye and A. Godzik, "Multiple flexible structure alignment using partial order graphs." *Bioinformatics*, vol. 21, no. 10, pp. 2362–9, May 2005.

[17] Y. Tseng, J. Dundas, and J. Liang, "Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns," *Journal of molecular biology*, vol. 387, no. 2, pp. 451–464, 2009.

[18] M. Shatsky, A. Shulman-peleg, R. Nussinov, and H. J, "Recognition of Binding Patterns Common to a Set of Protein Structures," *Lect Notes Comput Sc*, vol. 3500, pp. 440–455, 2005.

[19] S. Schmitt, D. Kuhn, and G. Klebe, "A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology," *J Mol Biol*, vol. 323, no. 2, pp. 387–406, Oct. 2002.

[20] M. Rosen, S. L. Lin, H. Wolfson, and R. Nussinov, "Molecular shape comparisons in searches for active sites and functional similarity." *Protein Eng*, vol. 11, no. 4, pp. 263–77, Apr. 1998.

[21] K. Kinoshita and H. Nakamura, "Identification of the ligand binding sites on the molecular surface of proteins," *Protein Sci*, vol. 14, pp. 711–718, 2005.

[22] J. Kovacs, P. Chacon, Y. Cong, E. Metwally, and W. W, "Fast rotational matching of rigid bodies by fast fourier transform acceleration of five degrees of freedom." *Acta crystallographica. Section D, Biological crystallography*, vol. 59, no. 8, pp. 1371–1376, 2003.

[23] A. Kahraman, R. J. Morris, R. a. Laskowski, and J. M. Thornton, "Shape variation in protein binding pockets and their ligands." *J Mol Biol*, vol. 368, no. 1, pp. 283–301, Apr. 2007.

[24] M. Mckay, W. Conover, and R. Beckman, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, pp. 239–245, 1979.

[25] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank." *Nucleic Acids Res*, vol. 28, no. 1, pp. 235–42, Jan. 2000.