Improving Accuracy in Binding Site Comparison with Homology Modeling

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

Abstract

Conformational changes make the comparison of protein structures difficult. Algorithms that identify small differences in protein structures to identify influences on specificity are particularly affected by molecular flexibility. However, such algorithms typically compare proteins with identical function and varying specificity, causing them to focus on closely related proteins rather than the remote evolutionary homologs sought by most comparison algorithms. This focus inspired us to ask if structure prediction algorithms, which more accurately predict the structures of close evolutionary neighbors, can be used to "remodel" existing structures with the same template, to make the comparison of their binding sites more accurate. Our results, on the enolase superfamily and the tyrosine kinases, reveal that this approach to error reduction is indeed possible, enabling our methods to identify influences on specificity in protein structures that originally could not be compared.

1. Introduction

Conformational flexibility affects accuracy in all methods for protein structure comparison. Most algorithms in this category seek to identify proteins with remote evolutionary and functional similarities (e.g. [1]–[3]), employing an assumption of conformational rigidity in order to make database searches more practical. Comparisons of this nature can be applied to whole protein structures (e.g. [4]–[6]) or focus on just crucial elements of protein binding sites (e.g. [7]–[9]). In both cases, small changes in backbone or sidechain conformation can obscure the strong similarity that is required to distinguish truly related proteins from the noisy background of randomly similar structures [10]–[12].

Molecular flexibility has a related effect on an emerging class of comparative algorithms that predict influences on binding specificity [13]–[15]. These algorithms analyze superposed binding cavities to identify regions that may be conserved, to accommodate the same molecular fragment, and other regions that vary, to create differences in binding specificity (Fig. 1). Like other methods, the detection of specificity determinants can be negatively effected by conformational change. Unlike other methods, however, algorithms that identify influences on specificity are generally examining very similar proteins rather than distant evolutionary homologs. This focus on similarity permits a novel approach that we propose here: to mitigate comparison errors from conformational variations by using protein structure prediction algorithms.

Specifically, this paper examines the hypothesis that homology modeling can be used to *remodel* close homologs in nearly identical conformations to enhance comparison accuracy, and thus the identification of structural influences on binding specificity. Naturally, the prediction of protein structures is also a source of error; even changes in rotamer conformation can alter binding cavity shape. Nonetheless, it is crucial to observe that structure prediction by homology modeling is more accurate when the structures of close homologs can be used as modeling templates [16], [17]. These methods generally preserve the structure of functional surfaces [18], as well. The comparison of very similar protein structures may thus be able to leverage the strengths of homology modeling algorithms to mitigate errors from conformational change.

We tested this hypothesis by examining redundant and nonredundant representatives of two families of well studied flexible proteins: the enolases and the tyrosine kinases. Structures from both families exhibit conformational variations in binding site shape that are known to cause differences in binding preferences. Both families also exhibit varying conformations that make their binding sites, as provided in existing crystal structures, incomparable. Using NEST [19], a rapid homology modeling package, and VASP [13], a protein structure comparison tool for identifying potential influences on specificity, we observed that remodeling

1. Corresponding Author

978-1-4673-2747-3/12/$31.00 ©2012 IEEE

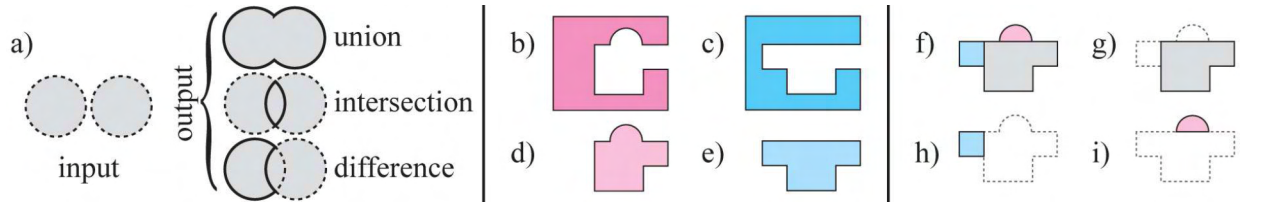

Fig. 1. Boolean Set Operations (a) on solid representations of protein structures (b,c) or aligned binding sites (d,e,f) can be used to detect similarities in binding sites (g) that accommodate the same molecular fragment or variations (h,i) that cause differences in specificity [13]–[15].

the structures of each family onto a single template enabled influences on specificity to be identified when they could not be, due to conformational flexibility. Together, these results point to new applications of protein structure prediction that will enhance the quality of protein structure comparisons in the future.

2. Related Work

Computational representations of protein structures often employ the assumption of conformational rigidity in structure comparisons. Many of these representations describe structure with points in three dimensions [2], [4]–[6], [20] and use geometric invariants (e.g. [4]), conserved matrices of inter-point distances [1], or topological similarities in geometric graphs [3], [21] to detect similar protein structures. A second class of point-based representations compare functional sites using "motifs" in three dimensions to represent atoms in catalytic sites [7]–[9], [22], evolutionarily significant amino acids [23], "pseudo-centers" representing protein-ligand interactions [24], and pseudoatoms representing amino acid sidechains [25]. If the points are assumed to remain rigidly associated with each other, point-based representations can be rapidly compared with least-squares methods [26], [27]. The rigidity assumption enables both these types of algorithms to rapidly scan databases of protein structures (e.g. [28]) to detect proteins with remote evolutionary similarities.

Relaxing the assumption of molecular rigidity to detect similar proteins in different conformations requires more sophisticated representations. Methods like Flexprot [29] and FlexSnap [30] represent protein structures as rigid components connected by virtual hinges, enabling their detection in different conformations. Posa [31] uses partial order graphs to construct correspondences between similar components shared among multiple proteins in different conformations. These representations are optimized for detecting remote evolutionary relationships despite differences in molecular conformation, by accounting for significant insertions and deletions and by allowing parts of the protein to be aligned in different positions.

In contrast to existing work, VASP [13], the comparison method applied here, compares the empty regions inside protein binding with Boolean set operations. Boolean set operations can detect unusually large differences [14] or unusually small similarities [15] in binding cavities that influence binding specificity. Much like existing methods, differences in conformation can make comparisons with solid representations inaccurate, as we observed on the START domains [13]. By normalizing the conformation of protein backbones with homology modeling, we present a novel strategy for mitigating the influence of conformation change on the accuracy of structure comparisons among similar proteins.

3. Methods

Our study integrates several existing methods to test our hypothesis. Here, we paraphrase these methods and explain how they are applied in the current setting. Overall, as input, we begin with a collection of protein structures for comparison. These structures exhibit the same fold and biochemical function, but may prefer to interact with different substrates. Furthermore, these structures may exhibit a range of backbone or side chain conformations, both active or inactive. In brief, our integrated approach begins by designating one structure as a *template* and modeling the other proteins with the template, using NEST [19]. We hypothesize that the effect of this modeling phase will be to reduce conformational differences between the model structures. We then structurally align the models, identify binding sites in each model, and compare the binding sites using VASP [13]. These steps are detailed below.

Model Building. Input for NEST is a template structure and a protein sequence alignment, which we generate with Ska [20]. With these inputs, and otherwise default parameters, NEST creates models with an "artificial evolution" approach that iteratively modifies the template structure with insertions, deletions, and mutations from the input alignment. For each modification, relaxation steps minimize van der Waals, hydrophobic, electrostatic, torsional, and hydrogen bonding potentials, and only the most favorable

modification is accepted in each iteration. Once the template has been completely modified into the query protein, the resulting model is returned as output.

Structural Alignment. Model structures are superposed onto the template, and for binding site using Ska [20], an algorithm for whole-structure superposition.

Binding Cavity Representation. Solid representations of binding cavities are generated by first aligning a structure or a model onto a designated structure with a bound ligand. In this data, the atrolactic acid bound to mandelate racemase from pseudomonas putida (pdb structure 1mdr) was used for cavity generation in the enolase structures. Among tyrosine kinases, the staurosporine bound to human Abelson kinase (pdb structure 2hz4) was used for cavity generation among tyrosine kinase structures.

When the structures or models are superposed onto the liganded structure, the position of the ligand indicates the binding site in each model. Using the aligned position of each model with the ligand, we generate a solid representation of each binding cavity: First, we generate a series of spheres with a 5 Å radius, centered at each ligand atom. Using VASP, we compute the Boolean union of these spheres. Second, we generate a molecular surface of the model using the Trollbase library from GRASP2 [2], which applies the classical rolling-probe technique [32] with a water-sized probe of radius 1.4 Å. The molecular surface is a closed surface that can be interpreted as a three dimensional solid. Using VASP, we use Boolean subtraction to subtract the molecular surface from the union of ligand spheres. Third, with the Trollbase library, we generate an "envelope surface" with a probe of radius 5.0 Å, based on an external cavity boundary used in SCREEN [33]. Using VASP, we compute the Boolean intersection of the envelope surface and what remains of the ligand spheres. The resulting region is a solid representation of the binding cavity on the model structure. This approach is detailed in [13].

Binding Site Comparison. One indicator of structural differences between two binding cavities is the existence of regions in one binding cavity that that can accommodate molecular fragments that will not fit in the other cavity. Given binding cavities A and B, we thus measure their similarity by using the volume of the largest contiguous region where the two cavities do not overlap (e.g. Figure 1h,i). We refer this region as the largest *fragment* between A and B. Fragments between similar cavities tend to have very small volumes, while fragments between cavities with different binding preferences are larger.

The largest fragment between A and B is determined by first computing the symmetric Boolean difference $(A - B) \cup (B - A)$. This set of operations may generate several disconnected fragments. Using a graph-based method described here [14], [34], we systematically isolate every contiguous region and measure its volume. The fragment with the largest volume (in Å^3) is returned.

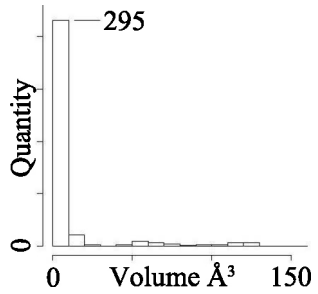

Fig. 2. Histogram of fragment volumes between enolases with identical binding preferences. 295 out of the 340 plotted here occupied less than 10 Å^3 in spatial volume.

Statistical Modeling. Fragments that occur between cavities with identical binding specificity are generally very small, due to structural similarity between the cavities. For example, among enolases with similar binding preferences (see homogenous enolase set, Figure 3), 248 out of 340 fragments were less than one cubic angstrom. 295 were less than 10 Å^3 (Figure 2). Similar distributions were observed for tyrosine kinases.

This left-skewed distribution of fragment volumes can be approximated closely by the log-normal distribution, which allows us to estimate the value of the distribution at any point on the positive x axis. As we shown earlier [14], [34], we can estimate the probability of observing a fragment equal or larger volume than a given fragment, which we refer to as the p-value. We describe a fragment whose volume has a p-value less than 0.05 as *statistically significant*.

The p-value estimates the probability of observing a fragment of a given size between two cavities assumed to have identical binding preferences. If that probability is too low, e.g. less than 0.05, then we reject our assumption as improbable, and predict the alternative: that the cavities exhibit different binding preferences. This prediction is based on our evaluation of the data, and not a statement of fact.

In this work, we evaluate the statistical significance of fragment volumes between modeled and unmodeled cavities. To evaluate the effect of modeling on fragment volume, we always train our statistical model on unmodeled cavities, as we did in earlier work [14], [15], [34], [35], so that we can assess the improvement in prediction accuracy when using remodeled structures, relative to unmodeled structures.

3.1. Data Set Construction

Protein Families. We test the effectiveness of our methods on the enolase superfamily and the tyrosine kinases. The wealth of research on both families enabled us to verify the accuracy of our predictions against established experimental evidence. These families were selected on the requirement that both families exhibit binding site flexibility that is visible in publicly available structures. We also require that both families exhibit variations in binding preferences based on well documented variations in their binding cavities. While differences in specificity may exist in the context of other ligands, our experimentation uses the experimentally determined binding preferences described below as a "gold standard" for similarities or differences in binding preferences.

Both families exhibit binding site flexibility that can have an integral impact on function. Members of the enolase superfamily have a flexible "capping domain" that influences specificity [36] and can close the active site. Tyrosine kinases exhibit several modes of conformational flexibility near the active site, including the "DFG flip" on the activation loop [37] and multiple conformations of the phosphorylation loop (the "P-loop") [38]. These conformational variations can hamper the comparison of enolase and tyrosine kinase binding sites, making it difficult to detect influences on specificity.

Proteins in the enolase superfamily catalyze reactions that abstract a proton from a carbon adjacent to a carboxylic acid. These reactions occur near the C-terminal ends of beta sheets in a conserved TIM-barrel, where amino acids act as acid/base catalysts to facilitate several different reactions [39]. Our demonstration focuses on three key differences in specificity that occur between the three enolase subfamilies: Enolases (EC 4.2.1.11), the eponymous subfamily, catalyze the dehydration of 2-phospho-D-glycerate to phosphoenolpyruvate [40], mandelate racemases (EC 5.1.2.2) convert (R)-mandelate to and from (S)-mandelate [41], and muconate-lactonizing enzymes (EC 5.5.1.1) catalyze the reciprocal cycloisomerization of cis,cis-muconate and muconolactone in muconate-lactonizing enzyme [39].

The tyrosine kinases (EC classes 2.7.10.1 and 2.7.10.1) transfer a phosphate group from adenosine triphosphate (ATP) to a tyrosine on an acceptor protein. This family of proteins plays essential roles in cell signalling (e.g. [42]), and due to their importance to cell growth and death, they are widely studied as targets for inhibitor design (e.g. [43], [44]). While there are many differences in binding preferences between kinases, our demonstration here focuses simply on the impact of the *gatekeeper* residue [45] on inhibitor specificity. Kinases with small gatekeeper residues can be targeted by many inhibitors, while kinases with large gatekeeper residues become resistant [46].

Enolase Superfamily (homogeneous):
Enolases: 1e9i, 1iyx, 1pdy, 2pa6, 2xsx, 3otr
Enolase Superfamily (homogeneous, redundant):
Enolases: 1ebh, 1els, 1nel, 2al2, 3enl, 7enl

Tyrosine Kinases (homogeneous):
Small Gatekeeper residue: 1qcf, 1fgi, 1fpu, 1fvr, 1gjo, 1irk, 1k2p, 1m14, 1m7n, 1qpc, 1r0p, 1t45, 1u4d, 1yvj, 1ywn, 2src
Tyrosine Kinases (homogeneous, redundant):
Small Gatekeeper residue: 2hz4, 2e2b, 2hyy, 2hz0, 2hzn, 2hzi, 2xyn, 2hmi, 3kf4, 3kfa, 3ms9, 3mss

Enolase Superfamily (heterogeneous) :
Enolases: 1e9i, 1iyx, 1pdy, 2pa6, 3otr, 1te6, 2xsx, **Mandelate Racemase:** 2ox4, **Muconate Lactonizing Enzyme: 2pgw**, **2zad**
Tyrosine Kinases (heterogeneous):
Small Gatekeeper residue: 2hz4
Large Gatekeeper residue: 1fvr, 1luf, 1rjb, 1sm2, 1snu, 1snx

Fig. 3. **PDB codes of structures used. Bolded structures were selected as templates.**

Selection. From the enolases and the tyrosine kinases, we constructed one dataset of proteins with similar binding preferences, which we refer to as a homogeneous dataset, and a second dataset of proteins with different classes of binding preferences, which we refer to as a heterogeneous dataset. Both datasets were first constructed with sequentially nonredundant proteins by eliminating pairwise sequence identity above 90 percent (measured with Clustalw [47]). Then, to increase conformational diversity, several additional nonidentical structures were added that had more similar sequences but very different conformations.

One or more structures were selected in each data set to be modeling templates. These selections were made based on the presence of a ligand in the template, which was used for generating a cavity in the template and all aligned models, and to ensure that models were made against holo structures. Next, all queries were modeled to the template with NEST, and then aligned to the template to preserve consistency. Finally, cavities were generated using the method described above.

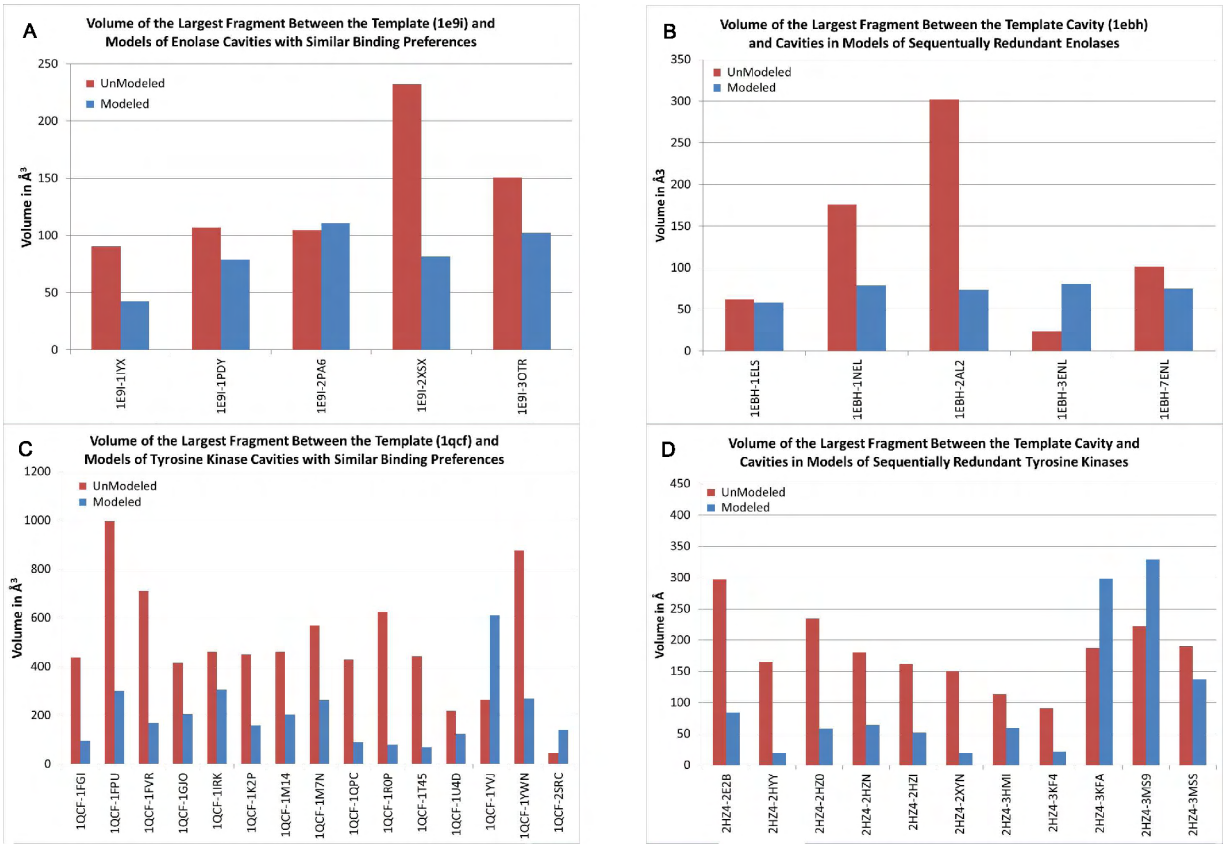

Fig. 4. Fragment volumes from cavities with similar binding preferences, before and after remodeling. Red bars indicate the volume of the largest fragment between the template cavity and cavities with similar binding preferences, before remodeling. Blue bars indicate the volume of the largest fragment between the template cavity and cavities with similar binding preferences, remodeled onto the template. Sequentially nonredundant (A) and redundant (B) enolases are shown on the top. Sequentially nonredundant (C) and redundant (D) tyrosine kinases are shown on the bottom.

4. Experimental Results

First, we demonstrate that remodeling protein structures that exhibit the same function and the same binding preferences in different conformations enables them to be more accurately compared. We then demonstrate that remodeling protein structures that exhibit the same function but different binding preferences does not make them indistinguishable from proteins with similar binding preferences.

4.1. Remodeling proteins with homogeneous binding preferences

We remodeled all members of the homogeneous enolase superfamily dataset onto the structure of saccharomyces cerevisiae enolase (1ebh). 4 out of 5 enolase cavities were more similar after remodeling than before. The homogeneous tyrosine kinase dataset was remodeled onto the structure of homo sapiens haematopoetic cell kinase (1qcf). 13 out of 15 tyrosine kinase cavities were more similar after remodeling, then before remodeling. The degree of increased similarity for both enolases and kinases can be seen in Figure 3.1a,c, which plots the volume of the largest fragment before and after modeling, for both datasets. In most cases, remodeling sequentially nonredundant proteins in different conformations yielded binding cavities that were much more similar than before.

Remodeling was similarly effective for reducing the impacts of conformational variations among very similar proteins (Figure 3.1b,d). Among enolases with greater than 90% sequence identity, the largest fragment dropped in volume in 4 out of 5 cases. Among kinases with greater than 90% sequence identity, the largest fragment dropped in volume in 9 out of 11 cases. When sequence identity is very high, cavity similarity is frequently but not necessarily enhanced by remodeling.

In several cases, the volume of the largest fragment fell a considerable amount before and after modeling.

We constructed a statistical model of fragment volume among sequentially nonredundant enolase and tyrosine kinase structures. We used that model to evaluate the significance of fragments between sequentially nonredundant enolases and kinases. In 1 out of 5 cases, the volume of the largest fragment between the template cavity and the enolase cavities was statistically significant before remodeling (2xsx), and not statistically significant afterwards. In 12 out of 15 cases, the volume of the largest fragment between the template cavity and the tyrosine kinase cavities was statistically significant before remodeling, and not statistically significant afterwards. Two of the remaining cases were always statistically insignificant (1u4d and 2src), and one case (1ywn) became more dissimilar after modeling. If we depended on statistical significance to differentiate cavity shapes without remodeling, our model would have been unable to distinguish 20% of similar-specificity enolases and 80% of similar-specificity tyrosine kinases from proteins with different binding preferences. Those errors were corrected by remodeling.

4.2. Remodeling proteins with heterogeneous binding preferences

We remodeled all members of the heterogeneous enolase superfamily against muconate cycloisomerase from Sinorhizobium meliloti (2pgw) and Thermotoga maritima (2zad). When using 2pgw as a template, the largest fragment was smaller after remodeling in 7 out of the 10 models. When using 2zad as a template, the largest fragment was smaller after remodeling in 8 out of the 10 models. These volumes are plotted in Figure 5. We also remodeled tyrosine kinases with large gatekeeper residues onto the structure of abelson tyrosine kinase from homo sapiens (2hz4). The largest fragment was smaller after remodeling in 6 out of 6 models. In general, remodeling slightly reduced structural dissimilarity between the cavities of proteins with identical function but different binding preferences. These differences are illustrated in Figure 6.

We also measured the statistical significance of each fragment generated above, before and after remodeling. Unlike the fragments generated from proteins with homogeneous binding preferences, described earlier, fragment volume did not change as much between the original structure and the remodeled structure. Fragments from original structures were often statistically significant, and fragments from remodeled structures were as well: the largest fragment for 20 out of the 20 enolases modeled on 2pgw and on 2zad were statistically significant after modeling. The largest

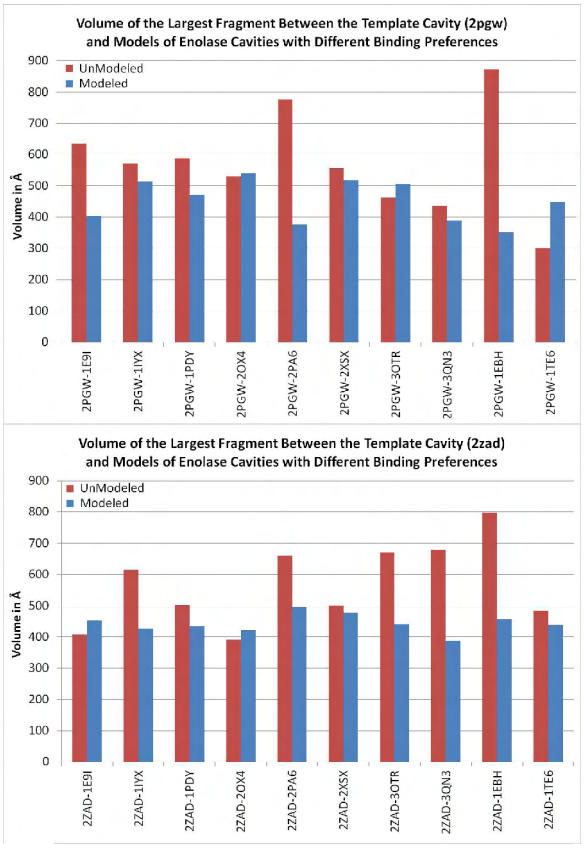

Fig. 5. Fragment volumes from enolase cavities with different binding preferences, before and after remodeling. Red bars indicate the volume of the largest fragment between the template cavity and cavities with different binding preferences, before remodeling. Blue bars; after remodeling. The top graph plots results with enolase template 2pgw, the bottom with 2zad.

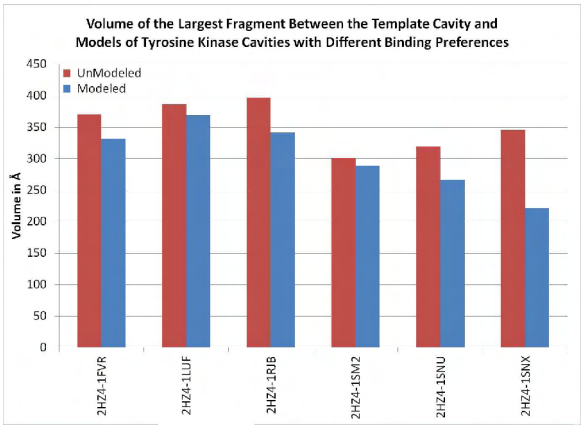

Fig. 6. Fragment volumes from tyrosine kinase cavities with different binding preferences, before and after remodeling. Red bars indicate the volume of the largest fragment between the template cavity (2hz4) and cavities with different binding preferences, before remodeling. Blue bars indicate the volume of the largest fragment between the template cavity and cavities with different binding preferences after remodeling.

fragment for 3 out of the 6 tyrosine kinases also remained statistically significant after modeling. These results suggest that the increased similarity induced by remodeling does not make it more difficult distinguish cavities with different binding preferences from those that have similar binding preferences, because cavities with different binding preferences remained structurally distinct in our data.

5. Conclusions

We have demonstrated a remodeling approach for comparing the structures of proteins with similar evolutionary backgrounds but different conformations. Our method exploits the fact that homology modeling algorithms are the most accurate when predicting protein structures that are very similar to a given template structure. This approach perfectly complements our intended application, the identification of structural influences on specificity, which is focused on the analysis of very closely related proteins that exhibit the same function but different binding preferences.

We demonstrated our initial results on sequentially nonredundant representatives of the enolase superfamily and the tyrosine kinases. Starting with protein structures in different conformations, we showed that remodeling onto the same template could normalize differences in cavity shape. In cases where binding cavities were closed or inactive, remodeled conformations returned them to an open state, allowing them to be accurately compared.

The accuracy of this comparison is evident in our results. Despite large differences in initial conformation, binding cavities in remodeled enolases and tyrosine kinases with similar binding preferences showed only statistically insignificant differences in shape. Thus, binding cavities with similar binding preferences were statistically indistinguishable from other binding cavities with similar binding preferences. In contrast, remodeled enolases and tyrosine kinases with different binding preferences exhibited binding cavities that remained largely different: Differences between these cavities were statistically significant relative to cavities with similar binding preferences. These results indicate that it is possible to accurately predict the presence of similar or different binding preferences from remodeled protein structures.

The approach we have demonstrated has considerable applications. Most importantly, it indicates that many protein structures that do not have the same conformation, due to differences in crystallographic method or due to the presence of absence of ligand, can still be compared in an accurate manner. Second, it suggests that the analysis of protein sequences that emerge from high throughput sequencing technologies and gene resequencing efforts can be analyzed from a structural perspective. Together with other sources of information, integrating volumetric comparisons and protein structure prediction algorithms may offer important advancements for protein engineering and protein function annotation.

References

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

[2] D. Petrey and B. Honig, "GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences." *Method Enzymol*, vol. 374, no. 1991, pp. 492–509, Jan. 2003.

[3] 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.

[4] R. Nussinov and H. J. Wolfson, "Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques." *Proc Natl Acad Sci U S A*, vol. 88, no. 23, pp. 10 495–9, Dec. 1991.

[5] C. A. Orengo and W. R. Taylor, "SSAP: Sequential Structure Alignment Program for Protein Structure Comparison," *Method Enzymol*, vol. 266, pp. 617–635, 1996.

[6] 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.

[7] 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." *Journal of Computational Biology*, vol. 14, no. 6, pp. 791–816, 2007.

[8] J. A. Barker and J. M. Thornton, "An algorithm for constraint-based structural template matching : application to 3D templates with statistical analysis," *Bioinformatics*, vol. 19, no. 13, pp. 1644–1649, 2003.

[9] R. B. Russell, "Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution." *J Mol Biol*, vol. 279, no. 5, pp. 1211–27, Jun. 1998.

[10] 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, "Geometric Sieving : Automated Distributed Optimization of 3D Motifs for Protein Function Prediction," in *Proceedings of The Tenth Annual International Conference on Computational Molecular Biology (RECOMB 2006)*, 2006, pp. 500–515.

[11] 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.

[12] V. Fofanov, B. Chen, D. Bryant, M. Moll, O. Lichtarge, L. Kavraki, and M. Kimmel, "A statistical model to correct systematic bias introduced by algorithmic thresholds in protein structural comparison algorithms," *2008 IEEE International Conference on Bioinformatics and Biomeidcine Workshops*, pp. 1–8, Nov. 2008.

[13] 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.

[14] 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 (BIBM), 2011, pp. 22–9.

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

[16] P. Koehl, M. Levitt *et al.*, "A brighter future for protein structure prediction," *nature structural biology*, vol. 6, pp. 108–111, 1999.

[17] D. Baker and A. Sali, "Protein structure prediction and structural genomics," *Science's STKE*, vol. 294, no. 5540, p. 93, 2001.

[18] J. Zhao, J. Dundas, S. Kachalo, Z. Ouyang, and J. Liang, "Accuracy of functional surfaces on comparatively modeled protein structures," *Journal of structural and functional genomics*, pp. 1–11, 2011.

[19] Z. Xiang and B. Honig, "Extending the accuracy limits of prediction for side-chain conformations1," *Journal of molecular biology*, vol. 311, no. 2, pp. 421–430, 2001.

[20] 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.

[21] J. F. Gibrat, T. Madej, and S. H. Bryant, "Surprising similarities in structure comparison." *Curr Opin Struct Biol*, vol. 6, no. 3, pp. 377–85, Jun. 1996.

[22] M. Moll, D. H. Bryant, and L. E. Kavraki, "The LabelHash algorithm for substructure matching." *BMC Bioinformatics*, vol. 11, no. 1, p. 555, Jan. 2010.

[23] B. Y. Chen, V. Y. Fofanov, D. M. Kristensen, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "Algorithms for structural comparison and statistical analysis of 3D protein motifs." *Pac Symp Biocomput*, vol. 345, pp. 334–45, Jan. 2005.

[24] 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.

[25] 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.

[26] 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.

[27] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, 1991.

[28] 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.

[29] M. Shatsky, R. Nussinov, and H. J. Wolfson, "FlexProt: alignment of flexible protein structures without a predefinition of hinge regions." *J Comput Biol*, vol. 11, no. 1, pp. 83–106, Jan. 2004.

[30] S. Salem, M. Zaki, and C. Bystroff, "Flexsnap: Flexible non-sequential protein structure alignment," *Algorithms for Molecular Biology*, vol. 5, p. 12, 2010.

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

[32] M. Connolly, "Solvent-accessible surfaces of proteins and nucleic acids," *Science*, vol. 221, no. 4612, pp. 709–713, Aug. 1983.

[33] M. Nayal and B. Honig, "On the Nature of Cavities on Protein Surfaces : Application to the Identification of Drug-Binding Sites," *Proteins: Struct. Funct. Genet.*, vol. 63, pp. 892–906, 2006.

[34] B. Y. Chen and S. Bandyopadhyay, "Modeling regionalized volumetric differences in protein-ligand binding cavities," *Proteome Science*, vol. 10, no. Suppl 1, p. S6, 2012.

[35] ——, "A regionalizable statistical model of intersecting regions in protein ligand binding cavities," *Journal of Bioinformatics and Computational Biology*, vol. 10, no. 3, p. 1242004, 2012.

[36] J. F. Rakus, A. A. Fedorov, E. V. Fedorov, M. E. Glasner, B. K. Hubbard, J. D. Delli, P. C. Babbitt, S. C. Almo, and J. A. Gerlt, "Evolution of enzymatic activities in the enolase superfamily: L-rhamnonate dehydratase." *Biochemistry*, vol. 47, no. 38, pp. 9944–54, Sep. 2008.

[37] Y. Shan, M. Seeliger, M. Eastwood, F. Frank, H. Xu, M. Jensen, R. Dror, J. Kuriyan, and D. Shaw, "A conserved protonation-dependent switch controls drug binding in the abl kinase," *Proceedings of the National Academy of Sciences*, vol. 106, no. 1, pp. 139–144, 2009.

[38] M. Saraste, P. Sibbald, A. Wittinghofer *et al.*, "The p-loop–a common motif in atp-and gtp-binding proteins." *Trends in biochemical sciences*, vol. 15, no. 11, p. 430, 1990.

[39] P. C. Babbitt, M. S. Hasson, J. E. Wedekind, D. R. Palmer, W. C. Barrett, G. H. Reed, I. Rayment, D. Ringe, G. L. Kenyon, and J. A. Gerlt, "The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids." *Biochemistry*, vol. 35, no. 51, pp. 16 489–501, Dec. 1996.

[40] K. Kühnel and B. F. Luisi, "Crystal structure of the Escherichia coli RNA degradosome component enolase." *J Mol Biol*, vol. 313, no. 3, pp. 583–92, Oct. 2001.

[41] S. L. Schafer, W. C. Barrett, A. T. Kallarakal, B. Mitra, J. W. Kozarich, J. A. Gerlt, J. G. Clifton, G. A. Petsko, and G. L. Kenyon, "Mechanism of the reaction catalyzed by mandelate racemase: structure and mechanistic properties of the D270N mutant." *Biochemistry*, vol. 35, no. 18, pp. 5662–9, May 1996.

[42] Z. Songyang, K. Carraway, M. Eck, S. Harrison, R. Feldman, M. Mohammadi, J. Schlessinger, S. Hubbard, D. Smith, C. Eng *et al.*, "Catalytic specificity of protein-tyrosine kinases is critical for selective signalling," *Nature*, vol. 373, no. 6514, pp. 536–539, 1995.

[43] D. Krause and R. Van Etten, "Tyrosine kinases as targets for cancer therapy," *New England Journal of Medicine*, vol. 353, no. 2, pp. 172–187, 2005.

[44] M. Deininger, E. Buchdunger, and B. Druker, "The development of imatinib as a therapeutic agent for chronic myeloid leukemia," *Blood*, vol. 105, no. 7, pp. 2640–2653, 2005.

[45] Y. Liu, K. Shah, F. Yang, L. Witucki, and K. M. Shokat, "A molecular gate which controls unnatural ATP analogue recognition by the tyrosine kinase v-Src." *Bioorganic & medicinal chemistry*, vol. 6, no. 8, pp. 1219–26, Aug. 1998.

[46] P. J. Alaimo, Z. a. Knight, and K. M. Shokat, "Targeting the gatekeeper residue in phosphoinositide 3-kinases." *Bioorganic & medicinal chemistry*, vol. 13, no. 8, pp. 2825–36, Apr. 2005.

[47] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez, "A new bioinformatics analysis tools framework at embl–ebi," *Nucleic acids research*, vol. 38, no. suppl 2, pp. W695–W699, 2010.