THE PROTEIN SOCIETY

G-LoSA: An efficient computational tool for local structure-centric biological studies and drug design

Hui Sun Lee[1]* and Wonpil Im[2]*

[1]Higuchi Biosciences Center, University of Kansas, Lawrence, Kansas 66047
[2]Department of Molecular Biosciences and Center for Computational Biology, University of Kansas, Lawrence, Kansas 66047

Received 29 November 2015; Revised 20 January 2016; Accepted 21 January 2016
DOI: 10.1002/pro.2890
Published online 27 January 2016 proteinscience.org

Abstract: Molecular recognition by protein mostly occurs in a local region on the protein surface. Thus, an efficient computational method for accurate characterization of protein local structural conservation is necessary to better understand biology and drug design. We present a novel local structure alignment tool, G-LoSA. G-LoSA aligns protein local structures in a sequence order independent way and provides a GA-score, a chemical feature-based and size-independent structure similarity score. Our benchmark validation shows the robust performance of G-LoSA to the local structures of diverse sizes and characteristics, demonstrating its universal applicability to local structure-centric comparative biology studies. In particular, G-LoSA is highly effective in detecting conserved local regions on the entire surface of a given protein. In addition, the applications of G-LoSA to identifying template ligands and predicting ligand and protein binding sites illustrate its strong potential for computer-aided drug design. We hope that G-LoSA can be a useful computational method for exploring interesting biological problems through large-scale comparison of protein local structures and facilitating drug discovery research and development. G-LoSA is freely available to academic users at http://im.compbio.ku.edu/GLoSA/.

Keywords: molecular recognition; local structure comparison; structural bioinformatics; computer-aided drug design

Introduction

One of the most remarkable protein features is their ability of reversible binding to other molecules. Protein responses to ligands are typically associated with a plethora of biological functions that are essential for life. A ligand can be any kind of molecules such as metal ions, substrates, partner proteins and/or nucleic acids, and drugs. Metal ion binding stabilizes protein structure, often gives rise to large conformational changes upon binding, and/or participates in catalysis.[1] Substrates bind at active sites of enzymes and are then chemically transformed into other molecules. Protein binding (i.e., protein-protein interactions) plays various roles in almost all biological activities, including, but not restricted to, signal transduction, molecule transport, gene regulation, catalytic enzymatic activities, muscle contraction, and structural roles.[2] Protein-nucleic acid interactions are also crucial in biological processes, ranging from replication and transcription to enzymatic events to miRNA machinery.[3–5] Drug compounds bind to proteins, regulating their functions so as to acquire beneficial effects to treat diseases.

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: NIH; Grant number: U54GM087519; Grant sponsor: KU; Grant number: GRF2301048; Grant sponsor: XSEDE; Grant number: MCB070009.

*Correspondence to: Hui Sun Lee, Higuchi Biosciences Center, University of Kansas, Lawrence, KS 66047. E-mail: huisun.cadd@gmail.com (or) Wonpil Im, Department of Molecular Biosciences and Center for Computational Biology, University of Kansas, Lawrence, KS 66047. E-mail: wonpil@ku.edu

Typically, molecular interactions with diverse ligands occur on local surface regions of proteins, though the shape and size of the local regions vary in terms of ligand types: for example, a center in a shell of several hydrophilic protein residues for metal ions,[6] a concave-shaped structure (i.e., "pocket") for small molecules, and noncontiguous, relatively large, flat surfaces for proteins.[7] Therefore, local structure-centric characterization of proteins rather than global structures is needed to better understand biology. Protein classification could be an illustrative example to show the needs. A major goal of SCOP (structural classification of proteins)[8] and CATH (class, architecture, topology, homologous superfamily)[9] is to understand the structural, functional, and evolutionary relationships among proteins by classifying domains according to their structure. This structure-based approach may be effective in detecting distant relationships across proteins. However, the classification by protein fold may be too conservative to do a fine classification of proteins with a similar function, yet distinct folds. Indeed, an all-against-all comparison of SCOP representatives showed the cases of evolutionary convergence to common functional sites from different folds,[10] indicating that overall protein structure dissimilarities do not necessarily imply dissimilarities in their functions.

The alignment and similarity measurement between protein structures are a key component for contemporary structural biology studies such as hierarchical classification of protein domains, protein function prediction,[11] protein structure prediction,[12] and drug discovery.[13] There are many publicly available computational tools for protein structure alignment and comparison such as DALI,[14] CE,[15] and TM-align.[16] However, these tools have been designed for global structure alignment, necessitating new computational methods that can align local structures and measure their structural similarities as a complementary or contrasting approach.

Considerable efforts have been made to develop efficient computational tools for local structure alignment. Consequently, a handful of methods are currently available. For example, SiteEngine is a geometric hashing-based pocket-comparison method. In this method, each structure is represented by physicochemically important points [i.e., pseudocenters or chemical feature points (CFPs)], and the structural similarity is measured based on geometric matching between surface patches.[17] The alignment and scoring algorithms were extended to I2I-SiteEngine, a method for structure alignment between two protein-protein interfaces.[18] ProBiS represents protein surfaces as CFPs and aligns them using a conserved geometry detected by maximum clique algorithm.[19] The structure alignment score is calculated using the root-mean-square deviation (RMSD) between aligned CFPs, the number of the aligned pairs, and alignment expectation value. The score is then standardized by *Z*-Score using precalculated alignment scores for all possible non-redundant Protein Data Bank (PDB) structure pairs.[20] CF-based alignment and scoring have a merit in that the CFs assigned in the structures are real interaction points for molecular recognition, but this approach requires more computational cost than Cα atom-based one due to a larger number of the structure-representing points in the alignment process. iAlign[21] and APoc[22] are computational methods for the structure alignment of protein-protein interfaces and protein pockets, respectively. They adopt a heuristic algorithm to align structures, where initial guessed alignments are generated from gapless sequence alignment, secondary structure alignment, and fragment superposition, and then the alignments are refined through iterative dynamic programming. Both methods provide size-independent structural similarity scores, whereas SiteEngine, I2I-SiteEngine, and ProBiS do not. The alignment and scoring in iAlign and APoc are based on the Cα atom positions (not CFPs), though interfacial contact patterns (in iAlign), the orientations of Cβ atoms, and residue-based chemical similarity (in APoc) were additionally used for more accurate scoring.

In this study, we introduce a new local structure alignment method, G-LoSA (Graph-based Local Structure Alignment), which is a generalization of our earlier algorithm that was used for template ligand identification[23] and ligand binding site (BS) prediction.[24] We have specially pursued the development of a computational tool for protein local structure alignment and similarity measurement that can be universally applied to any kinds of local structures (e.g., protein pockets and protein-protein interfaces) and provide a CF-based size-independent scoring function with reasonable computing efficiency. We first describe the alignment and scoring algorithms in detail. Representative examples and benchmark tests are then presented to illustrate robustness and applicability of G-LoSA to local structure-centric biological studies and drug design.

Results

G-LoSA scoring function

In G-LoSA, all possible alignments between two local structures are generated by iterative maximum clique search and fragment superposition (see Materials and Methods for the detailed algorithm), and the optimal alignment is determined by the maximum GA-score (G-LoSA Alignment score). The overall algorithm is schematically illustrated in Figure 1. A GA-score is a scoring function to quantify structure similarity between two local structures based on their CFPs.

$$\text{GA-score} = \text{Max}\left[\frac{1}{N_T}\left(\sum_{i}^{N_{ali}} \frac{q_i}{1+(d_i/d_0)^2}\right)\right] \qquad (1)$$

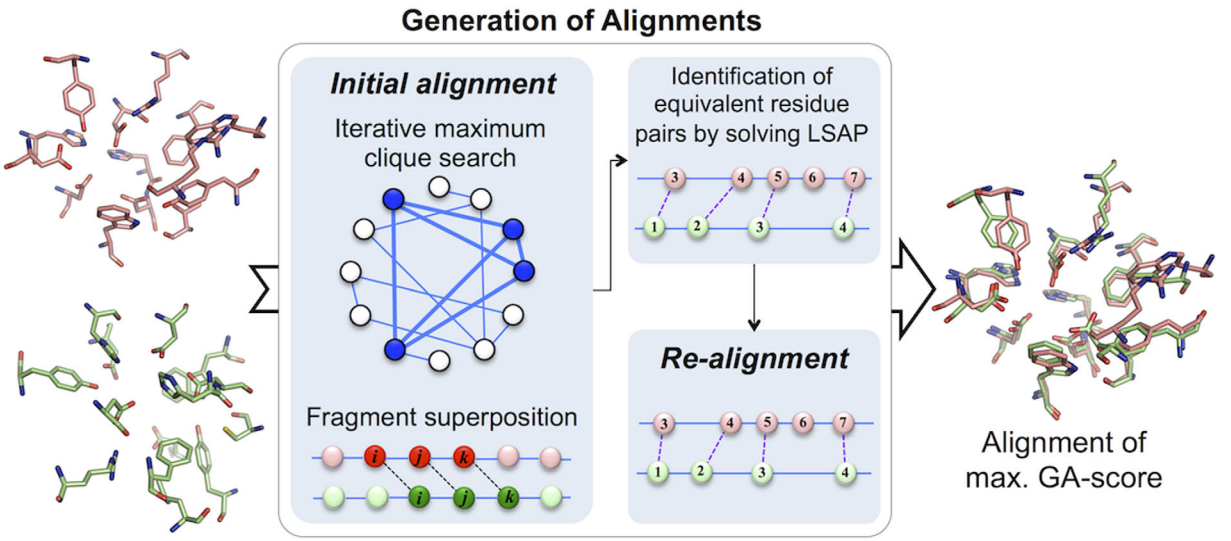

Figure 1. Schematic illustration of the alignment algorithm in G-LoSA.

$$q_i = \begin{cases} 1 & \text{if chemical feature types are identifical, or HD/OH or HA/OH} \\ 0.8 & \text{if HD/PC, HA/NC, OH/PC, OH/NC, or AR/AL} \\ 0.5 & \text{otherwise} \end{cases} \quad (2)$$

where "Max" denotes that the GA-score is the maximum of all possible alignments, N_T is the smaller number of CFPs between two local structures, and N_{ali} is the number of aligned CFPs. d_i is the distance between the CFPs in the ith pair. d_0 is the scaling factor to normalize the aligned distances. q_i is defined based on the chemical feature similarity of the ith CFP pair. We define seven different CFs for amino acids (Supporting Information Fig. S1). Hydrogen bond donor (HD), hydrogen bond acceptor (HA), hydroxyl group (OH), positively charged atom (PC), and negatively charged atom (NC) are defined by single atom points, while aromatic ring (AR) and aliphatic hydrophobic group (AL) by the geometric center of a set of atoms (see Supporting Information Table S1 for the detailed definition for each amino acid).

Figure 2 shows the average GA-scores calculated from all pairs of 2,454,439 random local structures as a function of number of CFPs. The raw GA-score (rGA-score) was calculated using a constant value (2.5 Å) for d_0. For the GA-score, a size-dependent scaling factor $d_0(N_T)$ was used instead, so that the average GA-score is not dependent on the size of the random structure pairs. The scaling factor was empirically obtained from curve fitting to the plots of the average d between an aligned CFP pair as a function of N_T.

$$d_0(N_T) = 0.27\sqrt{N_T - 6} + 0.98 \quad (3)$$

As shown in Figure 2, the mean GA-scores, normalized by $d_0(N_T)$ are independent of the number of CFPs (i.e., the size of the random local structures), but the rGA-score decreases from 0.69 to 0.35 as the number of CFPs increases. The average GA-score for a random local structure pair is 0.49. Table I shows the statistical significance of the GA-score derived from the random local structures. The GA-score distribution for all random local structures was modeled by the type I extreme value distribution (Gumbel distribution; Supporting Information Fig.

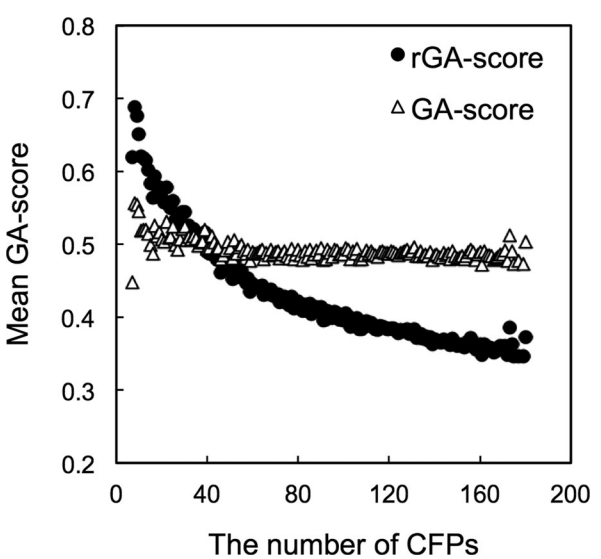

Figure 2. The average raw GA-score (rGA-score) and the GA-score of the random local structure pairs as a function of number of CFPs.

Table I. *Statistical Significance of the GA-Score Derived from the Random Local Structures*

GA-score	0.57	0.59	0.65	0.71	0.76	0.81
P-value	1×10^{-1}	5×10^{-2}	1×10^{-2}	1×10^{-3}	1×10^{-4}	1×10^{-5}

S2), and the *P*-values of representative GA-scores are given in Table I. A GA-score of 0.59 is significant at $P < 5 \times 10^{-2}$.

Local structure alignment by G-LoSA

Figure 3 shows four representative examples to illustrate the quality of G-LoSA local structure alignment. The alignments were obtained from PDB ligand/BS-structure library search by G-LoSA. For this search, the BS of a ligand TQ3 (5-phenylsulfanyl-2,4-quinazolinediamine) in PDB IDs: 1IA1 (chain id: A) was used as the query (target) structure. To put strict conditions on the structure library search, we excluded all homologous library proteins whose sequence identity is $> 30\%$ to the target protein. The four representative alignment pairs were chosen to have the same number of BS residues as the target BS (the number of BS residues is 13) and various GA-scores. Structural comparisons between the target BS and the identified template BS show a clear correlation between GA-score and their structural similarity.

To further provide insight into a relationship between GA-scores and BS structural similarities, Figure 3 also shows similarity scores between the target and template ligands, which are measured by the overlap ratio (R_O) defined by N_{OI}/N, where N_{OI} and N are the number of overlapped identical atoms and a total number of atoms in the ligand, respectively.[23] A library ligand is transferred into the target BS upon the superposition of its BS. If the distance between an atom of the target ligand and its nearest atoms of the template ligand is ≤ 1.2 Å and their atom types are identical, the ligand atom is defined as the overlapped identical atom. The R_O values of the four examples show a strong correlation with the GA-scores. Based on the fact that structurally similar pockets recognize similar ligands, the results also support the robustness of G-LoSA for quantitating similarity between different protein local structures (in this example, different small ligands binding pockets).

Benchmark validation for diverse types of local structures

We benchmark the G-LoSA performance in detecting biologically related local structures from experimental structures. To examine universal applicability of G-LoSA to various types of local structures, we evaluated its performance against four different

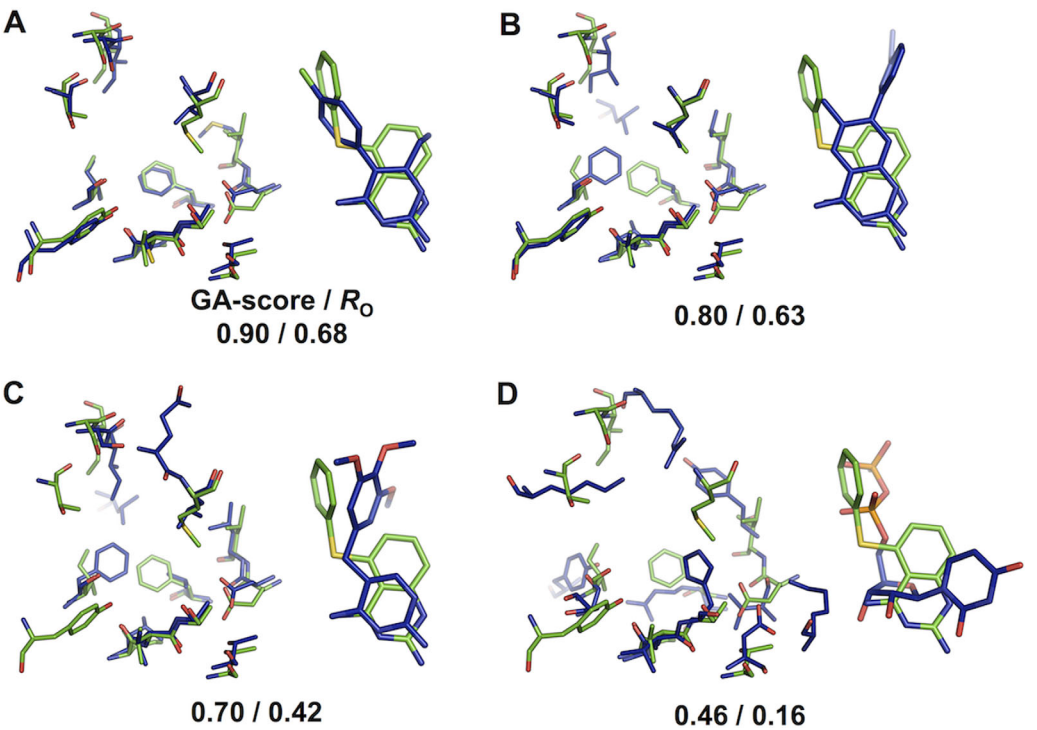

Figure 3. Representative examples to illustrate the relationship between GA-score and local structure similarity. Each library ligand/BS structure [blue (A) PDBs:2BL9, (B) 3SRQ, (C) 2W9G, and (D) 3OTK] is aligned to the target BS/ligand structure (green, PDB:1IA1) with its GA-score between BS and overlap ratio (R_O) between ligands.

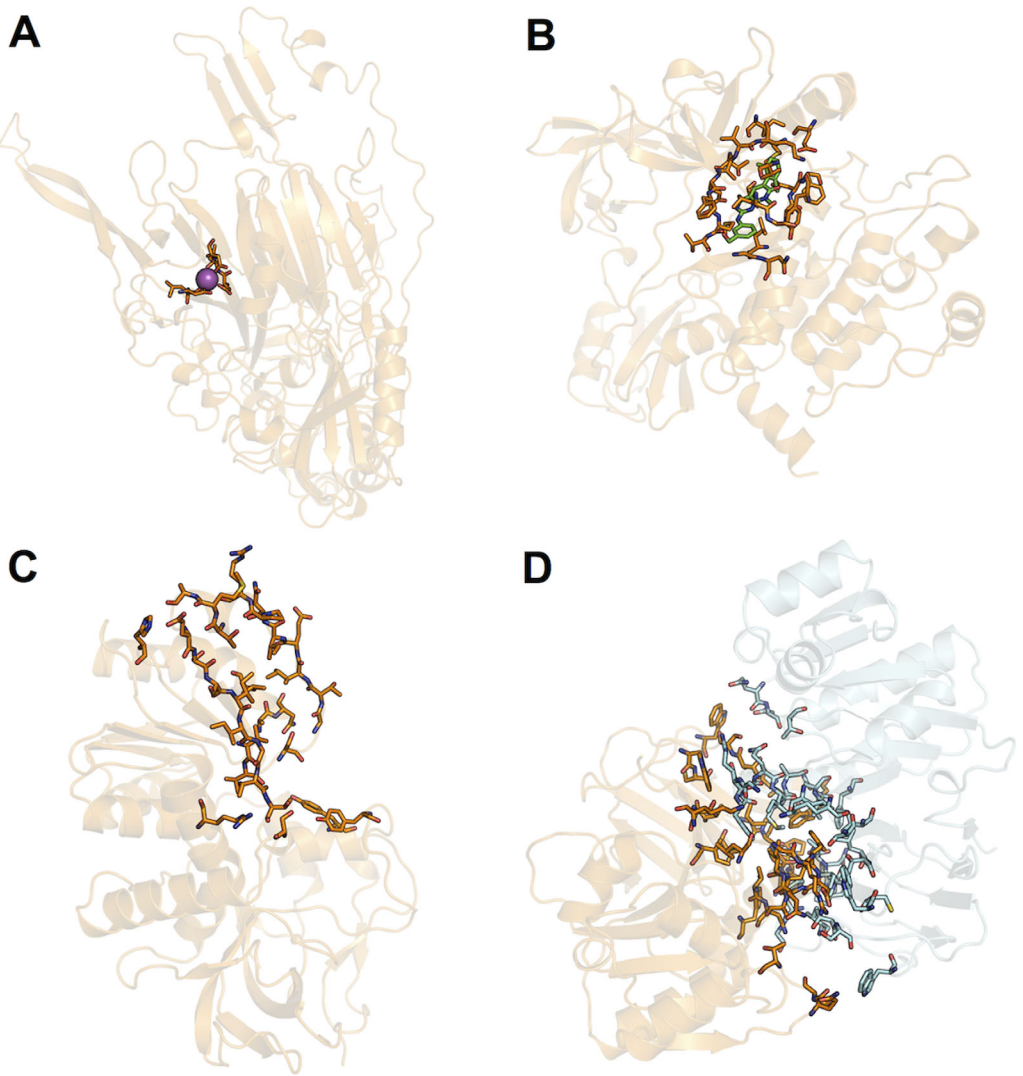

Figure 4. Structural illustration of four different benchmark local structure sets: (A) Ca^{2+}-BS, (B) small-molecule ligand BS set, (C) protein binding surface set, and (D) protein-protein interface set.

benchmark sets: Ca^{2+}-BS set, small-molecule ligand BS set, protein binding surface set, and protein-protein interface set (Fig. 4). The evaluation against the protein binding surface set uses the same homologous/nonhomologous pair set in the protein-protein interface set. Only difference is that for a given protein-protein interface pair (e.g., interface between chains A and B in PDB_1 vs. interface between chains C and D in PDB_2), the final GA-score for the protein binding surface benchmark evaluation is determined by the best GA-score between the surface of the first element (PDB_1_A) and either surface in the second element (PDB_2_C and PDB_2_D).

The quantitative comparisons are based on the receiver-operating-characteristic (ROC) curves from the prediction results and the area-under-curve (AUC) values. An AUC value of 1.0 signifies that the tool perfectly prioritizes homologous local structure pairs in terms of the similarity score, whereas a value of 0.5 implies random prediction. The G-LoSA performances are compared with those of APoc (for the Ca^{2+}-BS, small-molecule ligand BS, and protein binding surface sets) and iAlign (for the protein-protein interface set). APoc and iAlign were chosen for comparison because of their availability and reported outstanding performance for ligand BS and protein-protein interaction interfaces, respectively. PS-score (for APoc) and IS-score (for iAlign) normalized by smaller structure were used as the quantities for structure similarity measurement by these control tools.

Compared with APoc or iAlign, G-LoSA shows consistently better or comparable performance against the diverse benchmarks (Fig. 5), indicating its reliable performance regardless of the sizes or characteristics of target local structures. As shown in the ROC curve for the Ca^{2+}-BS set (Supporting Information Fig. S3), G-LoSA (AUC = 0.98) shows considerably better performance than APoc (AUC = 0.46). The worse performance by APoc may be due to its

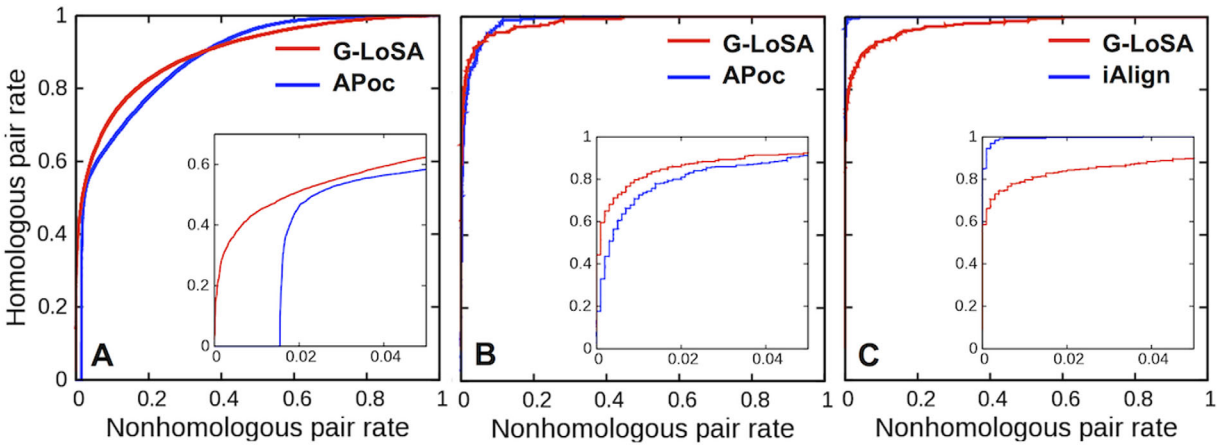

Figure 5. ROC plots for (A) small-molecule ligand BS, (B) protein binding surface, and (C) protein-protein interface benchmark sets. The insets are the ROC plots within a low true negative rate $\leq 5\%$.

alignment algorithm and scoring function that are optimized to deal with small ligand binding pockets with more than nine residues.[22] When APoc is applied to the small-molecule ligand BS set, it shows reliable performance [Fig. 5(A): AUC 0.90 for G-LoSA and 0.88 for APoc]. It is not surprising that APoc also shows comparable performance to G-LoSA even against the protein binding surface set [Fig. 5(B): AUC 0.98 for G-LoSA and 0.99 for APoc] in that protein binding surface is also a pocket associated with protein binding, though it is much more flat than small-molecular ligand BS in general. For the protein-protein interface set, where APoc is not applicable, G-LoSA shows comparable performance to iAlign [Fig. 5(C): AUC 0.98 for G-LoSA and 1.0 for iAlign]. Note that iAlign has its applicability limited only to protein-protein interfaces due to its original design for interfacial structural alignment and scoring.

Performance evaluation from the ROC plots could be more relevant within a very small nonhomologous pair rate range (e.g., $\leq$5%) because the high ability to recognize the homologous structure pairs at the beginning of a rank-ordered list is needed for practical tasks involved in a large structure library search. When we focus on the regime with a low nonhomologous pair rate $\leq 5\%$, G-LoSA shows reliable performances against the small-molecule ligand BS and protein binding surface sets [insets of Fig. 5(A,B)]. For the protein-protein interface set, iAlign outperforms G-LoSA, but the performance of G-LoSA is also reliable [inset of Fig. 5(C)], where G-LoSA achieves a homologous pair rate of 0.93 at nonhomologous pair rate of 0.05. The better performance of iAlign results from its algorithms optimized for protein-protein interface structures (e.g., calculations of contact overlap factors to consider the similarity of interfacial contacts between different interfaces). Overall, the analysis demonstrates that G-LoSA can be used to accurately quantify structure similarity for a given local structure pair regardless of the sizes and characteristics of the local structures of interest, suggesting its universal applicability to the characterization and classification of local structures for diverse biological studies.

Application to entire protein structures

We evaluate how accurately G-LoSA detects local structural conservation in entire proteins using the small-molecule ligand BS set. The original benchmark set consists of pairs of different small molecule BS structures. Instead of using the two BS structures, we performed structure alignment between the entire protein structure of one BS and the other BS structure (e.g., given a BS structure pair BS_PDB_1 and BS_PDB_2, G-LoSA aligns BS_PDB_2 onto the entire structure of PDB_1). This evaluation is more challenging in that alignment of a small structure onto much larger one has a high propensity of generating a false positive result than alignment between two small structures whose sizes are comparable. The ROC analysis (Fig. 6) demonstrates that G-LoSA outperforms APoc with more salient overall performance difference (AUC 0.86 for G-LoSA and 0.78 for APoc) than the above comparison of the two BS structure [Fig. 5(A)].

Figure 7 shows a representative example to illustrate a potential application of G-LoSA to whole protein search for template-based ligand BS, ligand structure, and protein BS prediction. For this illustration, we choose Ras-related protein 1 (Rap1; chain A in PDB:4DXA[25]) as a target protein, as this protein has both the ligand and the protein BS. 100 and 200 templates were first identified for ligand and protein BS prediction, respectively, in terms of GA-score. We excluded all homologous library proteins whose sequence identity is $>30\%$ to the target protein during the library searches. TM-scores between the target and template proteins were also

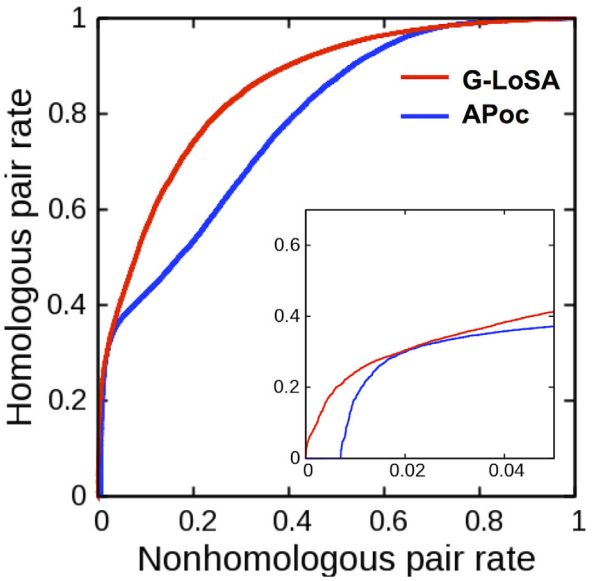

Figure 6. ROC plots for small-molecule ligand BS. In this analysis, alignment by G-LoSA and APoc was performed between the entire protein structure of one BS and the other BS structure, instead of using the two BS structures. The inset is the ROC plots within a low true negative rate $\leq 5\%$.

measured by TM-align. The templates were then resorted by an average of GA-score and TM-score to take into account the complementarity of global structure similarity in identifying good templates.[24] Five templates with the highest average scores were selected for each prediction (i.e., top five templates for ligand BS prediction and another top five templates for protein BS prediction). We measured the ratio of the correctly predicted BS residues over the total number of BS residues in the target protein (i.e., recall value) for the top five templates of each prediction. The predicted residues are defined as those that have any atom within 1 Å from any atom in the aligned template. The aligned structures of the top templates onto the target protein, the similarity scores, and recall values are shown in Supporting Information Figure S4 (for ligand BS prediction) and Supporting Information Figure S5 (for protein BS prediction). Our approach successfully predicts both sites [Fig. 7(A)] where the partner protein (Krev interaction trapped 1, KRIT1) forms a complex [Fig. 7(B), recall = 0.58 in the best template] and the native ligand (5′-guanosine-

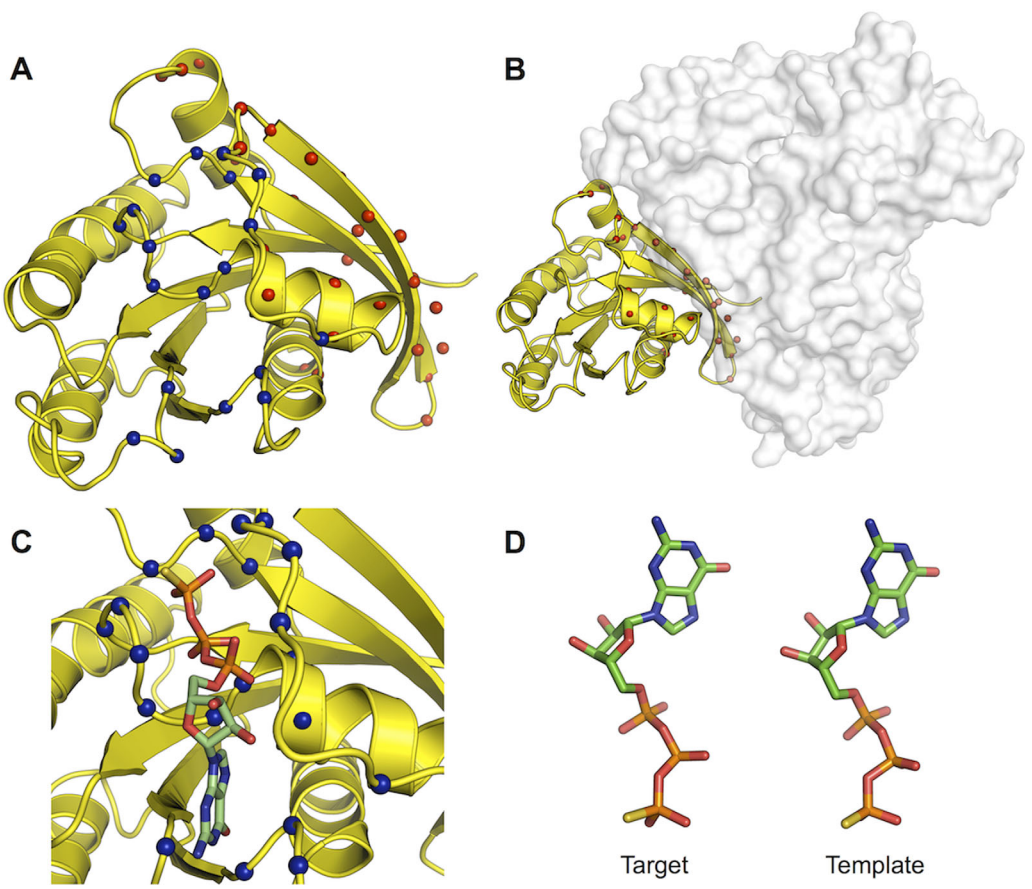

Figure 7. An example of template-based ligand and protein BS prediction by G-LoSA. A: The target protein structure (chain A in PDB:4DXA) shown in cartoon representation with predicted BS residues in sphere representation (blue for ligand BS and red for protein BS). B: The experimental structure of the target protein in complex with its partner protein. C: The native ligand structure in the target protein. D: A comparison of structure and conformation between the native and template ligands. The figures were prepared using the templates of the highest recall value for each prediction (template 1 for ligand BS prediction and template 4 for protein BS prediction).

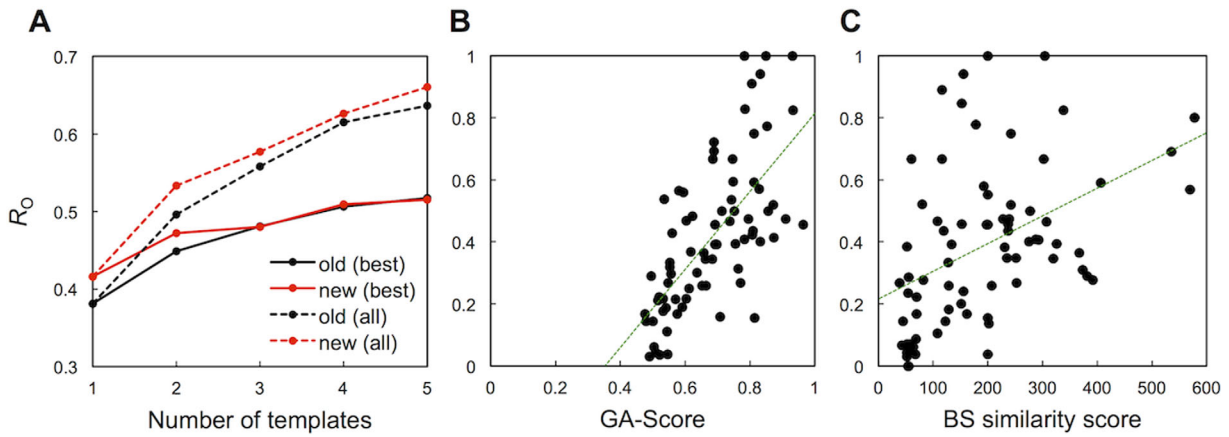

Figure 8. Performance comparison between the current and previous G-LoSA in identifying template ligands for the 75 benchmark targets from the Astex diverse set. A: The average R_O vs. the number of top templates. The best template was chosen among given top N templates in terms of R_O. The average R_O for multiple templates was measured by using all the top N templates. B: R_O vs. GA-score. C: R_O vs. BS similarity score used by the previous G-LoSA. For (B) and (C), the templates of the highest GA-score and BS similarity score for each benchmark target were used, respectively.

diphosphate-monothiophosphate) binds [Fig. 7(C), recall = 0.85]. In addition to the accurate prediction of the ligand BS, the template ligand perfectly regenerates the structure and conformation of the native ligand [Fig. 7(D)]. These results demonstrate the potential of G-LoSA to detect biologically important local regions on the entire surface of a given protein.

Comparison with previous G-LoSA in identification of ligand templates

Major extensions of G-LoSA algorithm in this study over our earlier version include (i) generation of more initial structure alignments using both multiple maximum clique solutions and fragment superposition, (ii) CF-based size-independent score of structure similarity, and (iii) usage of linear sum assignment problem (LSAP) algorithm to refine the initial alignments. In our previous study,[23] we applied the earlier version to searching for template ligands in a known protein-ligand BS structure library, aiming to design a ligand for a target protein. The performance evaluation against a benchmark set shows that using the currently available protein-ligand structure library, this approach can identify a single template ligand that is highly similar to the target ligand in more than half of the benchmark targets. In addition, multiple templates with partial similarity to the target ligand can also be identified for fragment assembly-based ligand design, leading to enhancement of the prediction performance. These results clearly show the potential application of G-LoSA to template-based ligand modeling for *de novo* ligand design. Here, we evaluate the ability of the new G-LoSA in identifying template ligands and compare the performance with the earlier version, using the same benchmark set (75 targets from the Astex diverse set[26]). The BS structure of each benchmark target was defined by a cutoff of 4.5 Å from the cognate ligand and used as a query BS structure to search for template ligands from the ligand BS structure library. In this case, the GA-score was normalized by the number of CFPs in the query BS structure to identify template ligands that can maximally cover the target ligand.

Figure 8(A) shows the average R_O between the target ligands and the best template ligands as a function of number of top templates (solid lines). The best template is a template with the highest R_O among the top N templates. The average R_O calculated using the coordinates from all top N templates is also plotted in the figure (dotted lines). The current G-LoSA outperforms the previous one for both the best template and multiple templates cases, indicating the ability of the current G-LoSA in more accurately rank-ordering for better templates.

Another advantage over the previous version is that GA-score can play a role as a confidence score for estimating the quality of predicted models from the templates. In Figure 8(B), the R_O values of the ligand templates from the best GA-scored BS for each target are plotted in term of GA-score. The results show that the GA-score has a reliable correlation with R_O (Pearson product moment correlation coefficient $R = 0.68$), suggesting that the quality of a ligand template could be inferred from the GA-score between the target and template BS, as also illustrated in Figure 3. The BS similarity score (S) used in the earlier G-LoSA is defined by $S = N^2/\text{RMSD}$, where N is the number of aligned library BS-structure residues. The RMSD is the root-mean-square deviation of the aligned residues pairs and calculated using the coordinates of Cα atoms and side-chain centroids. Unlike GA-score, this scoring

function does not provide the values of a standardized range, and it is dependent on the size of the local structures. Therefore, the BS similarity score in the previous G-LoSA is hard to be used as a confidence score to predict the quality of the ligand template [$R = 0.46$, Fig. 8(C)]. The benchmark validation clearly demonstrates that the current G-LoSA shows strong potential for computer-aided drug design.

Discussion and Conclusions

Accurate characterization and prediction of molecular interactions between proteins and diverse ligands in the context of their three-dimensional structures are central to better understanding the structure and function relationship of proteins and to developing new therapeutic agents. A comparative study of protein structures is a powerful approach to detecting structural conservation in the proteins, leading to disclosing novel biological insights. Efficient computational tools for local structure alignment and similarity measurement are imperatively needed as the molecular recognition by protein mostly occurs in a local region on the protein surface rather than the global structure.

This work presents G-LoSA, a method to align protein local structures in a sequence order independent way and to provide the GA-score, a size-independent quantity of structural similarity for a given local structure pair. In particular, the GA-score is calculated based on the CFs of each amino acid and can be applied to measure the structural similarity with the local structures of diverse sizes and characteristics, yet maintaining its length independency. Our validation results indicate that G-LoSA is a robust tool for local structure-centric comparative biology studies. In particular, G-LoSA is highly effective in detecting conserved local regions on the entire surface of a given protein. G-LoSA was also used to demonstrate its applicability to identifying template ligands from the PDB library for *de novo* drug design, showing strong potential for computer-aided drug design.

In G-LoSA, both the iterative maximum clique search and the fragment superposition are adopted to generate possible alignments between two different local structures. We examined G-LoSA on which alignment approach generates the maximum GA-score against the homologous pairs of the Ca^{2+}-BS, small-molecule ligand BS, and protein-protein interface sets. The results indeed show strong complementarity between both alignment approaches: 88 and 12% (iterative maximum clique search and fragment superposition, respectively) for the Ca^{2+}-BS set, 39 and 61% for the small-molecule ligand BS set, and 51 and 49% for the protein-protein interface set. The analysis also indicates a dominant role of the iterative maximum clique search in aligning the local structures that are smaller than small-molecule ligand BS.

G-LoSA has been designed aiming at its applications to high-resolution protein structures. However, our alignment algorithm is based on the Cα positions, and thus this design enables the alignment to be less sensitive to the conformational changes of side chains. Cheng *et al*. reported a computational tool, PCalign for measuring the structural similarity of protein-protein interfaces.[27] Their method adopts a coarse-grained approach in interface representation, alignment, and scoring, thus has a merit that its performance could be tolerant to low-resolution structures such as those solved by cryo-EM. Optimization of G-LoSA to efficiently deal with structural data in different resolutions could contribute to further enhancing our understanding of protein structural biology.

With the development of high-throughput experimental techniques, the size of data repositories of biological molecules has been dramatically increasing in the PDB. In addition, computer modeling efforts to generate high-resolution structures of diverse intermediate states are also enriching the universe of available biological structural data. We expect that G-LoSA could be harnessed to search for a huge protein structure database encompassing experimentally solved and predicted protein structures to explore interesting local structure-centric biological problems and facilitate drug discovery research and development.

Materials and Methods

Random local structure set

A set of random local structures was generated to derive a CF-based size-independent scoring function and to perform a statistical significance analysis. The overall procedure of the structure set preparation is shown in Supporting Information Figure S6. A list of non-redundant 21,744 protein chains was obtained using the first element of each cluster in the clustering analysis of all protein chains in the PDB[28] by blastclust with 30% sequence identity (http://www.rcsb.org/pdb/statistics/clusterStatistics.do, as of March 2014). The coordinates of the non-redundant protein chains were extracted from the corresponding PDB files. For each protein chain, a local structure was defined as residues within a radius centered at the Cα atom of a randomly selected residue. The radius was also randomly determined within a range of 3–20 Å. A set of the local structures was filtered out based on the resolution of X-ray structures ($\leq$3 Å) and the number of residues in each local structure (3–50). Fifty structures were randomly selected from the remaining set (14,776 local structures) for each number of residues (3–50, thus 2,400 local structures in total). All

possible pairs between the local structures (2,878,800 pairs) were generated and then filtered by a global structure similarity TM-score cutoff of 0.4 (measured by TM-align[16]) between the protein chains of each local structure pair, resulting in 2,454,439 pairs.

G-LoSA alignment algorithm

In the maximum clique search method, two given local structures (A and B) are represented by Cα atoms ($\mathbf{R}^{(A)}=\{\mathbf{r}_1^{(A)},\mathbf{r}_2^{(A)},\ldots,\mathbf{r}_M^{(A)}\}$ and $\mathbf{R}^{(B)}=\{\mathbf{r}_1^{(B)},\mathbf{r}_2^{(B)},\ldots,\mathbf{r}_N^{(B)}\}$, where $\mathbf{r}$ is the coordinate of Cα atom). All combinations of inter-structural pairs ($\mathbf{P}^{AB}=\{p_{11}(\mathbf{r}_1^{(A)},\mathbf{r}_1^{(B)}),\ p_{12}(\mathbf{r}_1^{(A)},\mathbf{r}_2^{(B)}),\ \ldots,p_{MN}(\mathbf{r}_M^{(A)},\mathbf{r}_N^{(B)})\}$) are generated using the representative points from the local structures. Two pairs $p_{ij}(\mathbf{r}_i^{(A)},\mathbf{r}_j^{(B)})$ and $p_{kl}(\mathbf{r}_k^{(A)},\mathbf{r}_l^{(B)})$ are selected from $\mathbf{P}^{AB}$ and then both distances $d(\mathbf{r}_i^{(A)},\mathbf{r}_k^{(A)})$ and $d(\mathbf{r}_j^{(B)},\mathbf{r}_l^{(B)})$ are calculated. If $|d(\mathbf{r}_i^{(A)},\mathbf{r}_k^{(A)})-d(\mathbf{r}_j^{(B)},\mathbf{r}_l^{(B)})|$ is less than a cutoff (d_{cut}), p_{ij} and p_{kl} are assigned to vertices of a product graph and connected by an edge. This procedure is applied to all pairs in $\mathbf{P}^{AB}$. The generated product graph is searched for the maximum clique, the largest subset of vertices in which all vertices are connected to all other vertices. We used an improved branch and bound algorithm for fast search for a maximum clique in a product graph.[29] In our case, solving the maximum clique problem for the product graph is equivalent to identification of the largest subset of structurally aligned points. Two local structures are superposed using the rotation matrix obtained from the aligned point sets. G-LoSA repeats this procedure three times, increasing d_{cut} from 1.5 to 2.5 Å by an increment of 0.5 Å (so-called the iterative maximum clique search), resulting in three alignments.

The second method to align two local structures is to use fragments from each local structure. A set of consecutive three residues (regardless of their sequence continuity) is extracted from each local structure by allowing (for smaller local structure) or not allowing (for larger local structure) overlap between different fragments. For each fragment pair, an alignment is obtained using equivalent residue pairs in the fragment pair.

To perform a structure alignment focused on conserved residues and to reduce computational cost, we incorporated an option to filter out dissimilar residue pairs from two local structures based on their BLOSUM62 score[30] and secondary structure identity during both the iterative maximum clique search and the fragment superposition. A secondary structure for a given residue is determined based on the Cα coordinates of five neighboring residues.[16] One can freely define a specific BLOSUM62 cutoff value (default = 0) for the similarity comparison and the minimum number of similar residues pairs in a fragment (default = 1) before structure alignment. All structure alignments by G-LoSA in this study were performed using the default values. The secondary structure comparison was only applied to structure alignment of protein binding surfaces and protein-protein interfaces to reduce computing costs.

The aligned residue pairs in the initial alignment by the iterative maximum clique search and the fragment superposition are identified using the shortest augmenting path algorithm to solve the linear sum assignment problem (LSAP).[31] If the distance between the Cα atoms of an aligned residue pair is >8 Å (an empirically optimized value), the pair is discarded from the aligned residue pair set. G-LoSA then again superposes the local structures by the rotation matrix for the updated aligned residue pairs. G-LoSA also identifies the aligned CFP pairs for GA-score calculation by solving the LSAP, based on the superposed local structure pair.

Diverse local structure benchmark sets

Ca^{2+}-BS set. To prepare a Ca^{2+}-BS set, we first downloaded all PDB files (as of April 2015) from the PDB and then collected all PDB protein chains containing Ca^{2+} ions. The Ca^{2+}-BS were then extracted from the coordinates of the Ca^{2+} containing protein chains using a cutoff distance of 4.5 Å between a Ca^{2+} ion and any heavy atom in a residue. Among Ca^{2+}-BS whose number of residues is between 3 and 6 (5,240 BS), ones from redundant proteins were removed using a sequence identity cutoff of 60%. A structure pair list was generated only using the remaining Ca^{2+}-BS (1,024 BS). Among the pairs, only pairs whose protein chains are homologous to each other were selected based on the 30% sequence identity cluster lists by blastclust, followed by further filtering using TM-score (≥0.7) and a distance between Ca^{2+} ions (≤3 Å) to select only Ca^{2+}-BS from structurally similar proteins. The remaining 550 pairs were finally used as a "homologous pair set" for Ca^{2+}-BS. For the "nonhomologous pair set," 500 protein chains were randomly selected from the PDB files and then local structures whose number of residues is between 3 and 6 were randomly extracted from each protein chain. One thousand pairs were randomly selected from the pair list and used as the nonhomologous pair set.

Small-molecular ligand BS set. The datasets used for this benchmark were derived from the APoc homologous/nonhomologous pair sets (subject/control sets in the original paper).[22] The homologous pair set consists of 38,066 pairs of pockets where proteins are at low sequence identity and contain the same or similar type of ligands. The nonhomologous pair

set contains the same number pairs of pockets that interact with dissimilar ligands in randomly selected proteins with low sequence or global structural similarity. The residues of a ligand BS were extracted from the protein using a cutoff distance of 4.5 Å between any ligand heavy atom and any protein heavy atom.

Protein-protein interface set. For a given protein complex, a protein-protein interface structure consists of two protein surfaces at the interface. The datasets used for this benchmark were derived from the iAlign homologous/nonhomologous pair sets.[21] In the dimer-597 set, the homologous and nonhomologous pair sets contain 373 biologically related, structurally similar protein-protein interfaces and 176,875 unrelated pairs, respectively. The protein-protein interfacial residues were extracted from the PDB file using a cutoff distance of 4.5 Å between any heavy atoms from individual proteins.

PDB structure libraries

To build a ligand BS structure library, the X-ray structures with resolution of >3 Å were eliminated from the library. DNA and RNA molecules were also removed, and ligand molecules in the PDB files were identified in the heteroatom section. Heteroatoms having an identical chain ID and sequence number were grouped into one heteroatom group. If a distance of any atom pair from different heteroatom groups was 1–2 Å, the two heteroatom groups were merged into one group and identified as multipart ligands. Metal ions, water molecules, and small molecular weight additives were removed by setting the minimum number of heavy atoms in a heteroatom group to 5. To only consider noncovalently bound ligands, if any atom in a heteroatom group was located within 2 Å from any protein atom, the heteroatom group was identified as a covalently linked ligand and removed from the library. If any atom of a residue in a protein is within 4.5 Å of its cognate ligand, the residue is defined as the BS residue. The library contains 100,856 ligand/BS-structure pairs.

For a protein BS structure library, if any atom of a residue in a protein is within 4.5 Å of any atom in its neighboring protein with a different chain ID in the asymmetric unit, the residue is defined as the protein BS residue. Redundancy of protein BS in each PDB file was removed using G-LoSA with a GA-score cutoff of 0.85. The library contains 457,669 protein BS.

References

1. Tainer JA, Roberts VA, Getzoff ED (1991) Metal-binding sites in proteins. Curr Opin Biotechnol 2:582–591.
2. Nooren IM, Thornton JM (2003) Diversity of protein-protein interactions. Embo J 22:3486–3492.
3. Nadassy K, Wodak SJ, Janin J (1999) Structural features of protein-nucleic acid recognition sites. Biochemistry 38:1999–2017.
4. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS (2010) Origins of specificity in protein-DNA recognition. Annu Rev Biochem 79:233–269.
5. Ha M, Kim VN (2014) Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol 15:509–524.
6. Yamashita MM, Wesson L, Eisenman G, Eisenberg D (1990) Where metal ions bind in proteins. Proc Natl Acad Sci USA 87:5648–5652.
7. Blundell TL, Sibanda BL, Montalvao RW, Brewerton S, Chelliah V, Worth CL, Harmer NJ, Davies O, Burke D (2006) Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. Philos Trans R Soc London B Biol Sci 361:413–423.
8. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32:D226–D229.
9. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35: D291–D297.
10. Russell RB (1998) Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 279:1211–1227.
11. Godzik A, Jambon M, Friedberg I (2007) Computational protein function prediction: are we making progress? Cell Mol Life Sci 64:2505–2511.
12. Moult J, Fidelis K, Zemla A, Hubbard T (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins 53:334–339.
13. Lee HS, Zhang Y (2012) BSP-SLIM: a blind low-resolution ligand-protein docking approach using predicted protein structures. Proteins 80:93–110.
14. Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. J Mol Biol 233:123–138.
15. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11:739–747.
16. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302–2309.
17. Shulman-Peleg A, Nussinov R, Wolfson HJ (2004) Recognition of functional sites in protein structures. J Mol Biol 339:607–633.
18. Shulman-Peleg A, Mintz S, Nussinov R, Wolfson HJ (2004) Protein-protein interfaces: recognition of similar spatial and chemical organizations. Lect Notes Comput Sci 3240:194–205.
19. Konc J, Janezic D (2010) ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. Bioinformatics 26:1160–1168.
20. Konc J, Cesnik T, Konc JT, Penca M, Janezic D (2012) ProBiS-database: precalculated binding site similarities and local pairwise alignments of PDB structures. J Chem Info Model 52:604–612.
21. Gao M, Skolnick J (2010) iAlign: a method for the structural comparison of protein-protein interfaces. Bioinformatics 26:2259–2265.

22. Gao M, Skolnick J (2013) APoc: large-scale identification of similar protein pockets. Bioinformatics 29:597–604.
23. Lee HS, Im W (2012) Identification of ligand templates using local structure alignment for structure-based drug design. J Chem Info Model 52:2784–2795.
24. Lee HS, Im W (2013) Ligand binding site detection by local structure alignment and its performance complementarity. J Chem Info Model 53:2462–2470.
25. Li X, Zhang R, Draheim KM, Liu W, Calderwood DA, Boggon TJ (2012) Structural basis for small G protein effector interaction of Ras-related protein 1 (Rap1) and adaptor protein Krev interaction trapped 1 (KRIT1). J Biol Chem 287:22317–22327.
26. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WT, Mortenson PN, Murray CW (2007) Diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem 50: 726–741.
27. Cheng S, Zhang Y, Brooks CL, III (2015) PCalign: a method to quantify physicochemical similarity of protein-protein interfaces. BMC Bioinformatics 16:33
28. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE (2011) The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 39:D392–D401.
29. Konc J, Janežič D (2007) An improved branch and bound algorithm for the maximum clique problem. MATCH Commun Math Comput Chem 58:569–590.
30. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919.
31. Derigs U, The shortest augumenting path method for solving assignment problems - motivation and computational experience. In: Monma CL, Ed. (1985) Algorithms and software for optimization. Basel: Baltzer.