On Monomeric and Multimeric Structures-Based Protein-Ligand Interactions

Yajun Dai, Yang Li, Liping Wang, Zhenling Peng, and Jianyi Yang

Abstract—Many ligands simultaneously interact with multiple protein chains in quaternary structure (QS). However, a significant number of previous studies on template-based modeling of protein-ligand interactions were based on monomeric structure (MS), which may suffer from incomplete binding information. The defects of using MS rather than QS have not been systematically studied before. In this work, based on molecular docking experiments and binding free energy estimations, we performed a large-scale comparison of the protein-ligand interactions in both forms of structures. We found that 1) about 18.6 percent biologically relevant ligands bind multiple chains in QS simultaneously. 2) For more than 95 percent complexes with multiple chains involved in the interactions, the binding free energy is lower for the QS form than the MS form. 3) For over 70 percent complexes with multi-chain binding pockets, docking with QS yields more accurate ligand conformations than with MS. While for about 1.82 percent complexes, accurate docking conformations were obtained by MS. Based on this work, it is encouraged to make use of QS rather than MS in future studies on protein-ligand interactions.

Index Terms—Biological unit, monomer, oligomer, molecular docking, protein-ligand interactions

1 Introduction

In the hierarchy of protein structure classification, there are two important levels: i.e., tertiary structure and quaternary structure (QS), which usually correspond to monomeric structure (MS) and multimeric structure, respectively. QS is believed to be the functional form of a protein but is difficult to obtain by experiments. In the Protein Data Bank (PDB) [1], most QSs were generated from the cystography data based on manual investigation and/or automated computational algorithms, such as PISA [2].

However, rather than using QS, many previous computational studies on protein-ligand interactions [3], [4], [5], [6], [7] were based on MS. This is because the existence of errors in PDB's QSs; and the difficulty in aligning multimeric structures. However, the protein-ligand interactions may be incomplete in MS, which inevitably affects the development and performance of corresponding algorithms. It remains largely unexplored about the influence of incomplete binding information to the studies of protein-ligand interactions.

In a recent analysis of the protein-ligand binding data in BioLiP [8], structures with different biological unit (BU) and asymmetric unit (AU) were skipped [9]. It was reported that about 20 percent homomers have multi-chain binding sites [9]. Abrusán and Marsh found that multi-chain binding sites and single-chain binding sites show different characteristics, which may affect the function evolution [9], the allosteric pathways [10], and the folding of protein complexes [11].

In this work, we collected a large set of highly reliable QSs from PDB and carried out a large-scale comparison of protein-ligand interactions under the forms of both structures. We aim to answer the following questions. 1) How common are multi-chain protein-ligand interactions? 2) What is the advantage of using QS over MS? 3) When is it acceptable to use MS?

2 Methods and Materials

2.1 Raw Set

It remains an unsolved problem of inferring QS from cystography data. Many efforts have been made to solve this problem, such as PISA [2], PiQSi [12], protCID [13], EPPIC3 [14] and QSbio [15]. A comprehensive description of these resources can be found in [16]. A random forest model was used to predict the biologically relevance of protein-protein interface (PPI) and a database of druggable cavities in PPIs was constructed in [17].

In this work, in order to reduce the impact of the incorrectness occurred in the generation of QS, we constructed our dataset based on the most recent work of QSbio [15], which predicts the reliability of each QS in PDB based on a combination of sequence homology and structure alignment. QSbio provides five confidence levels for the reliability of each QS in

- *Y. Dai, L. Wang, and J. Yang are with the School of Mathematical Sciences, Nankai University, Tianjin 300071, China. E-mail: dyj402@126.com, 1510079@mail.nankai.edu.cn, yangjy@nankai.edu.cn.*
- *Y. Li is with the School of Mathematical Sciences, Nankai University, Tianjin 300071, China, and also with the College of Life Sciences, Nankai University, Tianjin 300071, China. E-mail: younglee@mail.nankai.edu.cn.*
- *Z. Peng is with the Center for Applied Mathematics, Tianjin University, Tianjin 300072, China. E-mail: zhenling@tju.edu.cn.*

Manuscript received 1 Nov. 2019; revised 14 May 2020; accepted 2 June 2020. Date of publication 16 June 2020; date of current version 3 Feb. 2022. (Corresponding author: Jianyi Yang.) Digital Object Identifier no. 10.1109/TCBB.2020.3002776

PDB: *very high*, *high*, *medium*, *low*, and *very low*. Here we only selected the QSs with the highest confidence level, i.e., *very high*. For the 110,097 QS entries in QSbio, 51,050 are marked as *very high* confidence. When multiple very high-confidence QSs are available for a PDB structure, only the first one was kept. After removing obsolete QSbio entries (as PDB is updated weekly), we obtained 32,417 QSs which are used as the starting structures in this work. Note that the QSs used in this study are all homomers, because QSbio contains annotations for homomers only.

For each ligand in a QS, the ligand-binding residues were obtained based on the distance between the ligand and the residues' atoms in the structure. A residue is defined as a binding residue if one of the atomic distance between this residue and the ligand is smaller than a specified distance cutoff, i.e., the sum of the Van der Waal's radius of the two atoms plus 0.5 Å, which is usually in the range of [3.5, 4.5] Å. The value 0.5 Å is a tolerance distance proposed in the assessment of ligand-binding site prediction methods in the CASP9 experiment [18], which was also adopted in the literature [4], [6], [7], [8]. Based on this calculation we obtained ~0.2 million ligand-binding pockets, in which ~1.5 million residues are involved in ligand binding. Note that the above definition of binding residues is purely based on distance and does not take into consideration of other interactions, such as electrostatic interaction and hydrogen bond. Further consideration of other types of interactions may result in different sets of binding residues, which is worthy of investigating in future.

2.2 Selected Set

A subset of complex structures was selected from the raw set for further analysis as follows. Among the binding pockets in the raw set, those consisting of multiple chains were kept to compare the difference between MS and QS. The following steps were applied. 1) The following ligands were excluded: ligands with too few atoms (<10), peptides, nucleic acids and biologically irrelevant ligands [8]. From this filtering, the number of binding pockets was reduced to 48,503. 2) QSs with binding residues from only one chain or only a single binding residue in one chain were excluded, resulting to 9,022 binding pockets. This means that about 18.6 percent (= 9,022/48,503) of the biologically relevant protein-ligand interactions have multiple chains involved. 3) QSs containing too many protein chains (>24) were excluded because most of them represent virus particles. We got 8,932 binding pockets after this step. 4) Redundancy was further removed at 90 percent sequence identity (defined at chain level) for the remaining data. Finally, a set of 5,356 ligand-binding pockets was obtained, which consists of 1,916 unique proteins and 797 unique ligands. To mimic the previous studies in protein-ligand interactions that used single-chain based structure [3], [4], [5], [6], [7], the MS-based binding pocket was obtained by selecting the QS chain with the most number of binding residues.

2.3 Molecular Docking

There are many molecular docking software [19]. Here one of the most popular molecular docking software AutoDock Vina [20] was used to carry out the docking experiments. Besides AutoDock Vina, other docking software such as DOCK6 [21] and MDock [22] can be also used. For each structure, self-docking was performed in the forms of both MS and QS. When running AutoDock Vina, the search box was defined as 15×15×15 $Å^3$. The center of the search box was defined as the arithmetic mean of the coordinates of all the atoms in the binding residues. For other parameters in the program, the default values were used. For each complex, three ligand conformations were obtained. The first is the native conformation (N conformation) obtained from biological unit. The other two were generated based on docking with the input of MS (M conformation) and QS (Q conformation), respectively. Note that only the top conformation (ranked by the docking score in AutoDock Vina) was generated. Table 1 summarizes the notations used in this work.

TABLE 1
The Meanings of the Major Notations in This Work

Notation	Meaning
QS	Multimeric quaternary structure, which corresponds to the biological unit in PDB.
MS	Monomeric structure extracted from the QS. It contains single chain that has the most number of ligand-binding residues.
N conformation	The native conformation of the ligand in a complex structure.
Q conformation	The ligand conformation from self-docking with the native ligand structure against the QS.
M conformation	The ligand conformation from self-docking with the native ligand structure against the MS.
N-QS	A protein-ligand complex structure consisting of N conformation and QS.
N-MS	A protein-ligand complex structure consisting of N conformation and MS.
Q-QS	A protein-ligand complex structure consisting of Q conformation and QS.
M-MS	A protein-ligand complex structure consisting of M conformation and MS.
N/Q-chain	A single chain-based protein-ligand complex structure consisting of the ligand in N/Q conformation and the selected chain from QS.

2.4 Metrics

Two major metrics, binding free energy and ligand RMSD, are used to compare the above conformations. For binding free energy estimation, we used the scoring function X-Score [23], which was shown to be competitive and stable [24]. Many other scoring functions, such as ITScore [25] and DSX [26] can be used as well. The ligand RMSD was computed based on the software fconv [27]. Because errors happened during running one of the above programs for 526 structures, only 4,830 binding pockets (from 1,766 PDB entries) were used in the following analysis. The docking results and related structure files are available at http://yanglab.nankai.edu.cn/download/PL.

Note that BU rather than AU was used in our analysis. In general, there are three possible cases: BU is identical to AU; BU is a subset of AU; AU is a subset of BU. The first two cases suggest direct experimental support for QS and thus are more reliable. For the 1,766 PDB entries, 842 (mapped to

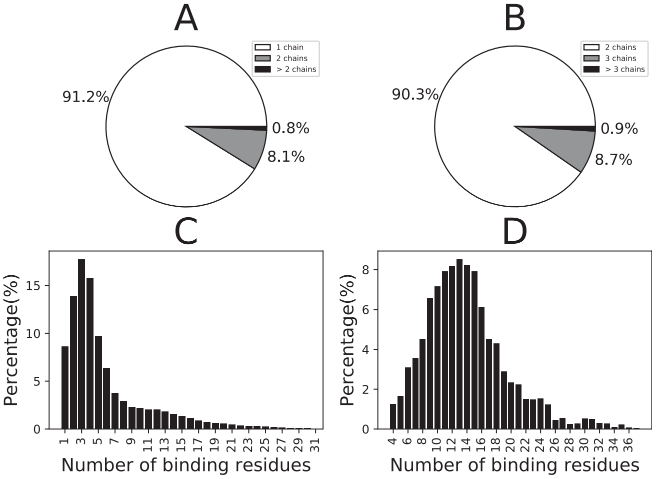


Fig. 1. Distributions of the protein-ligand binding data at chain and residue levels. (A)/(B) is for the chain distribution in the raw/selected set. (C)/(D) is for the residue distribution in the raw/selected set.

1,840 pockets) are from the first two cases and 924 (mapped to 2,990 pockets) are from the last cases. The average ligand RMSD for the QS from first two cases is 3.7 Å, which is slightly lower than that (i.e., 3.9 Å) for the QS from the last case, suggesting the importance of using more reliable QS in molecular docking studies.

3 Results and Discussions

3.1 Binding Pockets Consisting of Multiple Chains are Common

In the raw set, 91.2 percent binding pockets are comprised of residues from only one chain and the remaining 8.8 percent from multiple chains (Fig. 1A). This is mainly because the biologically irrelevant ligands, such as lipid, glycerol, acidic group and so on, are used as additives to facilitate crystallization experiments. These ligands tend to locate on the surface of the protein structure and have few contacts with the protein. For example, Figure S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2020.3002776, shows the QS of a membrane protein, which contains 30 lipids located on the surface of the structure and bound one chain only. These ligands may not perform biological function in the structure and are thus removed from the raw set to generate the selected set.

After removing peptides, nucleic acids and biologically irrelevant ligands, the ratio of multi-chain binding pockets increases to 18.6 percent. This means that 81.4 percent of the filtered ligands bind one chain only. For these structures, it is reasonable to use MS in the study of protein-ligand interactions [3], [4], [5], [6], [7]. However, there are still a notable portion of multi-chain binding pockets. These structures are not appropriate for MS-based algorithms, which may suffer from incomplete binding information and deserve a systematic investigation. In addition, we find that for the 285 structures in the core set of the PDBbind (v.2016) [28], about 35 percent of them contains more than one chain and about 14 percent binding pockets are formed by residues from multiple chains. This also suggests that binding pockets consisting of multiple chains are common.

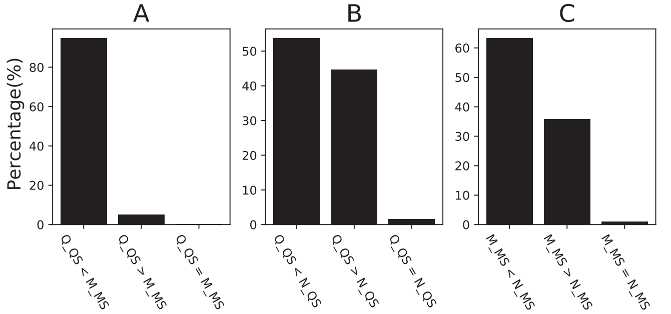


Fig. 2. Binding free energy comparisons between different ligand-protein complexes. (A) Comparison between the docked poses with MS and QS. (B) Comparison between the docked pose and the native pose in QS. (C) Comparison between the docked pose and the native pose in MS.

For the multi-chain binding pockets in the selected set, most of them (90.3 percent) consist of two chains (Fig. 1B). For example, Figure S2, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2020.3002776, presents the QS of a homodimer, in which the ligand locates at the interface between two chains (PDB ID: 4H82). In this structure, the ligand (compound 2b–2b) was designed by linking two ligands together chemically to control the crystal packing [29]. As the ligand binds to 15 residues of each chain, the MS form can be obtained by selecting any one of the two chains.

At the residue level, binding pockets with 3 (17.70 percent) or 4 residues (15.78 percent) are the most common in the raw set. About half of the binding pockets contain no more than 5 residues (Fig. 1C). This distribution seems to be unusual as a ligand (with the exception of ions and ion-like ligands) should be in contact with more residues to make the interaction stable. This is probably because many ligands (i.e., glycerol) are frequently used for structure determination and thus are not biologically relevant. Most of such ligands do not have strong interaction with the protein and some are even far away from the surface of the protein structure (14). No filtering of ligands was done in the raw set, which may lead to the unusual distribution in Fig. 1C. Thanks to the filtering process, more residues are involved in the ligand binding in the selected set than the raw set (13.9 versus 7.4 on average). The distribution for the number of binding residues in Fig. 1D resembles a normal distribution and the largest category is for pockets with 13 residues (8.46 percent).

3.2 Docking Using the Quaternary Structure is Favorable in General

For each complex in the selected set, the X-Score program [23] was used to compute the ligand-protein binding free energy. X-Score is an empirical scoring function which assumes that the binding free energy is determined by an additive functional form of four energy terms, i.e., Van de Waals interaction; hydrogen bond; deformation effect; and hydrophobic effect. A lower value of free energy usually indicates stronger binding affinity.

First, we compared the binding energy of the Q-QS and M-MS complexes. Fig. 2A shows for most of complexes (95 percent), the Q-QS complex has lower binding free energy than the M-MS complex, which can be explained by

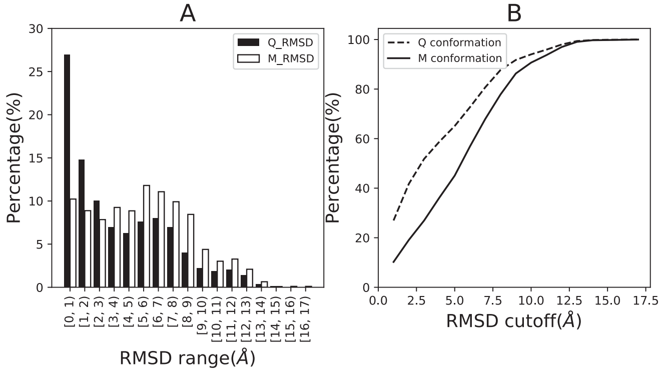


Fig. 3. RMSD comparison between the docking conformations with MS and QS. (A) is the RMSD distribution for the all the 4,830 complexes. (B) is the curve for the percentages of complexes with RMSD lower than the specified RMSD cutoffs.

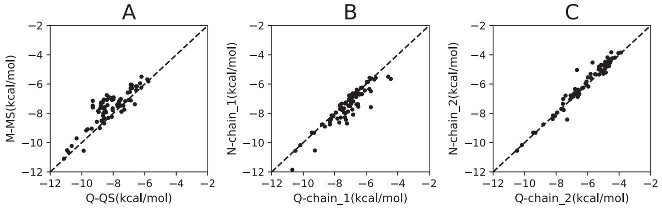


Fig. 4. Comparisons of the binding free energies of the ligand conformations for the 88 complexes. chain_1: the chain with the largest number of ligand-binding residues; chain_2: the chain with the biggest difference in binding free energy between the N and Q conformations.

the more residues participating in the interaction. Then, we compared the binding free energy of the N-QS and Q-QS, N-MS and M-MS in Figs. 2B and 2C, respectively. It is interesting to see that for the docking conformations (whether QS or MS), there are more than 50 percent complexes of docking conformations with even lower binding free energy than the corresponding complexes of native conformations.

For each complex from the selected set, the RMSD of the docking conformation to the native conformation is calculated. The RMSDs of the M and Q conformations for the complexes in the selected set are shown in Fig. 3. As shown in Fig. 3A, for most complexes (72.73 percent), the Q conformations are of lower RMSD than the M conformations. For each of the three RMSD bins below 3 Å, the percentage of the Q conformation ligands is much higher than the M conformation ligands. For example, for the very accurate ligand conformations with RMSD less than 1 Å, the percentage for Q conformation is around 27 percent, much higher than that for the M conformation (~15 percent). The mean RMSD of the Q conformations is 3.84 Å with a standard deviation of 3.31 Å while the mean RMSD of the M conformations is 5.43 Å with a standard deviation of 3.26 Å. Moreover, we computed the percentage of complexes with RMSD lower than a specified cutoff and the results are summarized in Fig. 3B. It shows there are more Q conformations than M conformations at all cutoffs. For very accurate conformations with RMSD $\leq$ 1.0 Å, the percentages of Q and M conformation are 26.97 and 10.22 percent, respectively; while for acceptable conformations with RMSD $\leq$ 2.0 Å, the percentages are 41.75 and 19.11 percent, respectively. These data show that docking with QS in general can achieve conformation with lower RMSD than using the MS.

On the one hand, we analyzed the correlation between experimental binding affinity data and the free energy of different complex structures. The PDBbind database hosts a comprehensive set of experimental binding affinity data for protein-ligand interactions. In the 2019 version of PDBbind, there are 17,679 protein-ligand complexes in total. A total of 233 samples from our dataset have binding affinity data in PDBbind. These data are used to measure the correlation between the binding energies of the docked conformations and the experimental binding affinities. The Pearson's correlation coefficient is 0.44 for the QS-docked conformations. In comparison, the Pearson's correlation coefficient is lower for the MS-docked conformations (i.e., 0.39). This also suggests the importance of using complete binding information for molecular docking, i.e., docking with complete QS is more acceptable.

On the other hand, Figure S3, available online presents a comparison between the solvent accessible surface areas (SASA) of ligands that were docked to QS and MS. It shows that ligands docked to MSs mostly have higher SASAs than those docked to QSs, suggesting the incompleteness of binding information in MS.

As show from Fig. 3, for $>$ 70 percent complexes, the Q conformation has a lower RMSD than the M conformation. Especially, there are 1,181 complexes with Q conformation RMSD $\leq$ 2.0 Å while M conformation RMSD $>$ 2.0 Å. Figure S4, available online shows an example where the Q conformation has a lower RMSD than the M conformation, which is for the protein "6,7-dimethyl-8-ribityllumazine synthase" (PDB ID: 1KYV). The QS is a homo pentamer and the ligand RBF is located at the interface between the chain A in green and the chain B in cyan (Figure S4A, available online). The ligand binds 9 and 4 residues in chains A and B, respectively. According to the description of the structure in [30], the residue His94 (lemon sticks in Figure S4B, available online) is highly conserved, which directly determines the interaction with the ligand by a stacking interaction. The distance between the aromatic ring of RBF and the side chain of this residue is 3.6 Å in the N conformation (red sticks in Figure S4B, available online), 3.5 Å in the Q conformation (blue sticks in Figure S4C, available online) and 4.0 Å in the M conformation (magenta sticks in Figure S4D, available online). Figure S4E, available online shows that docking using the QS results in a very accurate Q conformation with 0.34 Å RMSD (blue sticks), which are significantly lower than that of the M conformation (magenta sticks) from docking with the MS (RMSD = 7.03 Å).

3.3 Docking Using the Quaternary Structure is Not Always the Best Choice

As shown in Fig. 4, there are some complexes that the RMSDs of the M conformations are lower than the corresponding Q conformations. We are especially interested in those complexes, for which the M conformation RMSD is $\leq$ 2.0 Å while the Q conformation RMSD is $>$ 2.0 Å. In total, there are 88 (1.82 percent) such complexes.

For example, the QS (Figure S5A, available online, PDB ID: 2WGU) for the "human adenovirus serotype 37 fibre protein in complex binding with a sialic acid derivative" is a homo trimer [31]. The binding pocket is formed by 4 residues from chain A (cyan structure in Figure S5A, available online) and 3 residues from chain B (green structure in

Figure S5A, available online). The Q conformation (blue sticks in Figure S5B, available online) has a high RMSD (5.53 Å). In contrast, the M conformation (magenta sticks in Figure S5B, available online) is very accurate with 0.73 Å RMSD, which was obtained by docking with the MS of chain A. A closer inspection on the 'bad' Q conformation suggests that it in fact has a strong interaction with the other chain, showing good geometry complementarity with the structure of chain B (Figure S5C, available online). The binding free energy between the docking ligand and chain B receptor in the Q conformation is even lower than the N conformation (Figure S5D, available online) (-6.03 vs -5.86 kcal/mol). This explains that the 'bad' Q conformation was generated due to more constraints have been considered for chain B than chain A during the docking procedure.

The higher RMSD of the Q conformation for most of the 88 complexes can be explained similarly by the binding free energy. Fig. 4 shows a detailed comparison of the binding free energies of the ligand conformations for all the 88 complexes. We can see that the binding free energies of the Q-QS for 75 complexes (85 percent) are lower than the corresponding M-MS (Fig. 4A). This is due to the interaction with multiple chains in the Q conformation. For each QS, the chain with the largest number of ligand-binding residues is denoted by chain_1 (the same as the selection of MS); and the chain with the biggest difference in binding free energy between the N-chain and Q-chain complexes is denoted by chain_2. Fig. 4B shows that for most complexes (69 percent), the binding free energies of the Q conformations to chain_1 are higher than the corresponding N conformations. However, for 70 percent complexes, the binding free energies of the Q conformations to chain_2 are significantly lower than the corresponding N conformations, which explains the higher RMSDs of these Q conformations are due to the optimization of the interaction between chain_2 and the ligand. These conformations have the potential to be 'correct' as the binding free energy is even lower than the corresponding N conformation, which is just a snapshot from many dynamic states.

The proportion of the number of binding residues from chain_1 (over the total number of binding residues) was calculated. It suggests for 56 out of the 88 cases (64 percent), the proportion values are higher than 70 percent, i.e., with an uneven distribution of binding residues. On the contrary, for the cases that QS yield more accurate docking conformations, only 39 percent of them are of uneven distributions. This suggests that the evenness of binding residues distribution along receptor chains can influence the molecular docking as well.

In addition, we compared the BUs and AUs for these 88 cases and found that 54 of them do not have experimental QS. It is possible that such QSs may have errors or their native ligands are not in energetically optimal position (as analyzed above). In such case, it is reasonable to dock using MS.

3.4 The Answers to the Questions Raised in the Introduction

Based on the experiments and analysis conducted above, we should be able to answer the questions raised in the Introduction now.

1. How common are multi-chain protein-ligand interactions? For biologically relevant protein-ligand interactions, there are as high as 18.6 percent ligands to bind multiple chains simultaneously in the QS. Thus, it is very common for multi-chain protein-ligand interactions.
2. What is the advantage of using QS over MS? The QS is believed to be the functional form of the protein. In our experiments, for over 70 percent complexes, docking with QS yield more accurate ligand conformations than with MS. This is realized by optimizing the ligand interaction to multiple chains in QS, which is missed in MS.
3. When is it acceptable to use MS? Docking with QS is preferred as shown in the above conclusion. However, MS may be better for docking when it is clear that the protein's functional form is in monomer state, or there are obvious errors in the QS.

4 Conclusion

Based on molecular docking experiments and binding affinity estimations, we performed a large-scale comparison of the protein-ligand interactions under the forms of both MS and QS. We found 81.4 percent ligands interacts with only one chain. For these structures, it is reasonable to use MS in the study of protein-ligand interactions. However, there are as high as 18.6 percent ligands binding multiple chains in QS simultaneously. We found that for over 70 percent complexes with multi-chain binding pockets, QS-based docking yields more accurate ligand conformations than MS-based docking. It is thus highly recommended to make use of QS for future studies in protein-ligand interactions.

Acknowledgments

The work was supported by National Natural Science Foundation of China (NSFC 11871290 and 61873185), Fundamental Research Funds for the Central Universities, Fok Ying-Tong Education Foundation (161003), KLMDASRT and the Thousand Youth Talents Plan of China. Yajun Dai and Yang Li are contributed equally to this work.

References

[1] P. W. Rose *et al.*, "The RCSB protein data bank: Integrative view of protein, gene and 3D structural information," *Nucleic Acids Res*, vol. 45, pp. D271–D281, Jan. 4 2017.

[2] E. Krissinel and K. Henrick, "Inference of macromolecular assemblies from crystalline state," *J. Mol. Biol.*, vol. 372, pp. 774–797, Sep. 21, 2007.

[3] A. Roy, J. Yang, and Y. Zhang, "COFACTOR: An accurate comparative algorithm for structure-based protein function annotation," *Nucleic Acids Res.*, vol. 40, pp. W471–W477, Jul. 2012.

[4] J. Yang, A. Roy, and Y. Zhang, "Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment," *Bioinformatics*, vol. 29, pp. 2588–2595, Oct. 15, 2013.

[5] M. Brylinski and J. Skolnick, "A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation," *Proc. Nat. Acad. Sci. USA*, vol. 105, pp. 129–134, Jan. 8, 2008.

[6] M. N. Wass, L. A. Kelley, and M. J. Sternberg, "3DLigandSite: Predicting ligand-binding sites using similar structures," *Nucleic Acids Res*, vol. 38, pp. W469–W473, Jul. 2010.

[7] D. B. Roche, M. T. Buenavista, and L. J. McGuffin, "The FunFOLD2 server for the prediction of protein-ligand interactions," *Nucl. Acids Res*, vol. 41, pp. W303–W307, Jul. 2013.

[8] J. Yang, A. Roy, and Y. Zhang, "BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions," *Nucleic Acids Res*, vol. 41, pp. D1096–D1103, Jan. 2013.

[9] G. Abrusan and J. A. Marsh, "Ligand binding site structure influences the evolution of protein complex function and topology," *Cell Rep.*, vol. 22, pp. 3265–3276, Mar. 20 2018.

[10] G. Abrusan and J. A. Marsh, "Ligand-binding-site structure shapes allosteric signal transduction and the evolution of allostery in protein complexes," *Mol. Biol. Evol.*, vol. 36, pp. 1711–1727, Aug. 1 2019.

[11] G. Abrusan and J. A. Marsh, "Ligand binding site structure shapes folding, assembly and degradation of homomeric protein complexes," *J. Mol. Biol.*, vol. 431, pp. 3871–3888, Sep. 6 2019.

[12] E. D. Levy, "PiQSi: Protein quaternary structure investigation," *Structure*, vol. 15, pp. 1364–1367, Nov. 2007.

[13] Q. Xu and R. L. Dunbrack Jr., "The protein common interface database (ProtCID)–a comprehensive database of interactions of homologous proteins in multiple crystal forms," *Nucleic Acids Res.*, vol. 39, pp. D761–770, Jan. 2011.

[14] S. Bliven, A. Lafita, A. Parker, G. Capitani, and J. M. Duarte, "Automated evaluation of quaternary structures from protein crystals," *PLOS Comput. Biol.*, vol. 14, Apr. 2018, Art. no. e1006104.

[15] S. Dey, D. W. Ritchie, and E. D. Levy, "PDB-wide identification of biological assemblies from conserved quaternary structure geometry," *Nature Methods*, vol. 15, pp. 67–72, Jan. 2018.

[16] S. Dey and E. D. Levy, "Inferring and using protein quaternary structure information from crystallographic data," *Methods Mol. Biol.*, vol. 1764, pp. 357–375, 2018.

[17] F. Da Silva, G. Bret, L. Teixeira, C. F. Gonzalez, and D. Rognan, "Exhaustive repertoire of druggable cavities at protein-protein interfaces of known three-dimensional structure," *J. Med. Chem.*, vol. 62, pp. 9732–9742, Nov. 14 2019.

[18] T. Schmidt, J. Haas, T. Gallo Cassarino, and T. Schwede, "Assessment of ligand-binding residue predictions in CASP9," *Proteins*, vol. 79 no. Suppl 10, pp. 126–136, 2011.

[19] Z. Wang *et al.*, "Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: The prediction accuracy of sampling power and scoring power," *Phys. Chem. Chem. Phys.*, vol. 18, pp. 12964–12975, May 14 2016.

[20] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *J. Comput. Chem.*, vol. 31, pp. 455–461, Jan. 30 2010.

[21] W. J. Allen *et al.*, "DOCK 6: Impact of new features and current docking performance," *J. Comput. Chem.*, vol. 36, pp. 1132–1156, Jun. 5, 2015.

[22] S. Y. Huang and X. Zou, "Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking," *Proteins*, vol. 66, pp. 399–421, Feb. 1, 2007.

[23] R. Wang, L. Lai, and S. Wang, "Further development and validation of empirical scoring functions for structure-based binding affinity prediction," *J. Comput. Aided Mol. Des.*, vol. 16, pp. 11–26, Jan. 2002.

[24] Y. Li and J. Yang, "Structural and sequence similarity makes a significant impact on machine-learning-based scoring functions for protein-ligand interactions," *J. Chem. Inf. Model*, vol. 57, pp. 1007–1012, Apr. 24 2017.

[25] S. Y. Huang and X. Zou, "An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials," *J. Comput. Chem.*, vol. 27, pp. 1866–1875, Nov. 30, 2006.

[26] G. Neudert and G. Klebe, "DSX: A knowledge-based scoring function for the assessment of protein-ligand complexes," *J. Chem. Inf. Model*, vol. 51, pp. 2731–2745, Oct. 24, 2011.

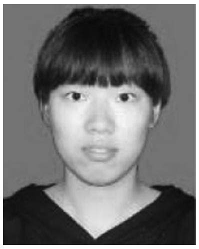
[27] G. Neudert and G. Klebe, "fconv: Format conversion, manipulation and feature computation of molecular data," *Bioinformatics*, vol. 27, pp. 1021–1022, Apr. 1, 2011.

[28] R. Wang, X. Fang, Y. Lu, and S. Wang, "The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures," *J. Med. Chem.*, vol. 47, pp. 2977–2980, Jun. 3, 2004.

[29] C. Antoni *et al.*, "Crystallization of bi-functional ligand protein complexes," *J. Struct. Biol.*, vol. 182, pp. 246–254, Jun. 2013.

[30] S. Gerhardt *et al.*, "The structural basis of riboflavin binding to Schizosaccharomyces pombe 6,7-dimethyl-8-ribityllumazine synthase," *J. Mol. Biol.*, vol. 318, pp. 1317–1329, May 17, 2002.

[31] S. Johansson *et al.*, "Design, synthesis, and evaluation of N-acyl modified sialic acids as inhibitors of adenoviruses causing epidemic keratoconjunctivitis," *J. Med. Chem.*, vol. 52, pp. 3666–3678, Jun 25 2009.



Yajun Dai received the BS degrees in applied statistics from Xidian University, in 2018, she is currently workimg toward the MS degree in the School of Mathematical Sciences at Nankai University, Tianjin, China. Her research interests include protein-ligand interaction, protein structure prediction, machine learning, and deep learning applications in bioinformatics.

Yang Li received the BE and PhD degrees from Nankai University, in 2012 and 2019, respectively. His research interests include protein-ligand interaction and drug design. He is currently a postdoc fellow at the China State Shipbuilding Corporation.

Liping Wang received the BS degree from Nankai University, in 2019, he is currently working toward the graduate degree in the Institute of Automation, Chinese Academy of Sciences. His reseach interests include artificial intelligence in data science.

Zhenling Peng received the PhD degree from the University of Alberta, in 2014. She is an associate professor with the Center for Applied Mathematics, Tianjin University, Tianjin, China. Her research interests include discovery and analysis of sequence-structure/disorder-function relationships in proteins, protein function prediction, and modeling.

Jianyi Yang received the PhD degree from the Nanyang Technological University. He is a professor with the School of Mathematical Sciences, Nankai University, Tianjin, China. He had his postdoctoral training in the University of Michigan. His research interests include protein structure and function prediction, protein structure alignment and RNA structure prediction. He has made significant contributions to the development of many widely used tools, including I-TASSER, trRosetta, COACH, COACH-D, BioLiP, mTM-align and so on. For more information, please visit http://yanglab.nankai.edu.cn/