Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Original Paper

Structural bioinformatics

LS-align: an atom-level, flexible ligand structural alignment algorithm for high-throughput virtual screening

Jun Hu[1,2], Zi Liu[1], Dong-Jun Yu[1,*] and Yang Zhang[2,*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094, [2]Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw, Ann Arbor, MI 48109-2218, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Sequence-order independent structural comparison, also called structural alignment, of small ligand molecules is often needed for computer-aided virtual drug screening. Although many ligand structure alignment programs are proposed, most of them build the alignments based on rigid-body shape comparison which cannot provide atom-specific alignment information nor allow structural variation; both abilities are critical to efficient high-throughput virtual screening.

Results: We propose a novel ligand comparison algorithm, LS-align, to generate fast and accurate atom-level structural alignments of ligand molecules, through an iterative heuristic search of the target function that combines inter-atom distance with mass and chemical bond comparisons. LS-align contains two modules of Rigid-LS-align and Flexi-LS-align, designed for rigid-body and flexible alignments, respectively, where a ligand-size independent, statistics-based scoring function is developed to evaluate the similarity of ligand molecules relative to random ligand pairs. Large-scale benchmark tests are performed on prioritizing chemical ligands of 102 protein targets involving 1,415,871 candidate compounds from the DUD-E (Database of Useful Decoys: Enhanced) database, where LS-align achieves an average enrichment factor (EF) of 22.0 at the 1% cutoff and the AUC score of 0.75, which are significantly higher than other state-of-the-art methods. Detailed data analyses show that the advanced performance is mainly attributed to the design of the target function that combines structural and chemical information to enhance the sensitivity of recognizing subtle difference of ligand molecules and the introduces of structural flexibility that help capture the conformational changes induced by the ligand-receptor binding interactions. These data demonstrate a new avenue to improve the virtual screening efficiency through the development of sensitive ligand structural alignments.

Availability: http://zhanglab.ccmb.med.umich.edu/LS-align/

Contact: njyudj@njust.edu.cn or zhng@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Computer-based virtual screening (VS) becomes increasingly prevalent in drug discovery, and has been exploited as a valuable and economic tool to identify new lead molecules, complementary to the expensive experimental screening. The computational VS methods can be grouped into two categories according to the information they use: protein-centric methods and ligand-centric methods (Roy and Skolnick, 2015). Starting from protein structures, the protein-centric methods can often achieve a better screening performance, as they enable the explicit evaluation of protein-ligand binding interactions through docking (Forli, et al., 2016; Patel, et al., 2017). Nevertheless, the performance highly depends on the quality of the receptor structure, as low-resolution models from protein structure predictions can often degrade the accuracy of the process when the experimental structure is not available (Zhang, 2009). In addition, they suffer from the inherent limitations of the applied protocols.

On the contrary, without relying the structure of proteins, the ligand-centric methods commonly employ known ligands as a seed to identify potential binders by 2D or 3D structural similarity comparisons (Eckert and Bajorath, 2007; Shang, et al., 2017). In the 2D structural virtual screening methods, a small molecule is generally represented as a vector (e.g., 2D fingerprint) with entries indicating the presence or absence of molecular features. These 2D-based methods are widespread in this field, as they can fast and easily figure out the similar active molecules. However, both molecule size and structural complexity can negatively impact

the screening performance of these methods (Holliday, et al., 2003; Willett, 2006). Compared to the 2D methods, the 3D-structure based ligand alignment programs can capture the physical and functional features required for the biological interaction, which is essential for scaffold hopping to infer new ligands starting from existing ligands (Hu, et al., 2017; Quintus, et al., 2009). The ability of scaffold hopping can enhance the VS performance by using the structural and physicochemical information to reduce the false negative rates. Moreover, the programs can provide useful insights for bioisostere replacement, cross-reactivity of existing drugs, and potential off-target interactions (Jennings and Tennant, 2007). Although 3D-based ligand-alignment methods are commonly more time-consuming than 2D-based methods, with the increase of computational power and the improvement of method efficiency, the time cost is becoming an increasingly less critical issue in their applications.

A variety of 3D-based methods have been developed for aligning small ligand molecules, which are mainly built on three different principles, including molecular interaction field-based (Cheeseright, et al., 2008), pharmacophore-based (Sperandio, et al., 2007), and shape-based alignment methods. Among them, the shape-based virtual screening methods, which seek to maximize the shape overlap between a pair of small molecules, have become particularly popular in recent studies. Most of the shape-based approaches use atom-centered, smooth Gaussian functions to model molecular volumes, due to the fact that it helps to gain rapid overlay. For example, Rapid Overlay of Chemical Structures (ROCS), whose source code is not yet open, employs molecular volumes defined by a Gaussian function and the ligand chemical nature (Grant, et al., 1996) for VS. Similarity, Align-It™, which is an open-source algorithm, utilizes a Gaussian description of molecular pharmacophores, although it uses a different optimization approach to seek for the best overlay (Taminau, et al., 2008). LIGSIFT is another open-source shape-based approach, which also uses a Gaussian function to model the molecular volumes for VS, followed by a short Metropolis Monte Carlo simulation to refine the overlap of target molecules (Roy and Skolnick, 2015). Despite the efficiency, there are two critical shortages in these 3D methods. First, since most of the methods are based on shape comparison, they do not provide the atomic-level alignment information, which is often critical for the detail VS data analysis and refinement. Second, these alignments are all rigid-body based, which does not allow flexibility and structural variation of the aligned ligand molecules; this can considerably limit their usefulness and efficiency in practical VS experiments, because many ligand molecules must adopt different shape conformations when bound with different protein receptors.

In this study, we propose a novel structural alignment algorithm, named LS-align, for atomic-level ligand structure comparison. LS-align contains two specific alignment modules, Rigid-LS-align and Flexi-LS-align. While the Rigid-LS-align module focuses on optimal rigid-body structure alignment, the Flexi-LS-align module allows flexible structural comparison, by considering various conformers deformed from the query ligand structures. To examine the strengths and weaknesses, the LS-align algorithm has been carefully benchmarked in a large-scale VS experiment, based on 102 targets from the DUD-E (the Database of Useful Decoys: Enhanced) database, with the result compared favorably with other state-of-the-art 3D-based virtual screening methods, including ROCS, Align-it™ and LIGSIFT. The on-line web server of LS-align, as well as the source code of the program, are made freely available at http://zhanglab.ccmb.med.umich.edu/LS-align/.

2 Methods

LS-align contains two modules, Rigid-LS-align and Flexi-LS-align, for rigid-body and flexible ligand structure alignments, respectively. It is noted that both modules neglect the hydrogen atoms and work only with the heavy atom structures. Details algorithms are described as follows.

2.1 Similarity scores of ligand molecules

Two scoring functions are exploited to quantify the similarity between ligand structures. The first is called LS-score, which solely consider the structural information of the atoms in the two molecules:

$$\text{LS-score} = \text{Max}\left[\frac{1}{N_{target}}\sum_{i=1}^{N_{ali}}\frac{1}{1+d_i^2/d_0^2}\right] \quad (1)$$

where N_{target} is the number of the heavy atoms of the target molecule, N_{ali} is the number of the aligned atom pairs, d_i is the distance between ith pair of aligned atoms and d_0 is a scale to normalize the match difference. 'Max' denotes the maximum value after optimal alignment. The value of LS-score lies between (0, 1), with more similar molecule pairs having a higher LS-score.

To examine the size dependence of the scale of LS-score, Figure 1 shows the average LS-scores calculated from 8,000 ligand pairs, which are randomly selected from the PDB library and have a pair-wise Tanimoto coefficient (calculated based on ligand fingerprint profiles) <0.3, as a function of the number (N_{min}) of the heavy atoms of the smaller ligand molecules. The raw LS-score (rLS-score) is calculated using a constant value (1.5Å) for d_0. The data shows a power-law dependence of the rLS-score on the molecule size, with magnitude of the rLS-score decreased by 1.6 times as the N_{min} increases from 10 to 50. Such a sensitivity to the ligand size for the random ligand pairs renders the absolute value of the rLS-score values meaningless.

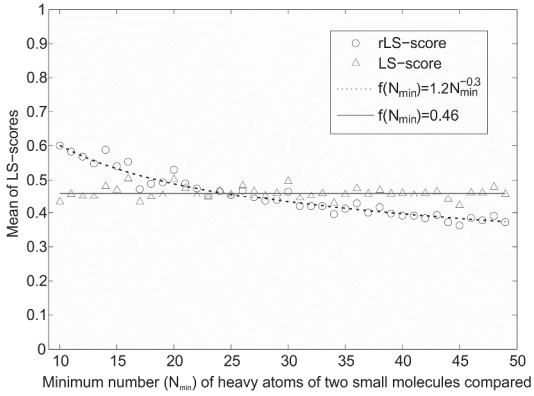

Fig. 1. The average rLS-score (circles) and LS-score (triangles) of random ligand pairs as a function of the minimum number (N_{min}) of the heavy atoms of two small molecules compared. For the rLS-score, a constant scale, d_0 = 1.5 Å, is used; for the LS-score, d_0 is calculated by Eq. (2). The dashed line is a nonlinear least square Marquardt-Levenberg fit of the rLS-score data to a power-law equation, $f(N_{\min}) = aN_{min}^b$, with *a=1.2* and *b=-0.3*. The solid line denotes the horizontal line of LS-score = 0.46.

To rule out the size dependence, following the idea of TM-score (Zhang and Skolnick, 2004) we introduce a size-dependent scale

$$d_0(N_{min}) = 0.55\sqrt[3]{N_{min}-9} + 0.15 \quad (2)$$

where the parameters are obtained empirically by making the corresponding power-law equation, $f(N_{min}) = aN_{min}^b$, to a horizontal line in a separate training dataset. The data in Figure 1 shows that the LS-score value with the new scale indeed flattens the molecule size dependence, with an approximately constant value of LS-score≈0.46 for unrelated random molecule pairs.

The second scoring function exploited is an extension of LS-score,

named PC-score, which combines the structural similarity with the atom mass and chemical bond similarities:

$$\text{PC-score} = \text{Max}\left[\frac{1}{N_{target}}\sum_{i=1}^{N_{ali}}\left(\frac{w_d}{1+d_i^2/d_0^2}+\frac{w_m}{1+\Delta m_i^2/\Delta m_0^2}+w_b BWJS_i\right)\right] \quad (3)$$

where Δm_i is the difference of the relative masses of the *i*th pair of aligned atoms, with Δm_0 being a constant value to scale Δm_i. $BWJS_i$ is a weighted Jaccard score which measures the similarity of the chemical bonds through which the *i*th pair of atoms interact with other atoms in the two molecules (e.g., *A* and *B*):

$$BWJS_i = \frac{G(S_{Ii})}{G(S_{Ai})+G(S_{Bi})-G(S_{Ii})} \quad (4)$$

where S_{Ai} and S_{Bi} represent the sets of chemical bonds that connect to the *i*th pair of atoms (A_i and B_i) from all other atoms in the molecules; S_{Ii} is the intersection set of the chemical bonds in S_{Ai} and S_{Bi} according to their bond type. Here, six types of chemical bonds, including Single, Double, Triple, Amide, Aromatic and Dummy bonds, are considered. $G(x)=\sum_{k=1}^{N_x} w(b_k)$, with N_x being the total number of chemical bonds of the bond set *x*, b_k the bond type of *k*th bond, and $w(b_k)$ the specific weight of b_k-th bond type. The specific weight for each bond type is listed in Table S1 in Supporting Information (SI). In Text S1, we present an illustrative example to help further explain the $BWJS_i$ score.

In Eq (3), w_d, w_m and w_b (=0.45, 0.10 and 0.45, respectively) are the weighting factors to balance the score terms, which are determined by maximizing the correlation between PC-score and the root-mean-square deviation of ligand structures, i.e., $RMSD_{LS}$, of the same set of 8,000 ligand pairs mentioned above. Here, $RMSD_{LS}$ for Ligand *A* and *B* is defined by

$$RMSD_{LS} = \max(R_{AB}, R_{BA}) \quad (5)$$

where R_{AB} is defined as

$$R_{AB} = \sqrt{\frac{1}{n_A}\sum_{i=1}^{n_A} \min_{j=1,\cdots,n_B} d_{ij}^2} \quad (6)$$

Here, n_A and n_B are the numbers of atoms of in *A* and *B*, respectively, and d_{ij} is the distance of *i*th atom on *A* and *j*th atom on *B*, after superposing *A* and *B* by Rigid-LS-align (described in Section 2.2), with the 'min' running through all atoms on *B*. Since only the minimum distance is considered, the $RMSD_{LS}$ allows redundant atom correspondence, which is different from the standard RMSD that considers deviation of *aligned* atom pairs. One advantage of $RMSD_{LS}$ over the *RMSD* is that the latter only considers the local structure deviation in the aligned regions and an optimization on *RMSD* can drive the optimization search for shorter alignments, while $RMSD_{LS}$ does not require pre-alignment of atom pairs and can count for the global similarity of two superposed molecules.

Here we note that we used $RMSD_{LS}$ as a gold standard to train the weight parameters of PC-score because we consider it an appropriate measurement of global structural similarity and we assume that a reasonable alignment should result in a low $RMSD_{LS}$ score. However, the $RMSD_{LS}$ cannot be directly used for optimizing structural alignment because the $RMSD_{LS}$ does not depend on alignment on its own. For this reason, we have selected to use PC-score instead of $RMSD_{LS}$ to optimize the structural alignment of ligand pairs. Another reason for the selection of PC-score is that PC-score contains additional chemical information that help assess the similarity of the chemical properties of the ligands. Nevertheless, as described above, the parameters used to define the PC-score are a fitting or mimic of the $RMSD_{LS}$ score which is generated by the superposition matrix that depends on the alignment.

Figure S1 shows the average PC-scores of the 8,000 random ligand pairs versus the size of the smaller ligand molecules compared (N_{min}). By coincidence, the PC-score with rescaled d_0 as Eq. (2) has also an approximately constant value of ≈0.46, which is independent of atom number for the random ligand pairs; but the rPC-score score with a constant d_0 has still a slight dependence on the molecule size (Fig. S1).

2.2 Rigid ligand structural alignment by Rigid-LS-align

2.2.1 Construction of initial ligand structural alignment

Given two ligand molecules (named as 'query' and 'template' for easy description), LS-align starts with the construction of initial structure alignments. Three kinds of quickly identified initial alignments are considered, which all employ an enhanced greedy search (EGS) algorithm to search against their score matrices (Fig. 2). The main difference in the three initial alignments lies at the score matrices, i.e.,

$$\begin{cases} S_{mass}(i,j) = 1/(1+\Delta m_{ij}^2/\Delta m_0^2) \\ S_{vdw}(i,j) = 1/(1+\Delta vdw_{ij}^2/\Delta vdw_0^2) \\ S_{bond}(i,j) = BWJS_{ij} \end{cases} \quad (7)$$

where Δm_{ij}, Δvdw_{ij}, and $BWJS_{ij}$ are, respectively, the relative mass difference, van der waals radius difference, and weighted Jaccard score of the chemical bonds between *i*th atom in the query and *j*th atom in the template molecule, which all do not require superposition.

Given a score matrix (denoted by $\boldsymbol{S}_{init}$, with row and column running along query and template respectively, Fig. 2), EGS first employs a greedy search strategy to generate an intermediary alignment under the constraint that each row and column can only be selected once. In this strategy, it starts with the selection of the max value in $\boldsymbol{S}_{init}$, e.g, at the lattice point (i, j), and aligns the *i*th atom in the query and the *j*th atom in the template. Accordingly, all elements at *i*th row and *j*th column are set to be invalid. Next, it selects again the max value of the remaining valid elements in $\boldsymbol{S}_{init}$, and aligns the corresponding atom pair in the query and template molecules, followed by the setting of all elements of the corresponding row and column as invalid in the matrix. Such procedure is repeated until no valid element remains in $\boldsymbol{S}_{init}$.

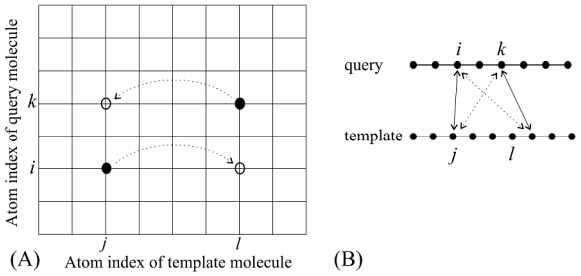

Fig. 2. Illustration of EGS algorithm. (A) and (B) are representations of the same swap movement of alignment, but one in 2D lattice system and another in 1D alignment.

The intermediary alignment by the greedy search strategy is apparently not an optimal solution, as the selection of local maximums at each step can ignore the optimal combinations of other elements that may have lower values. The second step of EGS is to improve the intermediary alignment through iterative switch movements. Let's denote $M = \{(i,j)_h\}_{h=1}^{H}$ as the alignment matrix, where $(i,j)_h$ is the *h*th pair of aligned atoms, and *H* is the total number of the aligned atom pairs. For a given aligned pair (i, j), we consider a switch of the alignment to another alignment pair (i, l) at the same row, where another existing alignment pair at the *l*th column must be switched to (k, j) to keep the condition that each atom can be aligned only once (Fig. 2). The score change due to this switch movement can be calculated by

$$\Delta S_l = [S_{init}(i,l) - S_{init}(i,j)] + \sum_{k=1}^{N_{row}} \lambda(k,l)[S_{init}(k,j) - S_{init}(k,l)] \quad (8)$$

where N_{row} is the number of rows in $\boldsymbol{S}_{init}$. $\lambda(k,l) = 1$ if M contains an aligned pair (k, l); or $\lambda(k,l) = 0$ otherwise. For the aligned pair (i, j), we calculate ΔS_l for all columns and select the maximum value, ΔS_{lmax}. If $\Delta S_{lmax} > 0$, we will update the hth aligned pair (i, j) in M to be (i, l_{max}), and change (k, l_{max}) to (k, j) in case that $\lambda(k, l_{max})$=1. Here, we do not consider the row-based switches since these are already included in the column-based movements due to the symmetry of alignment. Next, starting from the updated M, we repeat the above switch movements again to process the (h+1)-th aligned atom pair. When all H aligned pairs are all processed, this procedure is terminated if no aligned pair in *Ali* is modified; otherwise, we will repeat the switch refinement process for the H aligned pairs until the termination condition is reached. The final alignment will be returned when EGS terminated.

The EGS is processed for each of the score matrices in Eq. (7), from which three different initial alignments are constructed for the next step of heuristic iteration and refinement.

2.2.2 Alignment refinement through heuristic iterations

To find the optimal alignment of the query (Q) and template (T) structures that has the maximum similarity score according to LS-score (Eq. 1) or PC-score (Eq. 3), the three initial alignments obtained above are further refined through a heuristic iterative optimizing process that was inspired from that used in TM-align (Zhang and Skolnick, 2005) (Fig. 3).

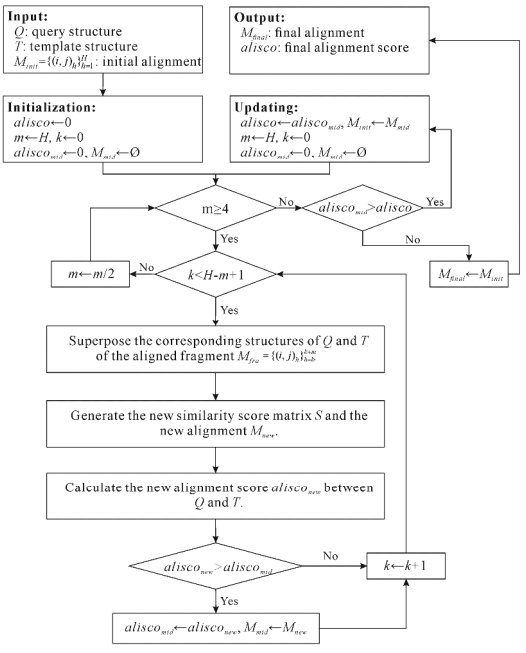

Fig. 3. Flowchart of the heuristic iteration procedure used by LS-align to search for the optimal alignments of ligand structures.

Starting with an initial alignment $M_{init} = \{(i,j)_h\}_{h=1}^{H}$, we can consider a fragment of m aligned atoms, e.g., $M_{fra} = \{(i,j)\}_{h=k}^{k+m}$, and superpose the aligned fragment of the query to the corresponding atoms of the template using the Kabsch's rotation matrix (Kabsch, 1976), where a new similarity score matrix can be calculated based on the Kabsch superposition, using either the LS-score: $S(i,j) = 1/(1 + d_{ij}^2/d_0^2)$, or the PC-score: $S(i,j) = w_d/(1 + d_{ij}^2/d_0^2) + w_m/(1 + \Delta m_{ij}^2/\Delta m_0^2) + w_b BWJS_{ij}$, for all atom pairs between query and template. Starting from the new scoring matrix $S(i,j)$, we employ the EGS algorithm (Fig. 2) to identify a new alignment between query and template, which will be used for creating a newer scoring matrix and a newer structure alignment. Such process will be repeated iteratively till the alignment is unchanged.

Because the converged alignment between query and template is sensitive to the selection of the fragment alignment, $M_{fra} = \{(i,j)_h\}_{h=k}^{k+m}$, we run multiple processes, with each process starting with m=H, H/2, ..., 4, respectively. When m<H, we will run a sliding window with size m through the H aligned atom pairs to create H-m+1 alignment fragments that will be used for iteration search. At each step of the heuristic iterations, an alignment score, on either LS-score or PC-score, will be recorded, where the alignment corresponding to the maximum score in all the iterations, from all three initial alignments, will be reported as the final resultant alignment, together with their alignment scores (Fig. 3).

2.3 Flexible ligand structural alignment by Flexi-LS-align

In practical biological processes, most ligand molecules need to change their conformation and shape, e.g. by rotating the rotatable bonds, to fit the diversity of binding pockets of different receptors. As an example, Figure S2 shows three conformers of the same ATP (Adenosine-5'-triphosphate) ligand, when they are bound with different protein-ATP complexes including the ABC transporter HlyB (PDB ID:1xef), RadA C-terminal ATPase (4a6x), and Cytoplasmic Asparaginyl-tRNA Synthetase (2xti). The conformations of the ATP are dramatically changed, with an average RMSD =3.99 Å between the three conformer pairs. To address the difficulty of rigid-body alignment on these cases, we propose a new module, Flexi-LS-align, for flexible structure alignment, in which multiple conformers of the ligands are created based on the superposed structure, by rotating the rotatable bonds. Consequently, multiple structural alignments are generated on the different conformers, where the alignment with the highest similarity score is selected as the final alignment.

To construct alternative conformers of ligands, Flexi-LS-align take as input the mol2 format file of the query structure, and then obtain a list of all the rotatable single bonds (if one of these bonds is broken, the molecule will be divide into two unrelated parts). For increasing the speed and efficiency of alignments, Rigid-LS-align first aligns the initial query conformer to the template and removes the well-aligned rotatable single bonds from the rotatable bonds list. If the number of the remaining rotatable single bonds is still larger than a pre-defined constant N_{RS} (N_{RS}=3 in this study), we select the N_{RS} most important bonds (i.e., rotating these bonds can maximally change the initial conformer) to construct the alternative conformers.

The selected bonds are rotated with various rotation angles to construct alternative conformers. The rotation angle of each selected bond ranges from -180° to 180° with a step size 60°. Here, we have examined the Flexi-LS-align with different step size from 10° to 90° on the 8,000 ligand pairs; it was found that the average PC-score increases with finer angle steps but the time cost increases exponentially, where the 60° step size gives a reasonable PC-score/time rate. Next, we select the top 10 most reasonable conformers, which have no atom pair with a distance less than 0.8 times of the van der Waals radius, as the candidate conformers of the query molecule used for flexible structure alignments. For each of the candidate conformer pair, an optimal alignment is obtained through the process described in Section 2.2. Finally, the conformer pairs resulting the highest LS-score or PC-score will be returned, together with their alignments and the alignment scores.

2.4 Statistical significance of molecular similarity

As the LS-score and PC-score cannot tell on their own how significant

the alignments are, we introduce an additional quality, *p*-value, which equals to the probability of having the specific scores above a certain value, to assess the statistical significances of the alignments relative to the random ligand structure comparisons.

In Figures S3 and S4, we show that the probability densities of LS-score and PC-score, which are calculated for the 1,000,000 pairs of randomly selected ligand structures from the PDB, follow well the type-I extreme value distribution (EVD) (Embrechts, et al., 1997):

$$f(z) = e^{-z-e^{-z}} \tag{9}$$

where $z = (s-\mu)/\sigma$ and s denotes the LS-score or PC-score. The location and scale parameters of the EVD can be calculated by

$$\begin{cases} \mu = a + b\ln N_Q + c\ln N_T \\ \sigma = d + e\ln N_Q + f\ln N_T \end{cases} \tag{10}$$

with N_Q and N_T being the numbers of the heavy atoms in the query and template molecules, respectively. The values of the parameters, *a*-*f*, are obtained by fitting Eq. (10) with the data in Figs. S3-4 through linear regression and listed in Table S2. Given the extreme value distribution of LS- and PC-scores, the *p*-value can be obtained by integrating Eq. (9):

$$p-\text{value} = \int_z^{\infty} f(z)\,dz = 1 - e^{-e^{-z}} \tag{11}$$

2.5 Benchmark dataset and evaluation indexes

The DUD-E (Mysinger, et al., 2012) database is used as a standard dataset to test the LS-align for VS prediction. DUD-E contains a list of 102 proteins, each with on average 224 active ligands. For each active ligand, there are around 50 similar but non-active ligands, called decoys, to challenge the VS procedure. The 102 proteins span diverse categories, including 26 kinases, 15 proteases, 11 nuclear receptors, 5 GPCRs, 2 ion channels, 2 cytochrome P450s, 36 other enzymes, and 5 miscellaneous proteins. To avoid bias of test, we remove the duplicate entries to make sure that each ligand has only one conformer in the database. The final list of all the 102 proteins along with the number of active and decoy ligands can be found in Table S3, with the whole dataset downloadable at https://zhanglab.ccmb.med.umich.edu/LS-align/Database.html.

To examine the performance of the structural alignment methods in VS, we first use the methods to match the seed ligand from the co-crystallized ligand-protein complex in DUD-E with all other active and decoy compounds associated with the same protein receptor. The alignment scores calculated by the methods are then used to rank the compounds. Two indexes, enrichment factor (EF) and hit rate (HR), are calculated by

$$\begin{cases} EF^{x\%} = \dfrac{TP^{x\%}/N^{x\%}_{selected}}{N_{actives}/N_{total}} \\ HR^{x\%} = \dfrac{EF^{x\%}_{actual}}{EF^{x\%}_{ideal}} \times 100 \end{cases} \tag{12}$$

where $x\%$ represents the fraction of screened library selected for evaluating the methods, which is usually set to 1%, 5%, and 10%. $TP^{x\%}$ and $N^{x\%}_{selected}$ are the number of true positives and the number of all candidate compounds in the top $x\%$ of the screening library selected by the testing methods. $N_{actives}$ and N_{total} denote the total numbers of the active molecules and all compound candidates screened, respectively. Thus, $EF^{x\%}$ represents the enrichment factor of the testing methods in prioritizing active compounds compared to the random selections at the cutoff $x\%$. $EF^{x\%}_{actual}$ and $EF^{x\%}_{ideal}$ are the enrichment factor values of the actual and the ideal scores, respectively, where $HR^{x\%}$ represents the ratio of the actual over the best possible score function for prioritizing the active compounds from the compound library.

In addition to these two evaluation indexes, we also use AUC, which is the area under the receiver operating characteristic (ROC) curve, to assess the tested methods. The AUC value varies from 0 to 1, with 0.5 indicating the performance of random selections.

3 Results and discussions

3.1 PC-score achieves better VS results than LS-score

Since the quality of the scoring function has impact on the ligand alignment and VS, we first compare the two proposed scoring functions, LS-score and PC-score, in Table 1, which lists the enrichment factor and AUC values on different scoring functions but using the same search engine from the Rigid-LS-align program.

It is shown that the PC-score outperforms the LS-score in all four evaluation indexes. The average $EF^{1\%}, EF^{5\%}, EF^{10\%}$, and AUC of PC-score are 20.1, 6.9, 4.3 and 0.74, which are 67.5%, 46.8%, 30.3% and 7.2% higher than that of LS-score, respectively. The *p*-values in student t-test are all below 10^{-6}, indicating that the differences are statistically significant. Figure S5 presents a head-to-head comparison of the AUC scores by PC- and LS-scores, where PC-score outperforms LS-score in 74 cases (or 72.5%), while LS-score does so in 28 cases out of the 102 DUD-E protein targets.

Table 1. VS results by LS- and PC-scores on DUD-E database. Values in parentheses are *p*-values in student t-test relative to PC-score.

Score function	$EF^{1\%}$(*p*-value)	$EF^{5\%}$(*p*-value)	$EF^{10\%}$(*p*-value)	*AUC*(*p*-value)
LS-score	12.0 ($<10^{-14}$)	4.7 ($<10^{-15}$)	3.3 ($<10^{-11}$)	0.69 ($<10^{-6}$)
PC-score	20.1	6.9	4.3	0.74

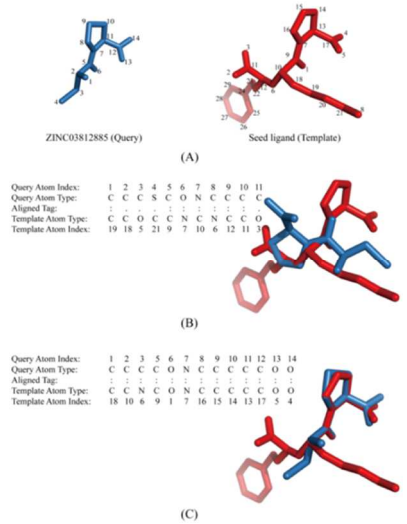

Fig. 4. Illustrative example of ligand structure alignments between ZINC03812885 (blue sticks) and seed ligand (red sticks) generated by Rigid-LS-aligns based on different scoring functions. (A) The original structures of the two molecules. (B) The alignment of the two ligands using LS-score-based Rigid-LS-align program. (C) The alignment of the two ligands using PC-score-based Rigid-LS-align program. The numbers mean the indexes of the heavy atoms in the original input file. ':' denotes strongly aligned atom pairs of the aligned distance <1Å. '.' means weakly aligned atom pairs of the aligned distance in 1-2 Å.

One of the major reasons for PC-score to outperform the LS-score is that LS-score is purely structure-based while PC-score includes mass and chemical bonding information of the aligned atoms in addition to the structure similarity. The mass and chemical binding information helps LS-align to find better matches between the molecules. Figure 4 shows a representative example to align an ACE active molecule ZINC03812885

and the seed ligand taken from ACE receptor structure (PDBID: 3bkl). The LS-score based algorithm identified a suboptimal alignment with only 7 out of 11 atom pairs that have the distance below 1 Å, resulting a RMSD=1.17 Å and LS-score=0.25, while the PC-score based program identified an alignment with all 13 atom pairs in a distance below 1 Å, which has a RMSD=0.44 Å and LS-score=0.62. Because LS-score does not contain chemical information, only 5 pairs of aligned atoms (C1, C2, C5, C9, C10) have identical atom type (or an atom identity =5/11=45%). Due to the guidance of mass and chemical bonds, the alignment by PC-score has 12 out of 13 aligned pairs with identical atom type (or an atom identity =92%). Such chemical restraints help to guide the Rigid-LS-align program to quickly identify optimal alignments that contain more chemically similar atom pairs, which often have a closer geometrical similarity. This more precise alignment of ligand molecules also helps increase the accuracy of the ligand selection in the VS experiments.

As PC-score contains three component terms, in Table S4 we examined the VS performance of the three terms from LS-, Mass- and BWJS-scores, in terms of the enrichment factors and AUC. It was shown that the performances of LS- and BWJS-scores are comparable and both consistently outperform that of Mass-score. However, a combination of all three terms, i.e., the PC-score, significantly outperforms all the individual scoring functions, demonstrating again the usefulness of the integration of multiple complementary information in the structural alignment search.

3.2 EGS identifies faster and more accurate alignments than the Hungarian algorithm

The Hungarian algorithm (Kuhn, 1955; Munkres, 1957) is a traditional combinatorial optimization approach, which has been widely used for solving the assignment problems including the sequence-order independent structure alignments. To compare EGS with the Hungarian algorithm, we constructed a new program (H-LS-align) by replacing EGS with the Hungarian algorithm in the LS-align. Figure S6 presents a head-to-head comparison of the two programs on the same set of 8,000 ligand pairs, in regard to the PC-score and the CPU time running on a 2.8 GHz IBM NeXtScale machine.

It is shown from Fig. S6A that LS-align generates alignments with a higher PC-score than H-LS-align in 5,167 out of the 8,000 cases, while H-LS-align outperforms LS-align in 1,956 cases. The average PC-scores identified by LS-align and H-LS-align are 0.408 and 0.387, with a *p*-value in the Wilcoxon signed rank t-test being 6.0×10^{-13}, suggesting that the PC-score difference of the two programs is statistically significant. Particularly, the LS-align is much faster than H-LS-align. As shown in Fig. S6B, the average CPU time by the LS-align and H-LS-align is 0.015s and 0.247s, where the longest running time for them is 0.73s and 10.94s, respectively. Overall, these data suggest that EGS can identify more accurate alignments within a much shorter time than the Hungarian algorithm, on the same framework of the LS-align program.

3.3 Comparison of Flexi-LS-align and Rigid-LS-align in VS

To examine the impact of conformational flexibility to the structural alignment programs, we collected 30,000 ATP molecule pairs randomly from the PDB library. Figure 5 lists the histogram of the $RMSD_{LS}$ by Rigid-LS-align and Flexi-LS-align. It is shown that Flexi-LS-align generated alignments with a $RMSD_{LS}$ <1 Å for 5,262 (about 17.5%) ATP pairs and below 2 Å for 24,476 (about 81.6%) ATP pairs, whereas Rigid-LS-align did so for 1,751 (about 5.8%) and 17,106 (about 57.0%) ATP pairs, respectively. The average $RMSD_{LS}$ of the alignments by Flexi-LS-align and Rigid-LS-align programs is 1.52 Å and 2.08 Å, respectively. Such reduction of $RMSD_{LS}$ by Flexi-LS-align is expected because a considerably large set of conformers have been used for the structural alignment comparisons; the data also confirms that the introduction of structural flexibility can indeed help find closer matches of the same ligand molecules from different binder systems.

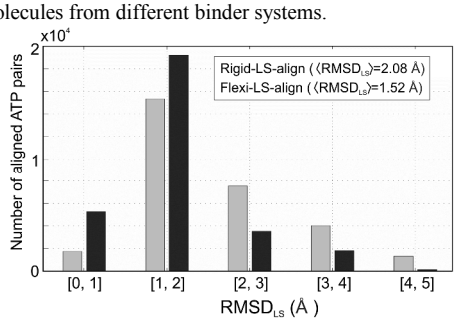

Fig. 5. Histogram of $RMSD_{LS}$ of the aligned ATP structure pairs by Rigid-LS-align and Flexi-LS-align, respectively, on 30,000 random ATP pairs from the PDB.

In Table 2, we further compare the VS results of Flexi-LS-align and Rigid-LS-align programs on the DUD-E dataset using the *EF* and *HR* evaluation indexes. It can be seen that Flexi-LS-align consistently outperforms Rigid-LS-align by generating higher enrichment factor values, no matter if LS-score or PC-score is used. When using LS-score, Flexi-LS-align can achieve an average $EF^{1\%}$ and $HR^{1\%}$ of 14.7 and 24.0, which are about 22.5% and 21.2% higher than that of Rigid-LS-align; if using PC-score, the $EF^{1\%}$ and $HR^{1\%}$ values by Flexi-LS-align are increased to 22.0 and 35.9, which are 9.5% and 8.8% higher than that by Rigid-LS-align, respectively.

Table 2. Virtual screening performance comparisons of Rigid-LS-align and Flexi-LS-align under different scoring functions on DUD-E database. 'LS' and 'PC' refer to 'LS-score' and 'PC-score' respectively.

Method	$EF^{1\%}$	$EF^{5\%}$	$EF^{10\%}$	$HR^{1\%}$	$HR^{5\%}$	$HR^{10\%}$
Rigid-LS-align (LS)	12.0	4.7	3.3	19.8	23.7	33.2
Flexi-LS-align (LS)	14.7	5.6	3.7	24.0	28.2	36.6
Rigid-LS-align (PC)	20.1	6.9	4.3	33.0	34.8	43.5
Flexi-LS-align (PC)	22.0	7.2	4.5	35.9	36.4	45.1

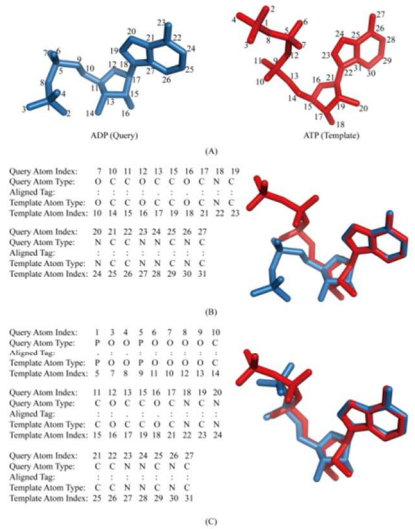

Fig. 6. Example of ligand structure alignments by Rigid-LS-align and Flexi-LS-align for ADP (blue sticks) and ATP (red sticks). (A) Original structures of ADP and ATP ligands. (B) Alignment of two ligands by Rigid-LS-align. (C) Alignment of two ligands by Flexi-LS-align. The numbers are the atom indexes in original structures. ':' denotes strongly aligned atom pairs of distance <1Å. '.' means weakly aligned pairs of distance <2 Å.

Figure 6 presents an illustrative example of alignment of the ADP to ATP molecules by Rigid-LS-align and Flexi-LS-align, respectively. Due to the adoption of the flexibility of the conformers, Flexi-LS-align identified a much higher number of the atom pairs (21 vs 16 to Rigid-LS-align) that are closely matched between the two molecules with a distance <1 Å. The ability in identifying more precise match of the ligand molecules partially explains why the Flexi-LS-align could have a better ability in generating higher enrichment factors in the VS experiment as demonstrated by the data in Table 2.

3.4 Comparison of LS-align with other methods in VS experiment based on rigid-body alignment

Most ligand structural alignment methods in literature do not consider the flexibility of ligand conformers. To have a fair comparison of LS-align with these programs, we first employ the Rigid-LS-align module to obtain the rigid-body alignments between the query and the template compounds in the library. Table 3 summarizes the VS results of Rigid-LS-align, compared to LIGSIFT (Roy and Skolnick, 2015), Align-It™ (Taminau, et al., 2008), and ROCS (OpenEye Inc) (Grant, et al., 1996), on the 102 DUD-E protein targets. Here, given the advantage of PC-score over LS-score in VS, we only discuss the version of LS-align using PC-score in the comparisons.

Table 3. VS results by Rigid-LS-align, LIGSIFT, Align-It™, and ROCS using single rigid-body conformers of molecules in the DUD-E. Shape sTC, chem sTC, and combo sTC are different modules used in LIGSIFT. Bold fonts highlight the highest value in each category.

Method	$EF^{1\%}$	$EF^{5\%}$	$EF^{10\%}$	$HR^{1\%}$	$HR^{5\%}$	$HR^{10\%}$
LIGSIFT (shape sTC)	8.6	3.6	2.6	13.8	17.9	26.3
LIGSIFT (chem sTC)	13.3	4.9	3.4	21.3	24.7	33.8
LIGSIFT (combo sTC)	11.9	4.5	3.1	19.6	22.8	31.5
Align-It™	14.2	5.6	3.7	23.4	28.8	37.4
ROCS	13.2	4.5	2.9	21.3	22.4	29.0
Rigid-LS-align	**20.1**	**6.9**	**4.3**	**33.0**	**34.8**	**43.5**

Comparing to LIGSIFT, Align-It™ and ROCS, Rigid-LS-align achieves a considerably higher *EF* score, indicating Rigid-LS-align can recognize more active molecules in the rigid-body based VS experiments. The average $EF^{1\%}$, $EF^{5\%}$, and $EF^{10\%}$ values of Rigid-LS-align are 20.1, 6.9, and 4.3, which are about 41.5%, 23.2%, and 16.2% higher than that of the second-best method Align-It™, respectively. Rigid-LS-align also outperforms LIGSIFT, Align-It™, and ROCS concerning $HR^{1\%}$, $HR^{5\%}$, and $HR^{10\%}$. The average $HR^{1\%}$, $HR^{5\%}$, and $HR^{10\%}$ of Rigid-LS-align are 33.0, 34.8, and 43.5, whereas the average $HR^{1\%}$, $HR^{5\%}$, and $HR^{10\%}$ of the second-best method (Align-It™), are 23.4, 28.8, and 37.4, respectively.

In addition to enrichment factor, Figure S7 presents a head-to-head comparison of Rigid-LS-align versus the three control methods on the AUC values. Out of the 102 DUD-E targets, there are 70, 65, and 86 cases where Rigid-LS-align has a higher AUC than LIGSIFT, Align-It™, and ROCS, respectively. To examine the significance of the differences between the methods, we list in Table S5 the *p*-values in the student's t-test using various assessment indexes (EF, HR and AUC), where the *p*-values between Rigid-LS-align and the three control methods are all below 10^{-3}, indicating that the difference is statistically significant.

In Table 4, we split the DUD-E datasets into four categories, i.e., kinases (26 targets), proteases (15 targets), nuclear receptors (11 targets), and GPCRs (5 targets). It is found that Rigid-LS-align has a slightly lower performance in proteases and GPCRs than kinases and nuclear receptors; but its enrichment score is higher than the control methods through all categories of proteins. Taking kinases as an example, the $EF^{1\%}$, $EF^{5\%}$, and $EF^{10\%}$ of Rigid-LS-align are 19.0, 6.5, and 4.2, which are about 21.0%, 12.1%, and 13.5% higher than Align-It™, 57.0%, 47.7% and 35.5% higher than LIGSIFT, and 46.2%, 62.5%, and 61.5% higher than ROCS, respectively.

To further examine the robustness of the programs on different ligands, in addition to the use of co-crystallized ligands, we performed a new experiment that uses each of the active ligands (~224 per protein target) as the seed ligand to rank other decoy compounds in the DUD-E datasets. Figure S8 shows a head-to-head comparison of the average $EF^{1\%}$ value for each target by Rigid-LS-align versus that by the three control methods. Out of the 102 protein targets, there are 96, 80, and 84 cases where Rigid-LS-align has a higher average $EF^{1\%}$ than LIGSIFT, Align-It™, and ROCS, respectively. The *p*-values of the difference between Rigid-LS-align and the three control methods in student *t*-test are all below 10^{-6}, indicating the difference is statistically significant. Figure S9 lists the average $EF^{1\%}$, $EF^{5\%}$, and $EF^{10\%}$ values over all the 102 targets achieved by Rigid-LS-align (22.4, 7.1, and 4.3), which are again higher than that by the control methods, i.e., 50.3%, 47.9%, and 34.4% higher than that of LIGSIFT, 22.4%, 22.4%, and 19.4% higher than that of Align-ItTM, and 25.8%, 34.0%, and 34.4% higher than that of ROCS, respectively.

Table 4. *EF* values by Rigid-LS-align, LIGSIFT, Align-It™, and ROCS using single rigid-body conformers on four protein categories. Bold fonts highlight the highest value in each category.

Subset (#proteins)	Method	$EF^{1\%}$	$EF^{5\%}$	$EF^{10\%}$
Kinases (26)	LIGSIFT (chem sTC)	12.1	4.4	3.1
	Align-It™	15.7	5.8	3.7
	ROCS	13.0	4.0	2.6
	Rigid-LS-align	**19.0**	**6.5**	**4.2**
Proteases (15)	LIGSIFT (chem sTC)	8.2	4.1	3.0
	Align-It™	5.5	3.4	2.5
	ROCS	6.3	2.6	2.0
	Rigid-LS-align	**15.4**	**6.3**	**4.3**
Nuclear receptors (11)	LIGSIFT (chem sTC)	16.9	6.1	4.1
	Align-It™	11.8	5.0	3.3
	ROCS	16.3	6.0	3.8
	Rigid-LS-align	**22.2**	**7.2**	**4.6**
GPCRs (5)	LIGSIFT (chem sTC)	7.3	3.7	2.6
	Align-It™	3.2	2.6	2.0
	ROCS	7.0	3.1	2.2
	Rigid-LS-align	**16.6**	**5.5**	**3.6**

3.5 Comparison of LS-align with other methods in VS experiment based on flexible alignment

Since the consideration of structural flexibility has shown positive impact on the VS performance, here we compare the Flexi-LS-align module with the LIGSIFT, Align-It™, and ROCS programs. Because these control methods do not have option for flexible alignment, we first generate a set of alterative conformers by the Flexi-LS-align and then apply the control programs on the conformers. Table 5 shows the VS results of the four programs on the same set of conformers from Flexi-LS-align.

Table 5. VS results by Flexi-LS-align, LIGSIFT, Align-It™, and ROCS using the same set of alternative conformers from Flexi-LS-align. Values in parentheses are *p*-values in student t-test relative to Flexi-LS-align.

Method	$EF^{1\%}$ (*p*-value)	$EF^{5\%}$ (*p*-value)	$EF^{10\%}$ (*p*-value)	AUC (*p*-value)
LIGSIFT (chem sTC)	13.0 ($<10^{-12}$)	4.8 ($<10^{-12}$)	3.1 ($<10^{-12}$)	0.70 ($<10^{-12}$)
Align-It™	13.2 ($<10^{-12}$)	4.9 ($<10^{-12}$)	3.2 ($<10^{-12}$)	0.66 ($<10^{-10}$)
ROCS	13.7 ($<10^{-10}$)	4.6 ($<10^{-13}$)	2.9 ($<10^{-16}$)	0.61 ($<10^{-19}$)

Flexi-LS-align	**22.0**	**7.2**	**4.5**	**0.75**

The average $EF^{1\%}$ by Flexi-LS-align is 22.0 on the 102 DUD-E proteins, which is higher than that of LIGSIFT (13.0), Align-It™ (13.2), and ROCS (13.7), all with a significant *p*-value ($<10^{-10}$). Among the 102 targets, a high enrichment score with $EF^{1\%} > 30$ was observed for Flexi-LS-align on 26 DUD-E targets, including ESR1, FAK1, GCR, MK01, SAHH, WEE1, KIF11, PGH2, PYRD, XIAP, HIVPR, TYSY, ACES, KITH, MMP13, PNPH, THB, ADA, COMT, DEF, FABP4, FPPS, HMDH, KPCB, MET, and PUR2; on the contrast, there are 9, 11, or 10 targets that are achieved by LIGSIFT, Align-It™, or ROCS at this level of performance (see Table S6 for detailed *EF* values by the programs).

It is of interest to note that the use of flexible conformers generated by Flexi-LS-align does not improve much the performance of control methods compared to using the rigid-body alignment. In fact, the average $EF^{1\%}$ values of LIGSIFT (13.0) and Align-It™ (13.2) using multiple Flexi-LS-align conformers are slightly lower than that using rigid-body conformers (13.3 and 14.2), although ROCS is slightly improved (13.2 vs 13.7). One reason for this difference might be that the multiple conformers generated by Flexi-LS-align, which involve the conformer selections specified by the Rigid-LS-align alignment that is specifically beneficial to LS-align search, might not satisfy the requirement of the conformation diversity of the control methods.

To examine this possibility, we further compare Rigid-LS-align with the control methods on a new set of diverse conformers created by the software OMEGA (OpenEye Inc) (Hawkins, et al., 2010) using the default settings from the co-crystallized seed ligands. Table 6 show the VS performances by the four programs based on 50 and 10 OMEGA-generated conformers, respectively. Indeed, the conformers from OMEGA improve the performance of LIGSIFT and Align-It™. They also increase the performance of Rigid-LS-align and ROCS, compared to that using a single conformer. But the increase on LS-align is not as high as that using the 10 flexible conformers created by Flexi-LS-align; this is consistent with the fact that the conformer generation procedure, which decides the rotation bonds and angles based on the first round of LS-align rigid-body superposition, has been specifically optimized for maximizing the Flexi-LS-align performance.

Table 6. VS performance comparisons of Rigid-LS-align, LIGSIFT, Align-It™, and ROCS using 50 and 10 OMEGA-generated conformers of database molecules on 102 DUD-E targets. Bold fonts highlight the highest value in each category.

Method	$EF^{1\%}$	$EF^{5\%}$	$EF^{10\%}$	$HR^{1\%}$	$HR^{5\%}$	$HR^{10\%}$
Using 50 OMEGA conformers						
LIGSIFT (chem sTC)	20.3	6.7	4.1	32.7	33.6	41.3
Align-It™	16.0	5.4	3.4	25.7	27.0	34.4
ROCS	13.9	4.7	2.9	22.4	23.4	29.5
Rigid-LS-align	**21.9**	**7.3**	**4.5**	**36.5**	**36.5**	**44.7**
Using 10 OMEGA conformers						
LIGSIFT (chem sTC)	18.3	6.1	3.9	29.4	30.6	38.8
Align-It™	15.4	5.4	3.5	24.6	27.2	34.6
ROCS	14.3	4.8	3.0	23.0	23.9	29.9
Rigid-LS-align	**21.8**	**7.2**	**4.5**	**36.1**	**36.2**	**44.4**

Nevertheless, the performance of Rigid-LS-align on the OMEGA conformers is still relatively higher than that of the control methods on the same conformer set. Using 50 OMEGA-generated conformers, for example, the $EF^{1\%}$ and $HR^{1\%}$ of Rigid-LS-align are 21.9 and 36.5, which are about 7.9% and 11.6% higher than that of LIGSIFT, 36.9% and 42.0% higher than that of Align-It™, and 57.6% and 62.9% higher than that of ROCS, respectively (see detail in Table S7). Using 10 OMEGA-generated conformers, the $EF^{1\%}$ and $HR^{1\%}$ of Rigid-LS-align are 21.8 and 36.1, which are 19.1% and 22.8% higher than that of LIGSIFT, 41.6% and 46.7% higher than that of Align-It™, and 52.4% and 57.0% higher than that of ROCS, respectively.

In Figure S10, we further examine the AUC histogram of the four programs using the 50 OMEGA conformers, which shows that the distribution of AUC score by Rigid-LS-align is generally shifted to the region of larger AUC values. The average AUC for Rigid-LS-align, LIGSIFT, Align-It™, and ROCS are 0.740, 0.714, 0.664, and 0.601, respectively, which demonstrate again the advantage of the LS-align program in terms of both conformational search engine and scoring function.

3.6 Comparison of 3D- and 2D-based methods on VS

Since 3D based structure alignments are generally more time-consuming than 2D based approaches, to examine the justification of the time investment we present in Table S8 a brief comparison of the VS results by the four 3D based methods, with the most widely used 2D fingerprint method, named 'TC-FP', in which the similarity score of two ligand molecules is measured by the Tanimoto coefficient of their fingerprints (Willett, 2006). Here, the fingerprint vector of each molecule is generated by OpenBabel (O'Boyle, et al., 2011); and 50 conformers by OMEGA are used for the 3D methods.

Interestingly, despite the fast speed and simplicity of implementation, the performance of TC-FP is only slightly worse than (or largely comparable to) the best 3D structure alignment result from Rigid-LS-align, and outperforms all of other 3D-ligand comparison methods. Nevertheless, we believe that there are several reasons for which the 3D methods are still valuable. First, since they are based on different principles, the VS results by the 2D and 3D approaches are highly complementary. As shown in Figure 7, a simple linear combination with TC-FP can have the majority of the 3D methods significantly outperform the 2D method in VS. For example, with a 50/50 combination, TC-FP+Rigid-LS-align can generate an average $EF^{1\%}$ value of 26.0, which is significantly higher than either of TC-FP (20.5) or Rigid-LS-align (21.9), with a *p*-value $=4.2\times10^{-8}$ and 2.0×10^{-8}, respectively.

Second, the LS-align has been designed mainly for comparing the structures of ligand molecules, where VS is only one of its applications. For example, the 3D structural alignments could help visualize and analyze the physical and function features required for the biological interactions, which is essential for scaffold hopping (Hu, et al., 2017; Quintus, et al., 2009). In Figure S11, we present six examples of scaffold hopping, in which the analogous compounds can be easily identified by Rigid-LS-align but cannot be found using 2D-based approaches.

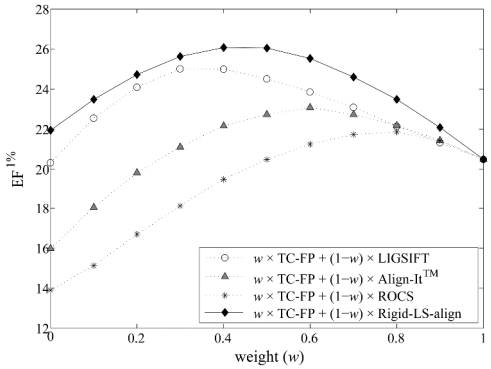

Fig. 7. The variation curves of the average $EF^{1\%}$ value versus weight to linearly combine the 2D and 3D scores. The data are generated by using 50 OMEGA-generated conformers on the DUD-E dataset. The best combinations that result in the highest $EF^{1\%}$ value (26.0, 25.0, 23.1, 21.9) have the weight of 0.5, 0.3, 0.6, and 0.8 for Rigid-LS-align, LIGSIFT, Align-ItTM, and ROCS, respectively (see Table S8).

4 Discussions

We have developed a novel algorithm, LS-align, for atomic-level, sequence-order independent comparison of ligand molecule structures. The large-scale experimental tests on the DUD-E database show that LS-align outperforms many of the state of the art 3D-based ligand alignment methods in virtual screening by prioritizing ligand compounds with a higher enrichment score and hitting rate.

Detail data analyses show that the superior performance of LS-align in VS mainly stems from the use of the sensitive PC-score which combines multiple information of structural and physicochemical features with optimized weighting schemes. With the same right-body alignment scheme, the combination of multiple features can increase the enrichment score by 67.5% compared to the purely structure-based scoring function. The second advantage of the LS-align is the introduces of structure variations in the alignment search, which further improves the screening performance by nearly 10% in the benchmark test. Although this improvement is not as significant as the score function changes, it provides the potential to count for the binding-induced conformational changes in the ligand molecules and therefore break through the barrier of the rigid-body alignments in prioritizing such difficult cases in VS. Finally, the EGS-based heuristic iterative search algorithm is helpful for fast and accurately identifying the reasonable alignments from a huge number of possible alternatives. Currently, for a pair of ligand molecules each with 30 heavy atoms, it takes about 30 ms and 0.5 s for obtaining rigid-body and flexible alignments, respectively, on a 3.5 GHz Intel-Xeon processor.

Compared to the current 3D ligand comparison methods in literature, which are mainly based on shape overlap of molecule structures, another advantage of the proposed LS-align is its ability in generating atom-specific alignments, which provides the convenience for visualization and analyses of the ligand structural comparisons. This is also critical for the application to ligand-protein docking and scaffold hopping as illustrated in Figure S11.

Nevertheless, it is noted that the LS-align method has still room for further improvement. For example, the current LS-align program only considers single-bond rotations with discrete rotation angles for flexible conformer generations. In our virtual screening experiment, we have observed several cases, in which the best candidates with a similar chemical formula was missed by LS-align due to the conformational changes that are not recovered by the Flexi-LS-align. Detailed analyses showed that these missed cases are mainly due to the limit of single-bond rotation and the big size of discrete rotations. Fig. S12 shows two examples in which better alignments could be achieved when considering additional degree of ring flexibility and/or with a finer rotation bond-angle. Including multiple chemical bond rotations with finer variation steps in the flexible alignment process should apparently help LS-align to achieve more accurate structure alignment and enhance the usefulness in the VS experiment; but merely reducing the step size of rotations or adding extra degrees of freedom in ligand flexibility will exponentially increase the time cost of the alignment search process. Nevertheless, given the importance of the virtual screen experiments and the role that the ligand flexibility plays in VS, such effort is worthy of exploring in future studies.

Funding

This work was supported in part by the National Natural Science Foundation of China (No. 61373062, 31628003), the Natural Science Foundation of Jiangsu (No. BK20141403), the Fundamental Research Funds for the Central Universities (No. 30916011327), China Scholarship Council (No. 201606840087), National Institute of General Medical Sciences (GM083107 and GM116960), and National Science Foundation (DBI1564756).

Conflict of Interest: none declared.

References

Cheeseright, T.J., *et al.* FieldScreen: virtual screening using molecular fields. Application to the DUD data set. *Journal of chemical information and modeling* 2008;48(11):2108-2117.

Eckert, H. and Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug discovery today* 2007;12(5):225-233.

Embrechts, P., Kluppelberg, C. and Mikosch, T. Modelling extremal events for insurance and finance. Berlin: Spring Verlag 1997.

Forli, S., *et al.* Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nature protocols* 2016;11(5):905-919.

Grant, J., Gallardo, M. and Pickup, B. A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* 1996;17(14):1653-1666.

Hawkins, P.C., *et al.* Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *Journal of chemical information and modeling* 2010;50(4):572-584.

Holliday, J.D., *et al.* Analysis and display of the size dependence of chemical similarity coefficients. *Journal of chemical information and computer sciences* 2003;43(3):819-828.

Hu, Y., Stumpfe, D. and Bajorath, J. Recent advances in scaffold hopping. *J. Med. Chem* 2017;60(4):1238-1246.

Jennings, A. and Tennant, M. Selection of molecules based on shape and electrostatic similarity: proof of concept of "electroforms". *Journal of chemical information and modeling* 2007;47(5):1829-1838.

Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* 1976;32(5):922-923.

Kuhn, H.W. The Hungarian method for the assignment problem. *Naval Research Logistics (NRL)* 1955;2(1‐2):83-97.

Munkres, J. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics* 1957;5(1):32-38.

Mysinger, M.M., *et al.* Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry* 2012;55(14):6582-6594.

O'Boyle, N.M., *et al.* Open Babel: An open chemical toolbox. *Journal of cheminformatics* 2011;3(1):33.

Patel, H., Brinkjost, T. and Koch, O. PyGOLD: a python based API for docking based virtual screening workflow generation. *Bioinformatics* 2017.

Quintus, F., *et al.* Ligand scaffold hopping combining 3D maximal substructure search and molecular similarity. *BMC Bioinformatics* 2009;10(1):245.

Roy, A. and Skolnick, J. LIGSIFT: an open-source tool for ligand structural alignment and virtual screening. *Bioinformatics* 2015;31(4):539-544.

Shang, J., *et al.* HybridSim-VS: a web server for large-scale ligand-based virtual screening using hybrid similarity recognition techniques. *Bioinformatics* 2017.

Sperandio, O., *et al.* MED-SuMoLig: a new ligand-based screening tool for efficient scaffold hopping. *Journal of chemical information and modeling* 2007;47(3):1097-1110.

Taminau, J., Thijs, G. and De Winter, H. Pharao: pharmacophore alignment and optimization. *Journal of Molecular Graphics and Modelling* 2008;27(2):161-169.

Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* 2006;11(23-24):1046-1053.

Zhang, Y. Protein structure prediction: when is it useful? *Curr Opin Struct Biol* 2009;19(2):145-155.

Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004;57:702-710.

Zhang, Y. and Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* 2005;33(7):2302-2309.