JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

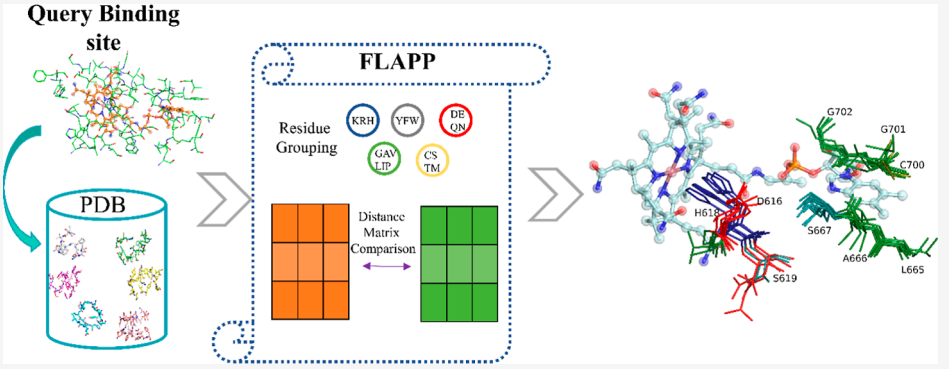
pubs.acs.org/jcim Article

Fast Local Alignment of Protein Pockets (FLAPP): A System-Compiled Program for Large-Scale Binding Site Alignment

Santhosh Sankar, Naren Chandran Sakthivel, and Nagasuma Chandra*

Cite This: J. Chem. Inf. Model. 2022, 62, 4810−4819 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



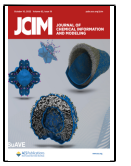
ABSTRACT: Protein function is a direct consequence of its sequence, structure, and the arrangement at the binding site. Bioinformatics using sequence analysis is typically used to gain a first insight into protein function. Protein structures, on the other hand, provide a higher resolution platform into understanding functions. As the protein structural information is increasingly becoming available through experimental structure determination and through advances in computational methods for structure prediction, the opportunity to utilize these data is also increasing. Structural analysis of small molecule ligand binding sites in particular provides a direct and more accurate window to infer protein function. However, it remains a poorly utilized resource due to the huge computational cost of existing methods that make large-scale structural comparisons of binding sites prohibitive. Here, we present an algorithm called FLAPP that produces very rapid atomic level alignments. By combining clique matching in graphs and the power of modern CPU architectures, FLAPP aligns a typical pair of binding sites at ∼12.5 ms using a single CPU core, ∼1 ms using 12 cores on a standard desktop machine, and performs a PDB-wide scan in 1−2 min. We perform rigorous validation of the algorithm at multiple levels of complexity and show that FLAPP provides accurate alignments. We also present a case study involving vitamin B12 binding sites to showcase the usefulness of FLAPP for performing an exhaustive alignment-based PDB-wide scan. We expect that this tool will be invaluable to the scientific community to quickly align millions of site pairs on a normal desktop machine to gain insights into protein function and drug discovery for drug target and off-target identification and polypharmacology.

1. INTRODUCTION

Proteins are the primary architects of biological processes, and the study of their three-dimensional structures provides a means to understand them at the highest resolution. Protein function has traditionally been elucidated by biochemical assays, but the genomic era has provided us with protein sequences at an unprecedented rate, leading to a large gap between protein identification and understanding their function. Computational approaches allow us to bridge the gap. Typically, such computational approaches utilize various measures of similarity inferred to transfer knowledge from annotated proteins.[1] Sequence-based comparison methods such as BLAST and HMMER offer a great heuristic in rapidly identifying sets of proteins that perform similar functions by mining curated databases such as PFAM and transferring annotations from known proteins to query proteins.[2−4] On the other hand, structural alignment evaluates similarity by considering their three-dimensional shapes. While alignments that use fold information are more effective than traditional sequence alignments, those that assess similarity in the binding sites provide a higher resolution and more direct means to function assignment.

Many algorithms have been reported for comparing binding sites. These broadly fall into: (i) alignment-free and (ii) alignment-based. Alignment-free comparison methods use physicochemical descriptors (e.g., hydrogen bonding patterns)

Received: July 28, 2022
Published: September 19, 2022



https://doi.org/10.1021/acs.jcim.2c00967
J. Chem. Inf. Model. 2022, 62, 4810−4819

Graphical Overview of FLAPP

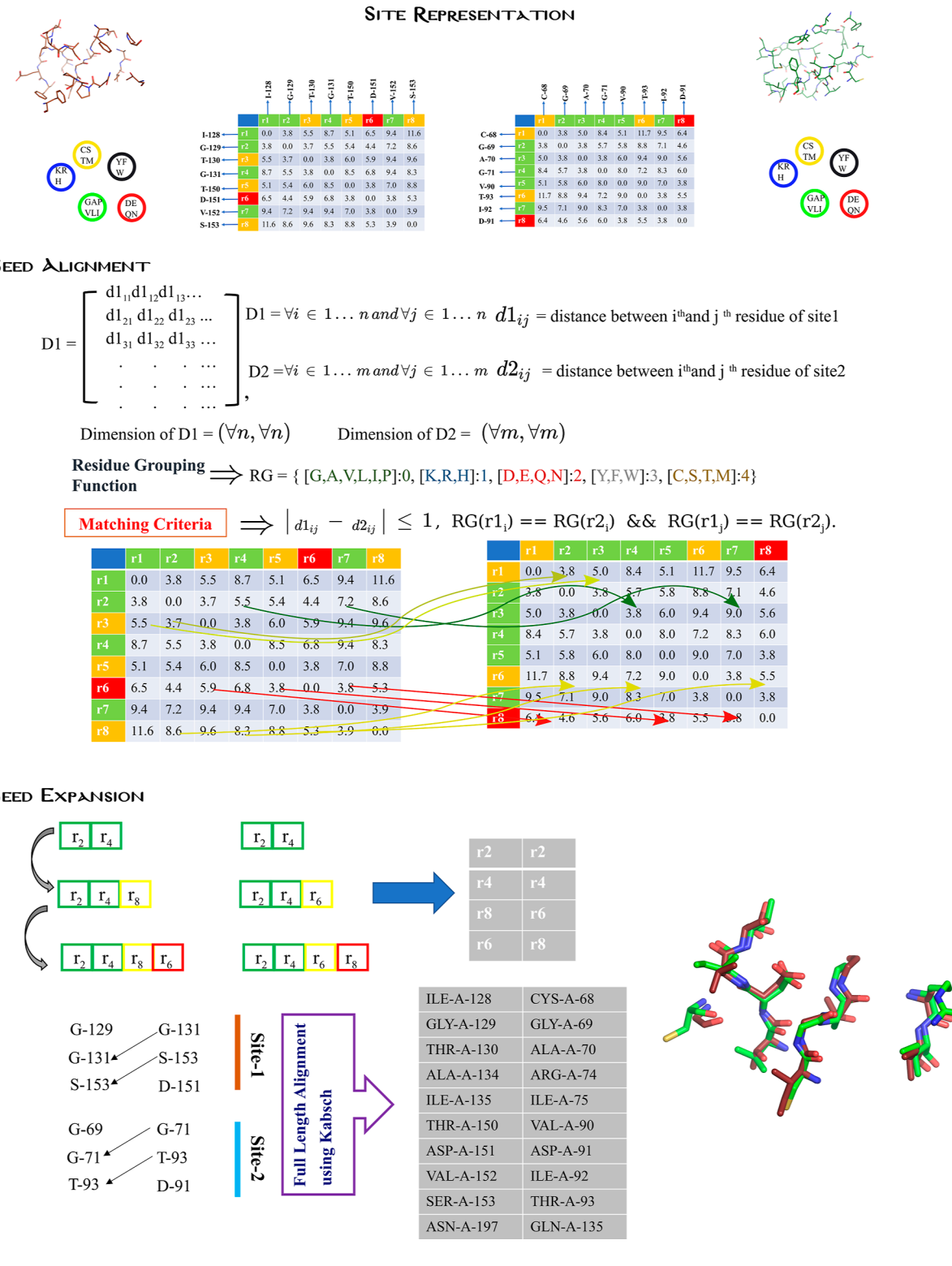


Figure 1. Graphical overview of the FLAPP algorithm. The FLAPP algorithm consists of three main modules: (i) binding site representation, (ii) seed selection, and (iii) seed expansion. FLAPP first represents binding sites as Cα distance matrices. Using physicochemical grouping and a distance metric from eq 1, it first selects residue pairs that are similar from both binding sites as seed primers. These primers are then extended with residues as long as the similarity criteria (i.e., eqs 1 and 2) are maintained. The seeds identified after the seed expansion step are used to get an optimal site rotation matrix via least square superposition using the Kabsch algorithm.

https://doi.org/10.1021/acs.jcim.2c00967
J. Chem. Inf. Model. 2022, 62, 4810−4819

for assessing similarity between binding sites,[5−7] whereas alignment-based methods aim to generate 3D alignments that give residue−residue correspondences between the sites (e.g., PocketAlign, G-LoSA, SiteMotif, and SiteEngine[8−11]). Due to the sequentially discontinuous nature of the binding sites, popular sequence or fold alignment algorithms such as dynamic programming are not readily amenable for aligning binding sites, and hence, alignment-free methods are usually preferred for this purpose. Dynamic programming is an algorithmic technique that breaks down a problem into simpler subproblems, solutions for which are computed and stored, in order to reduce time complexity and thereby achieve significant optimization. In bioinformatics, Needleman−Wunsch and Smith−Waterman are the two widely adopted dynamic-programming algorithms that efficiently solve the complexity involved in the sequence alignment. Despite the speed of alignment-free methods (e.g., PocketMatch, RAPMAD), alignment-based methods (e.g., PocketAlign) are ideal because they provide residue-level alignments. However, their runtime performance remains a bottleneck for their application in large-scale site comparison exercises. For example, SiteEngine (2005) and PocketAlign (2011) take ∼2 min to align a pair of sites.[8,11] Alternatives such as G-LoSA and SiteMotif are able to bridge this gap by doing one alignment per second or less.[9,10] Nevertheless, they become infeasible when the number of comparisons is in the order of a million or more. This highlights the need for a fast, accurate, and scalable site alignment algorithm.

In this work, we present a new algorithm, Fast Local Alignment of Protein Pockets (FLAPP), for rapid and accurate alignment of binding sites, facilitating proteome-wide or PDB-wide scans. The algorithmic design leverages the modern CPU architecture and efficient data structures to generate alignments comparable to state-of-the-art methods. We demonstrate the superiority of FLAPP by benchmarking against the best available methods currently. Lastly, we showcase the efficiency of FLAPP by deriving motifs in vitamin B12 ligand binding sites and exhaustively scanning the motifs against all binding pockets in PDB.

2. METHODS

2.1. Algorithmic Implementation. A "binding site" is defined as the set of residues in a protein that forms a cavity or pocket capable of housing a small molecule ligand enabling its binding to the protein. If the binding site is present in PDB as complexed to the ligand in question, we refer to it as a "binding site", where as if it is not present as a liganded complex with the given ligand but predicted to do so, we refer to it as a "pocket", consistent with the wide usage in literature. For proteins that are complexed with ligands, the binding site represents those residues that are present within 4.5 Å of any ligand atom. In cases where the structure of protein is solved in an apo form (i.e., without ligand), then the pockets can be identified using established pocket prediction algorithms such as PocketDepth, FPocket, and SiteHound.[12−14] Both "pocket" and the "binding site" represent the similar meaning in the context of this study.

The FLAPP implementation (Figure 1) constitutes three modules: (i) site representation, (ii) seed selection, and (iii) optimal alignment building.

2.1.1. Site Representation. The site representation module constructs a distance matrix for the binding site. For a given binding site of N residues, FLAPP first extracts Cα positions and constructs a distance matrix (D) of NxN dimensions that contains the euclidean distances for all N^2 residue-pairs. Next, it groups all 20 amino acids into five standard groups based on the physicochemical properties. Group-0: (G,A,V,L,I,P); group-1: (K,R,H); group-2: (D,E,Q,N); group-3: (Y,F,W); and group-4: (C,S,T,M). The matrix D can also be thought of as a graph G(V,E), V representing the set of Cα atoms in each of the N residues and E representing all-pair distances among them.

2.1.2. Seed Primer Selection. The second module exhaustively samples the sites to identify all possible "seeds", which are progressively grown to find matching cliques. A clique represents a set of vertices in a graph that form a completely connected subgraph. These are then screened to find the optimal alignment in the next module. From the adjacency matrices $D1$ and $D2$, it selects all distances that are similar based on two criteria: (i) the magnitude of the difference between two distances is less than 1 Å (eq 1), (ii) the residues being compared belong to the same residue grouping (eq 2). Distances that satisfy these criteria are chosen as "seed primers". The seed primers represent the common elements in a given pair of sites and by definition would form cliques.

$$\text{RMSD} = \sqrt{(d_1 - d_2)^2} \Rightarrow \{d_1 \forall D_1, d_2 \forall D_2 \leq 1\ \text{Å}\}, \quad \text{where } d_1 \in D1 \text{ and } d_2 \in D2 \tag{1}$$

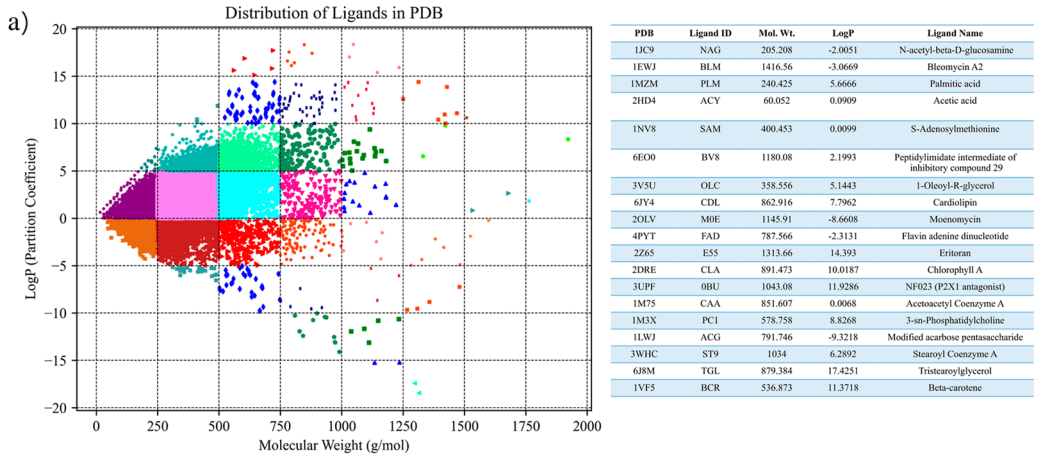
$$\text{RG}: R_1 \rightarrow R_2 \text{ and } (R_1, R_2) \in R. \quad R\{[\text{G, A, V, L, I, P}]: 0, [\text{K, R, H}]: 1, [\text{D, E, Q, N}]: 2, [\text{Y, F, W}]: 3, [\text{C, S, T, M}]: 4\} \tag{2}$$

2.1.3. Seed Growth. At this point, each seed primer contains two residues from each site based on eqs 1 and 2. To add the next residue, we incorporate a strategy where we seek residues starting from the residue closest to the seed primer. To accomplish this, we construct a sorted distance list for the trailing seed residue (D^{sort}). Each row of $D1^{\text{sort}}$ contains distances between any ith residue of pocket-1 with all other residues in pocket-1. The program scans this list in the ascending order to find matching distances in $D2^{\text{sort}}$. At this step, the seeds are progressively expanded as follows: an empty ArrayList is created to store the seeds and subsequent residues pairs are then added to the "ArrayList" such that (a) the distances match according to our criteria (eqs 1 and 2), (b) the newly added residues in both the structures have equivalent groups and (c) the residues in the ArrayLists from both the structures form matching cliques of size >3. Residue pairs not associated with any clique are stored separately in a look-up table called the "HashTable", which is consulted at each iteration to avoid repeated traversals of the same paths. The loop is iterated recursively until all residues are scanned. This step yields "expanded seeds", which provide candidate alignments to be screened in the next step.

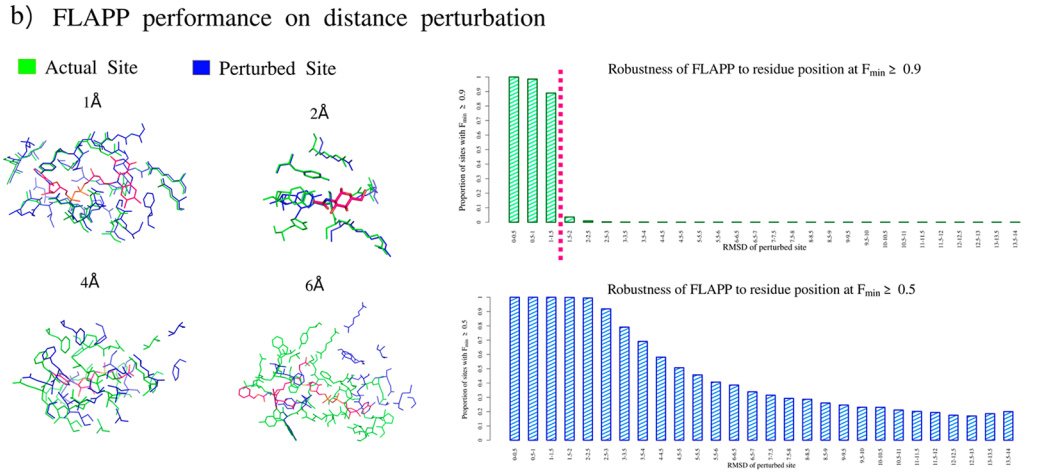
2.2. Least Square Superposition Using Kabsch Algorithm. The candidate alignments are evaluated using the Kabsch least square superposition algorithm, from which the largest candidate (max n) with the least RMSD (<1 Å) is taken as the optimal alignment of the two pockets. The Kabsch alignment is performed iteratively over all possible local alignments and only reports the maximum alignment.[15]

2.3. Alignment Scores. The number of possible residues that can align between two binding sites is bounded by the size of the smaller of the two sites being compared. In order to make comparison of binding site pairs more uniform, we define two

https://doi.org/10.1021/acs.jcim.2c00967
J. Chem. Inf. Model. 2022, 62, 4810−4819



PDB	Ligand ID	Mol. Wt.	LogP	Ligand Name
1JC9	NAG	205.208	-2.0051	N-acetyl-beta-D-glucosamine
1EWJ	BLM	1416.56	-3.0669	Bleomycin A2
1MZM	PLM	240.425	5.6666	Palmitic acid
2HD4	ACY	60.052	0.0909	Acetic acid
1NV8	SAM	400.453	0.0099	S-Adenosylmethionine
6EO0	BV8	1180.08	2.1993	Peptidylimidate intermediate of inhibitory compound 29
3V5U	OLC	358.556	5.1443	1-Oleoyl-R-glycerol
6JY4	CDL	862.916	7.7962	Cardiolipin
2OLV	M0E	1145.91	-8.6608	Moenomycin
4PYT	FAD	787.566	-2.3131	Flavin adenine dinucleotide
2Z65	E55	1313.66	14.393	Eritoran
2DRE	CLA	891.473	10.0187	Chlorophyll A
3UPF	0BU	1043.08	11.9286	NF023 (P2X1 antagonist)
1M75	CAA	851.607	0.0068	Acetoacetyl Coenzyme A
1M3X	PC1	578.758	8.8268	3-sn-Phosphatidylcholine
1LWJ	ACG	791.746	-9.3218	Modified acarbose pentasaccharide
3WHC	ST9	1034	6.2892	Stearoyl Coenzyme A
6J8M	TGL	879.384	17.4251	Tristearoylglycerol
1VF5	BCR	536.873	11.3718	Beta-carotene



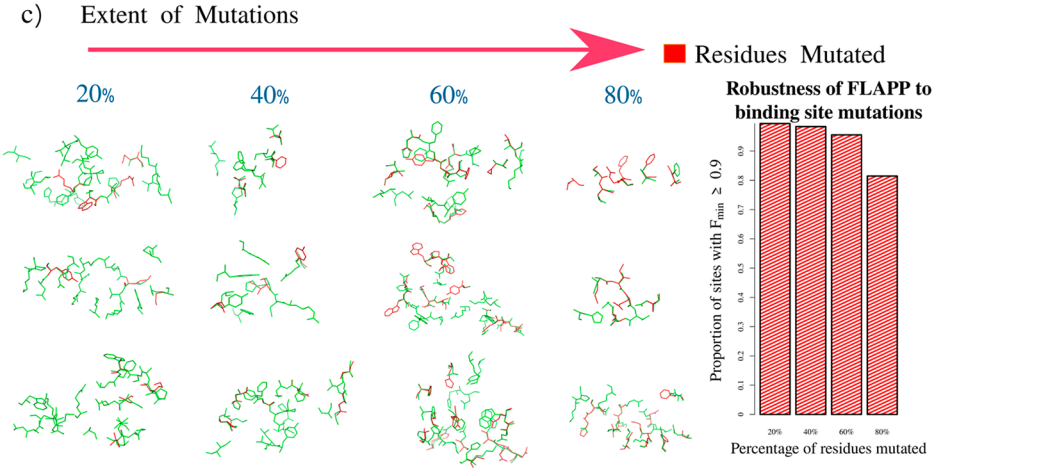


Figure 2. Assessing the sensitivity of FLAPP to different perturbations. (a) Grouping of ligands in PDB based on molecular weight (g/mol) and partition coefficient (Log *P*). The plot is partitioned into grids of 250 g/mol on the *x*-axis and 5 units of Log *P* on the *y*-axis. 19 diverse ligands were chosen and one binding site for each was chosen randomly. (b) The residues in the 19 binding sites were randomly perturbed so as to generate 3000 synthetic sites each, with RMSDs ranging from 0 to 14 Å. The sensitivity of FLAPP in aligning residues in the original site with the perturbed sites is explored at two F_{min} scores (0.5 and 0.9). (c) Similarly, synthetic binding sites that had a proportion of their residues mutated were generated for each of the 19 binding sites.

scores: F_{min} and F_{max}. F_{min} is defined as the ratio of the number of aligned residues to the number of residues in the smaller binding site in the pair, while F_{max} is defined as the ratio of the number of aligned residues to the number of residues in the larger binding site in the pair. Here, smaller and larger binding sites are defined with respect to the number of residues.

$$F_{min} = \frac{\#\text{ aligned residues}}{\#\text{ residues in smaller site}}$$

$$F_{max} = \frac{\#\text{ aligned residues}}{\#\text{ residues in larger site}}$$

https://doi.org/10.1021/acs.jcim.2c00967
J. Chem. Inf. Model. 2022, 62, 4810−4819

F_{min} and F_{max} are always expected to range between 0 and 1, both inclusive.

2.4. Code Optimization. The algorithm was implemented using Python (Anaconda distribution 3.9). Many arithmetic operations such as calculating RMSD, singular value decomposition (SVD), for Kabsch are implemented using NumPy functions. For numerical calculations, Intel's SVML (short vector math library) library has been utilized. The entire program was wrapped with Numba Just-In-Time compiler, which allowed the Python code to be compiled into machine code specific to each system architecture (LLVM compiler).[16] Moreover, the code has been written in a format that facilitates Numba to compile into the Single Instruction Multiple Data (SIMD) form. As a consequence of using Numba, the code had to be written such that the memory was pre-allocated for a few operations. For example, the "append" function of python which adds new elements to a list is not supported by Numba. This is because the function "append" creates a dynamic array whose size is not fixed and can be modified during execution. Incorporating such functions is not supported in compilation mode as we are required to provide the array size in advance. To circumvent this, a matrix of dimension 1000 was created with zero padding. The index of the array is then accessed and replaced sequentially via an incremental counter. The HashTable variable, which is used to check if the cliques have been traversed before, implements this logic.

The operations that further helped in optimizing the algorithm are as follows: (i) the use of sorting at each seed selection which allows for considering local alignments as nearby residues are aligned first, (ii) creation of a LookUp table (size of $n \times 100$, where n represents the number of residues of site-1 and 100 corresponds to the index of the matched residues of site-2; at the start of execution, the LookUp matrix is initialized with zeros. For every match detected, a counter variable was incremented to one which will replace zeros with the index pointing to the residue number of site-2) to keep track of paths that have already been explored and hence avoid futile traversals, (iii) use of the Intel SVML library for computing the dot products in the construction of rotation matrices for the calculation of RMSD.

3. RESULTS

We present a fast and accurate algorithm FLAPP for aligning binding sites in protein structures. An alignment of a pocket-pair (25 residues each) takes 12 ms (~1/80 s) on a single desktop processor (Intel i5-8400 2.80 GHz). The speed of obtaining alignments at the atomic level enables accurate searching at a PDB-scale.

3.1. Evaluating the Accuracy of FLAPP. *3.1.1. Robustness of FLAPP to Minor Perturbations in the Binding Site.* To evaluate the robustness of FLAPP to geometric and compositional changes in binding sites, we systematically tested the (i) sensitivity of FLAPP to perturbations in binding site residue positions and (ii) sensitivity of FLAPP to mutations in binding site residues. In order to ensure that our analysis is not biased by the choice of ligand, we systematically selected a diverse set of sites (corresponding to 19 ligands) by constructing a 2D matrix based on molecular weights and partition coefficients of all ligands in PDB and selecting representatives from different regions (Figure 2a).

We first tested FLAPP's tolerance to perturbations by synthetically generating 3000 new binding sites from each of the 19 diverse sites by randomly perturbing their residue positions. The perturbations were done such that the newly generated binding sites had a bounded RMSD ranging between 0 and 14 Å with respect to the original site. Site alignments were then evaluated by FLAPP for each perturbation. FLAPP aligns perturbed sites with their parent sites with perfect F_{min} and F_{max} scores up to an RMSD of 1.5 Å in all cases (Figure 2b and Supporting Information, Figure S2) indicating that it is robust to minor perturbations in residue positions. Our analysis mimics perturbations seen in real life scenarios that can occur due to minor crystallographic variations or conformational changes due to ligand binding or even positional flexibility in NMR data. As the perturbations cross 6 Å, the scores drop below 0.5. This is expected as the RMSD increases, residue-to-residue alignments become weaker.

Next, we tested the sensitivity of FLAPP to change in binding site residue types, by systematically generating mutations in the site. For each of the 19 binding sites, we generated 1000 new sites for each site by randomly mutating some proportion of residues to other residues. FLAPP successfully aligns sites even after mutating up to 60% of their residues (Figure 2c). These perturbations are designed to mimic comparison of sites in homologous proteins and situations such as polymorphisms and disease associated mutations, which are typically seen only in a small portion of binding sites. The fact that our algorithm is robust upto 60% of changes in the site implies that residue grouping does not adversely affect its performance in achieving optimal alignments. The use of the Kabsch algorithm after the selection of the seeds ensures that residue−residue correspondences missed due to grouping are salvaged at the final alignment step (Supporting Information: PerturbationAnalysis-Supplementary.xlsx).

3.1.2. Evaluating FLAPP's Accuracy Across Levels of Increasing Complexity of Comparison. In order to systematically evaluate the alignment accuracy, FLAPP was tested on diverse pairs of binding sites. Here, we grouped binding sites into three levels of complexity (Easy, Medium, and Hard targets) based on sequence and structure similarities. (A) The easy targets include 7563 protein pairs from pdb_95 that share high sequence identity (>95%), belong to the same SCOP family and bind to the same ligand.[17] As the proteins are nearly identical, their binding sites are expected to be the same. FLAPP successfully showed identical alignments in 98% ($F_{min} \geq 0.8$ and alignment length >10) (Supporting Information, Figure S1a and Supporting Information: Level-Supplementary.xlsx). (B) The medium complexity targets were built using 729 protein pairs which do not have any similarity in sequence (identity <30%), but exhibit structural similarity (belong to the same SCOP superfamily).[18] The idea is that structurally similar and sequentially dissimilar proteins binding to the same ligand tend to share commonality in their binding sites. Out of 729, 416 pairs consist of binding sites that bind to the same ligand, which we consider as true positives. The remaining 313 are binding sites recognizing different ligands (Tanimoto coefficient <0.5) and are considered to be negative controls. FLAPP achieves an ROC of 0.93 in this data set, indicating that it can successfully handle binding sites from structurally similar, but sequentially dissimilar proteins (Supporting Information, Figure S1b and Supporting Information: Level-Supplementary.xlsx). (C) The hard target data set contains protein pairs that are diverged both at the sequence and the structure but recognize common ligands. Because it is the same ligand that is getting recognized, the proteins could possess some commonality in their site. We took such proteins from previously published literature where

https://doi.org/10.1021/acs.jcim.2c00967
J. Chem. Inf. Model. 2022, 62, 4810−4819

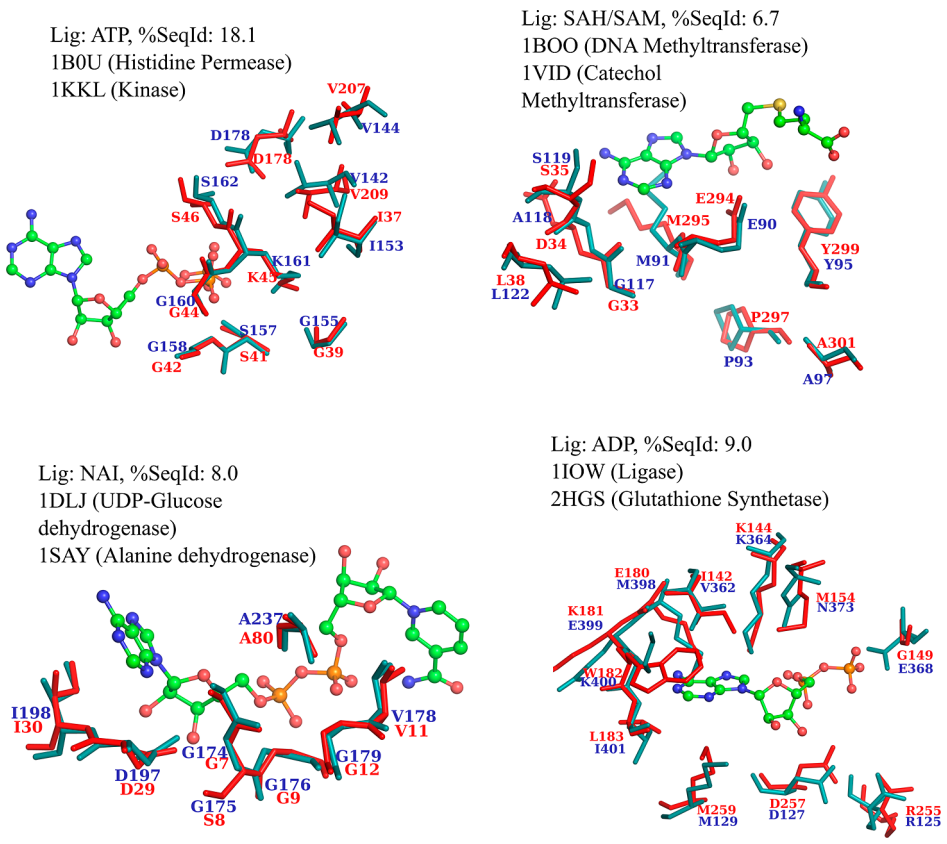


Figure 3. Alignments produced by FLAPP in the hard validation set. Each binding site pair has less than 20% sequence identity and proteins belong to different folds. All four pairs have sequence order reversal in the sequence. FLAPP successfully identifies similarities within each pair and outputs accurate residue−residue matches.

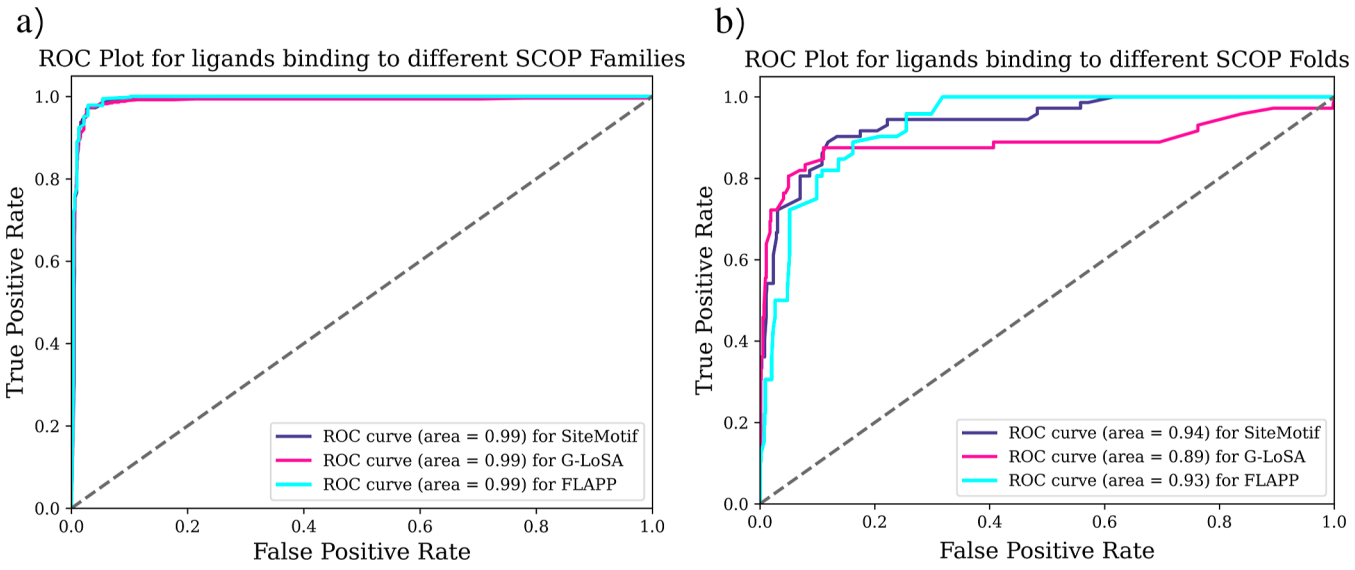


Figure 4. Alignment accuracy of FLAPP over known site alignment tools (SiteMotif and G-LoSA). Two sets of binding site pairs are analyzed, (a) 2421 pairs, of which 514 are true positives (similar sites) and 1907 are true negatives. All three methods are seen to fare well; (b) binding sites that exhibit only partial similarity between them capturing the length of alignment of each method, 5383 pairs are considered of which 72 are true positives and 5311 true negatives.

the authors explored a case of convergent evolution and reported four pairs of proteins that are structurally distinct but possess good commonality at their binding sites.[19] The mentioned pairs also intriguingly display sequence order reversal at the binding site which conventional sequence and structural alignment approaches fail to align. FLAPP is able to successfully align all these pairs. The residue−residue correspondence reported by our method is the same as that described by the authors. The optimal alignments reported by FLAPP for these pairs are shown in Figure 3. This in fact illustrates the importance of site-based alignment over traditional sequence or structural comparison. In addition to these targets at three

https://doi.org/10.1021/acs.jcim.2c00967
J. Chem. Inf. Model. 2022, 62, 4810−4819

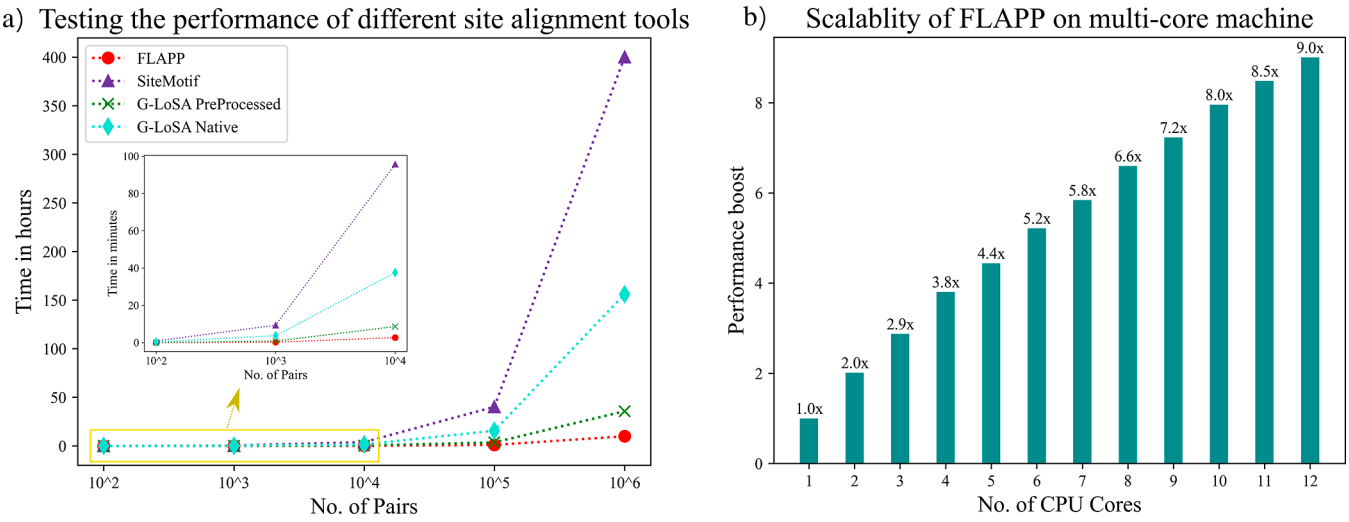


Figure 5. Execution time of FLAPP over known site alignment. (a) Performance comparison of FLAPP, SiteMotif, and G-LoSA on different data set sizes. The run times of FLAPP, G-LoSA (preprocessed and native), and SiteMotif were compared on sets of randomly selected binding site pairs. Sets of different sizes varying from 100 pairs to 0.1 million pairs were used to test the scalability of the method. FLAPP was able to align 1,000,000 pairs in 1.9 h (~6 ms per alignment). Our method is able to achieve at least 4X performance compared to the closest competitor G-LoSA PreProcessed and 20× faster than the actual G-LoSA program while maintaining accuracy on par with the SiteMotif. (b) Scalability of FLAPP to multiple CPU cores: the multi-processing enabled version of FLAPP was run on many cores and the performance boost was obtained. FLAPP efficiently utilizes all available CPU cores in order to scale further. At 12 cores, FLAPP aligns 1080 pairs per second (~1 ms per pair).

levels, we also constructed a data set of 50,000 binding site pairs that are dissimilar (sequence identity <30%, different SCOP folds) and bind to different ligands (Tanimoto Coefficient <0.5). These serve as negative controls for evaluating specificity. FLAPP did not yield any significant alignment among these pairs, indicating high specificity (Supporting Information: Level-Supplementary.xlsx). Put together, our results show that FLAPP has high sensitivity and specificity.

3.1.3. Benchmarking of Accuracy and Sensitivity of FLAPP against Other Methods. Next, we benchmark the accuracy of FLAPP against two site alignment algorithms, G-LoSA and SiteMotif,[9,10] that are currently the best alignment-based algorithms. Other methods that produce optimal alignments, for example, PocketAlign and SiteEngine, are not included as they need around 5 min per comparison. G-LoSA and SiteMotif on average take less than a second. To enable unbiased, systematic analysis, we assess the alignment accuracy of FLAPP, SiteMotif, and G-LoSA on two data sets of non-redundant binding sites. Data set-1 comprises a set of 514 pairs of proteins that bind to the same ligand and belong to the same SCOP superfamily but different SCOP families. This is essentially an obvious case because the sequence and the structures share a high degree of similarity at the SCOP superfamily level. As expected, all three methods fare very well on this data set (Figure 4a). Data set-2 comprises a pair of proteins that binds to similar ligands but belong to different SCOP folds. This data set can be considered as inherently difficult as the proteins share no similarity in the sequence and the structure. We compute the length of alignment reported by each method as our metric of accuracy. SiteMotif shows the best performance in Data set-2, followed closely by FLAPP and then G-LoSA (Figure 4b). However, FLAPP and SiteMotif consistently achieve longer alignments compared to G-LoSA (Supporting Information: Data sets-Supplementary.xlsx.)

3.2. Runtime Performance. *3.2.1. Benchmarking the Runtime of FLAPP.* Traditional alignment based site comparison methods, despite being accurate, are often many orders of magnitude slower than alignment free methods. This is because alignment operations such as rotation, translation, and least squares superpositions are computationally expensive. The FLAPP algorithm and implementation is designed to mitigate these factors. Large-scale analyses of binding sites typically involve the use of alignment-free methods such as PocketMatch and RAPMAD because of their quick execution times of ~1/250 s and ~1/100 s for comparing a site pair. A comparison of 1 million binding site pairs is expected to complete in around 70 min. In comparison, SiteMotif or G-LoSA, would take around 6 days. Alignment-free methods output a score that represents the similarities between the sites, while alignment-based methods provide us with full length residue−residue correspondences. FLAPP bridges the gap between these sets of tools, bringing in the accuracy of alignment-based methods while competing with the speed of alignment-free methods. To obtain an accurate estimation of the performance, FLAPP was measured against different sets ranging from 10^2 to 10^6 sites taken randomly from PDB. The performance was timed with G-LoSA and SiteMotif. G-LoSA can be executed in one of two ways: either with or without file pre-processing. We term the former as "G-LoSA-Preprocessed" and the latter as "G-LoSA-Native". The native version processes the input PDB files on each run using a Java program, while the input files are processed beforehand in the preprocessed version. From Figure 5a, it is clear that FLAPP tops the list as the fastest alignment program, outperforming both SiteMotif and G-LoSA by a huge margin. Even the preprocessed version of G-LoSA is 4× slower than our method.

3.2.2. Parallelized Implementation Scales Well Across Multiple CPU Cores. The speed benchmarking analysis showed that FLAPP on average takes 10 ms to align a pair of sites on a single desktop computer. In other words, the single threaded version of FLAPP aligns 120 pairs per second. This is by far the fastest implementation of alignment-based site comparison software that exists. However, modern CPUs are designed to have many cores, which permits the concurrent execution of individual statements across multiple processes. We therefore

https://doi.org/10.1021/acs.jcim.2c00967
J. Chem. Inf. Model. 2022, 62, 4810−4819

aimed to utilize all available CPUs to further accelerate our algorithm. The FLAPP code is wrapped in a multiprocessing module that will assign each pair of pockets to a different computing resource. To assess the performance, we randomly generated 1 million pairs of binding sites from the PDB for which serial implementation took 140 min to complete. The performance of the parallel implementation was evaluated as the ratio of the time taken by the serial version (T_s) to time taken by parallel version (T_p) at different no. of CPU cores (N), where $N \geq 1$ (Figure 5b). At $N = 2$, FLAPP runs approximately 2× faster than the serial version. This implied that FLAPP can scale with an increase in the number of CPUs. Scaling was however not linear because the multiprocessing introduces CPU overheads. Utilizing 12 cores, FLAPP was able to align 1080 site pairs per second, a ninefold improvement over the serial version. At $N = 12$, FLAPP sets a new benchmark as the first alignment-based approach that needs only 1 ms to compare two sites. A standard desktop machine equipped with a Intel Core i7-10700 was used for this analysis.

3.3. New Application Capabilities that Emerge out of the Speed: Case Study of Vitamin B12 Binding Sites. We showcase the performance of FLAPP in achieving a PDB-wide scan of 3D binding sites for vitamin B12 binding sites as an example. Vitamin B12 is a large ligand and is known to bind to diverse protein structures.[20] We performed a two-level systematic analysis: (i) to find a vitamin B12 binding motif and (ii) to scan the motif against PDB. First we used SiteMotif, a recently developed method from our group for multiple site alignment. A total of 83 proteins were extracted from the PDB that are complexed with vitamin B12. Identical sequences within them were removed using a sequence identity cutoff of 95 which resulted in 51 proteins. We then removed redundancies at structural levels by doing all-to-all 3D structural alignment using the TM-align program.[21] TM-align, in addition to doing structural alignment, returns a TM-score to numerically quantify the extent of similarity between proteins. As reported by the authors, TM-score > 0.5 indicates that structures being compared are likely to be similar. We used the same cutoff and grouped all 51 proteins into 12 structure-based clusters, which indicated that the vitamin B12 ligand binds to proteins from 12 different structural families (Supporting Information B12-Supplementary.xlsx). Next, we checked if the sites of 12 proteins share any commonality in their binding sites using FLAPP. We observed that 7 of the 12 share significant similarity in their sites (alignment length > 10, RMSD < 1 Å) (Supporting Information, Figure S3) despite belonging to different SCOP folds, suggesting this to be a predominant binding site type for binding vitamin B12. We examined the binding sites for the remaining five clusters and found that they were very different from the other groups, and as they formed singletons, we did not consider them for further analysis. The selected seven vitamin B12 proteins are 1XRS, 2REQ, 3IV9, 3KOX, 5C8A, 5CJV, and 6WTE, showed an average sequence identity of 20.8% among them and a mean structural deviation of 18.2 Å, in consonance with their diversity. We then examined the positional conservation among these using SiteMotif (by aligning each of them onto 3KOX taken as a representative). This yielded a structural motif (Figure 6) that contained a histidine coordinating with the cobalt, an aspartic acid, and serine. These residue conservation with vitamin B12 has been reported earlier.[20,22]

We then performed a one-versus-all binding site alignment between the vitamin B12 motif and the entire set of known

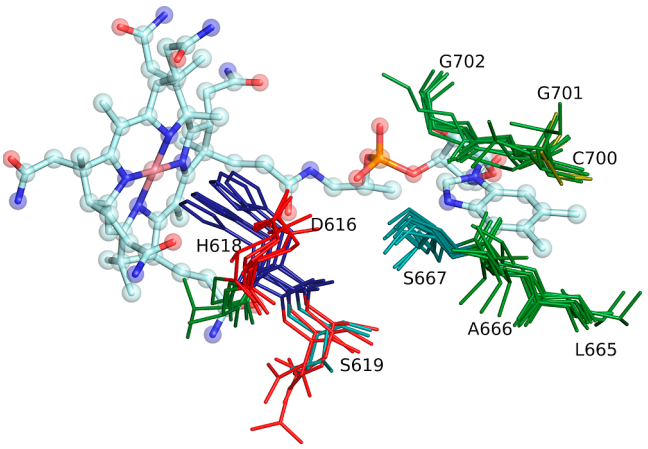


Figure 6. Deriving the structural motif for vitamin B12 binding sites. For the set of seven unique vitamin B12 proteins, the binding sites are extracted and the conserved residues are identified using SiteMotif. The motif thus derived is used with FLAPP for large-scale analysis such as a PDB-wide search. Vitamin B12 is represented in ball and stick and the conserved residues are represented in sticks. Residues are colored according to their physicochemical property.

binding sites from PDB using FLAPP. FLAPP successfully recovered all seven known binding sites containing the motif. We tested if the motif is present in all vitamin B12 binding proteins including those in homologous proteins that are not crystallized with vitamin B12. To do this, we established an unbiased scan of the vitamin B12 structural motif against pockets from an extensively augmented pocketome PocketDB that we have previously developed.[23] PocketDB, consists of high confidence putative pockets for all structures in PDB (PDBv-2014) using pocket prediction algorithms (139,718 pockets). FLAPP scanned the vitamin B12 motif against 139,718 pockets of PocketDB (978,026 alignment operation). FLAPP completed the scan on a 12 core machine in 10 min. FLAPP reported 26 hits for the vitamin B12 motif. These 26 proteins were known to utilize vitamin B12 for its function but the structures complexed with vitamin B12 are available only for 12 of them. This indicates that FLAPP was able to correctly identify the apo form of the vitamin B12 binding proteins. An exhaustive site alignment exercise, such as one which involves PDB level scale, will provide a number of insights that were previously considered to be impossible due to the inherent complexity of the current algorithms during runtime. Here, we demonstrate that FLAPP can perform such analysis on a normal desktop system in reasonable time.

4. DISCUSSION

Biological reactions are driven by biomolecular interactions. Elucidating the characteristics of interactions governing between proteins and ligands can provide functional insights as to why cells behave the way they do or which proteins a particular ligand can bind to. The demand to address such questions has gained significant attention as it drives us to demystify novel protein targets and cellular pathways that are currently unknown to science. This is even more important if the ligand being investigated is a drug molecule or a probable lead candidate. Often, precise identification of small molecule binding sites provide direct answers into understanding the mechanism by which proteins recognize ligands.

https://doi.org/10.1021/acs.jcim.2c00967
J. Chem. Inf. Model. 2022, 62, 4810−4819

We present here a highly scalable algorithm for binding site comparison that provides fast and accurate residue-level alignments and facilitates rapid identification of similar binding sites at a proteome level. Because binding sites provide a direct window into understanding the function of proteins, this opens up the possibility for extensive protein annotation by knowledge transfer based on site similarity. Alignments at the binding site level allow us to probe similarity at the residue level, providing us with much higher confidence in our annotation process. PDB currently (v 30 April 2022) contains 189,915 structures—of which 68,968 are crystallized with specific ligands.[15] Furthermore, pocket prediction methods will expand the repertoire of potential ligand recognition sites by several folds.[12,13,23,24] FLAPP was tailor made for binding site comparisons at a scale that meets the pace at which data are being generated. Our benchmarks show that FLAPP produces alignments comparable to the existing state-of-the-art at blazing speed of 1/80th of a second per comparison. FLAPP accomplishes this speed as a consequence of various optimizations at the implementation level which includes, grouping of residues, hash table to keep track of traversed path, SIMD instructions whenever possible, and lastly the use of compiler libraries to compile the whole program based on the system specific architecture. In addition, FLAPP supports multiprocessing and can harness the power of multiple CPU cores to provide alignments at a speed of about 1 ms per pair on a standard 12 core desktop computer.

Similarities at the binding site level can be used in both protein-centric fashion to identify potential ligands that can bind to it and in a ligand-centric fashion to identify different proteins that can potentially bind the small molecules. The most obvious use of such a capability is to annotate ligand-binding ability at a proteome-wide scale. In addition, binding site alignments have the potential for immense application in various aspects of drug discovery—specifically in drug repurposing and in identifying drug off-targets.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00967.

> Assessing the accuracy of FLAPP using data sets constructed to query the ability to find an alignment in specific scenarios; assessing the sensitivity of FLAPP to distance perturbation at F_{min} 0.7; binding site similarity of the 12 selected vitamin B12 binding site representatives; sensitivity of FLAPP to minor perturbations in residue position and type explored using 19 diverse binding sites binding diverse ligands; validation of FLAPP at varying levels of complexity; comparison of FLAPP to state-of-the-art methods G-LoSA and SiteMotif; and case study of vitamin B12 binding sites (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Nagasuma Chandra − *Department of Biochemistry and BioSystems Science and Engineering, Indian Institute of Science, Bangalore 560012 Karnataka, India;* orcid.org/0000-0002-9939-8439; Email: nchandra@iisc.ac.in

Authors

Santhosh Sankar − *Department of Biochemistry, Indian Institute of Science, Bangalore 560012 Karnataka, India;* orcid.org/0000-0002-2755-5052

Naren Chandran Sakthivel − *Department of Biochemistry, Indian Institute of Science, Bangalore 560012 Karnataka, India;* orcid.org/0000-0001-7728-0985

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00967

Notes

The authors declare no competing financial interest.

Code availability: The source code of FLAPP is available at https://github.com/santhoshgits/FLAPP.

■ ACKNOWLEDGMENTS

We acknowledge support from the Bioinformatics grant, Department of Biotechnology, Government of India (BT/PR40187/BTIS/137/3/2021). N.C.S. is funded by the Council of Scientific & Industrial Research, Government of India [Senior Research Fellow—File no: SPM-07/079(0287)/2019-EMR-I].

■ REFERENCES

(1) Loewenstein, Y.; Raimondo, D.; Redfern, O. C.; Watson, J.; Frishman, D.; Linial, M.; Orengo, C.; Thornton, J.; Tramontano, A. Protein Function Annotation by Homology-Based Inference. *Genome Biol.* **2009**, *10*, 207.

(2) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(3) Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7*, No. e1002195.

(4) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J.; Finn, R. D.; Bateman, A. Pfam: The Protein Families Database in 2021. *Nucleic Acids Res.* **2021**, *49*, D412−D419.

(5) Yeturu, K.; Chandra, N. PocketMatch: A New Algorithm to Compare Binding Sites in Protein Structures. *BMC Bioinf.* **2008**, *9*, 543.

(6) Krotzky, T.; Grunwald, C.; Egerland, U.; Klebe, G. Large-Scale Mining for Similar Protein Binding Pockets: With RAPMAD Retrieval on the Fly Becomes Real. *J. Chem. Inf. Model.* **2015**, *55*, 165−179.

(7) Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein-Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623−637.

(8) Yeturu, K.; Chandra, N. PocketAlign A Novel Algorithm for Aligning Binding Sites in Protein Structures. *J. Chem. Inf. Model.* **2011**, *51*, 1725−1736.

(9) Lee, H. S.; Im, W. G-LoSA: An efficient computational tool for local structure-centric biological studies and drug design. *Protein Sci.* **2016**, *25*, 865−876.

(10) Sankar, S.; Chandra, N. SiteMotif: A Graph-Based Algorithm for Deriving Structural Motifs in Protein Ligand Binding Sites. *PLoS Comput. Biol.* **2022**, *18*, No. e1009901.

(11) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. SiteEngines: Recognition and Comparison of Binding Sites and Protein-Protein Interfaces. *Nucleic Acids Res.* **2005**, *33*, W337−W341.

(12) Kalidas, Y.; Chandra, N. PocketDepth: A New Depth Based Algorithm for Identification of Ligand Binding Sites in Proteins. *J. Struct. Biol.* **2008**, *161*, 31−42.

(13) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinf.* **2009**, *10*, 168.

(14) Ghersi, D.; Sanchez, R. Improving Accuracy and Efficiency of Blind Protein-Ligand Docking by Focusing on Predicted Binding Sites. *Proteins* **2009**, *74*, 417−424.

(15) Kabsch, W. A. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976**, *32*, 922−923.
(16) Lam, S. K.; Pitrou, A.; Seibert, S. Numba: A LLVM-Based Python JIT Compiler. *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC—LLVM '15*; ACM Press: Austin, Texas, 2015; pp 1−6.
(17) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(18) Lo Conte, L. SCOP: A Structural Classification of Proteins Database. *Nucleic Acids Res.* **2000**, *28*, 257−259.
(19) Ausiello, G.; Peluso, D.; Via, A.; Helmer-Citterich, M. Local Comparison of Protein Structures Highlights Cases of Convergent Evolution in Analogous Functional Sites. *BMC Bioinf.* **2007**, *8*, S24.
(20) Sukumar, N. Crystallographic Studies on B12 Binding Proteins in Eukaryotes and Prokaryotes. *Biochimie* **2013**, *95*, 976−988.
(21) Zhang, Y. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302−2309.
(22) Tollinger, M.; Konrat, R.; Hilbert, B. H.; Marsh, E. N. G.; Kräutler, B. How a Protein Prepares for B12 Binding: Structure and Dynamics of the B12-Binding Subunit of Glutamate Mutase from Clostridium Tetanomorphum. *Structure* **1998**, *6*, 1021−1033.
(23) Bhagavat, R.; Sankar, S.; Srinivasan, N.; Chandra, N. An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3D Structure. *Structure* **2018**, *26*, 499−512.e2.
(24) Hernandez, M.; Ghersi, D.; Sanchez, R. SITEHOUND-Web: A Server for Ligand Binding Site Identification in Protein Structures. *Nucleic Acids Res.* **2009**, *37*, W413−W416.

https://doi.org/10.1021/acs.jcim.2c00967
J. Chem. Inf. Model. 2022, 62, 4810−4819