JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

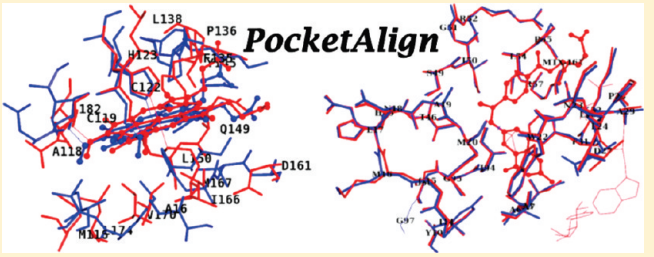
PocketAlign A Novel Algorithm for Aligning Binding Sites in Protein Structures

Kalidas Yeturu and Nagasuma Chandra*

Bioinformatics Centre, Indian Institute of Science, Bangalore-560012, India

Supporting Information

ABSTRACT: A fundamental task in bioinformatics involves a transfer of knowledge from one protein molecule onto another by way of recognizing similarities. Such similarities are obtained at different levels, that of sequence, whole fold, or important substructures. Comparison of binding sites is important to understand functional similarities among the proteins and also to understand drug cross-reactivities. Current methods in literature have their own merits and demerits, warranting exploration of newer concepts and algorithms, especially for large-scale comparisons and for obtaining accurate residue-wise mappings. Here, we report the development of a new algorithm, PocketAlign, for obtaining structural superpositions of binding sites. The software is available as a web-service at http://proline.physics.iisc.ernet.in/pocketalign/. The algorithm encodes shape descriptors in the form of geometric perspectives, supplemented by chemical group classification. The shape descriptor considers several perspectives with each residue as the focus and captures relative distribution of residues around it in a given site. Residue-wise pairings are computed by comparing the set of perspectives of the first site with that of the second, followed by a greedy approach that incrementally combines residue pairings into a mapping. The mappings in different frames are then evaluated by different metrics encoding the extent of alignment of individual geometric perspectives. Different initial seed alignments are computed, each subsequently extended by detecting consequential atomic alignments in a three-dimensional grid, and the best 500 stored in a database. Alignments are then ranked, and the top scoring alignments reported, which are then streamed into Pymol for visualization and analyses. The method is validated for accuracy and sensitivity and benchmarked against existing methods. An advantage of PocketAlign, as compared to some of the existing tools available for binding site comparison in literature, is that it explores different schemes for identifying an alignment thus has a better potential to capture similarities in ligand recognition abilities. PocketAlign, by finding a detailed alignment of a pair of sites, provides insights as to why two sites are similar and which set of residues and atoms contribute to the similarity.



INTRODUCTION

In the post genomics era, recognizing similarities and detecting relationships among protein molecules is a fundamental task. Functional characterization of protein molecules can be achieved at different levels, that of sequence, the whole structure, or at the level of binding sites.[1,2] What ultimately matters for a given protein is whether it has the ability to perform a given function and not what means it uses to achieve this capability. Binding sites can be easily viewed as the functional or operative parts of protein molecules.[3] A given function could be conserved simply because a substructure such as at the binding site is conserved conferring a particular ligand recognition ability. It is of immense interest therefore to compare protein molecules at the level of their binding sites irrespective of whether or not they share sequence or fold-level similarities. A number of cases are known where binding sites and hence function is conserved in protein molecules of different sequence and fold families, serine proteases being good examples.[4,5] As the structural databases are growing at high paces,[6] the need to carry out comparisons across datasets so as to effectively and efficiently mine such similarities, is also rising. With the availability of structural information on proteins and the availability of methods for prediction and comparison of binding sites, genome scale target validation can be carried out. Such a possibility has been explored in a novel drug discovery pipeline reported from our laboratory, targetTB,[7] where binding sites are detected in modeled structures of proteins using PocketDepth[8] in the pathogen, *Mycobacterium tuberculosis*, and the host followed by a large-scale comparison of the sites using PocketMatch (PM) utilizing cluster computers. In addition to the large-scale scenario, comparison of binding sites finds application in detecting determinants of ligand recognition[10,11] in a family of proteins, such as carbohydrate binding proteins. Given the wide array of applications, it is necessary to have highly effective algorithms for comparison of binding sites.

Comparing binding sites at the three-dimensional (3D) level, however, is by no means a trivial task. Additional challenges are presented in comparing binding sites due to three main reasons: (i) Binding sites are often made of residues discontinuous in sequential space, resulting in a need for searching across an exponential number of possible alignments; (ii) binding sites are

Received: March 18, 2011
Published: June 11, 2011

fairly small, consisting of only a handful of residues, leaving not much opportunity to detect a signal or to unambiguously measure the quality of the alignment; and (iii) only a portion of the binding sites could be similar and yet retain a given function. Often, there is no prior systematic information about which residues may be crucial for ligand binding and hence for function in different proteins, making it difficult to weight importance of individual residues. The residues in a site however are all situated in a single spatial zone forming the binding site, making it possible to employ methods that utilize spatial proximity for comparison.

A handful of methods are available in literature for binding site comparison.[12−19] They can be largely classified into those based on: (i) maximal common subgraph search (MCSS) methods, where sites are represented as graphs with atoms as nodes and Euclidean distances among them as edges; maximal common subgraphs in a pair of sites are then identified, which indicate the extent of similarity in them using methods available for that purpose;[17−22] and (ii) geometric hashing methods, in which three-atom sets forming triads are superposed, appropriate rotation and translation matrices computed to find the largest number of matching atoms between two sites.[16,23,24] CAVBASE,[13] the method by Kinoshita and Nakamura,[25] and ProBis[17−19] are examples of the first method, whereas SitesBase[26] and the method by Minai and co-workers[27] are examples of the second. As with any algorithm of this nature, operating in the 3D space, these algorithms also have a high computational cost and are not directly amenable for high-throughput analysis and often work only on predefined datasets. The site representations in some of these methods set at the level of individual atoms are seen to lead to conflicts in final residue mapping in some cases, also making it difficult to recognize moderate or part similarities in a pair of sites, in some cases.

There are a few indirect methods reported for comparison of sites, which utilize features, such as shape descriptors, charge distribution, hydrophobicity, surface area, or site volumes. These methods are inherently fast and provide an overall indication of the extent of similarity between two sites, but an important drawback with these methods is that residue-wise correspondences between the two sites are not obtained.

We recently developed a shape signature-based algorithm PM[9,28] that quickly compares a pair of given binding sites by measuring similarities in their overall shapes. In this method, sites are represented as sorted distance lists of different types capturing chemical nature and geometry of the binding site. Such an abstraction made PM suitable for carrying out large-scale, high-throughput comparisons. However, it does not compute or output explicit information on residue-wise correspondences. Residue- and atom-wise correspondences become necessary to map differences in the sites to the sequence level information, and more importantly to exploit differences or similarities between sites for designing specific ligands and also to rationalize mechanisms of enzyme function and drug action alike. Here we present a new algorithm PA (PA), for aligning a pair of binding site structures which leverages advantages of the fundamental concepts used in PM at the same time overcoming the problem of obtaining residue-wise correspondences, which is essential to judge the quality of alignment. Geometric perspectives are computed and compared with each other, enabling exploration of identifying different residue pairings and grouping of the sets of pairings into mappings. Sensitivity of the alignment is significantly increased by refining initial seed alignments. Though PA has been built on the concepts of PM at the initial stage for computation of all pair distance lists, all the subsequent steps in the PA workflow are new algorithms developed to comprehensively explore the alignment space, identify best seeds, and refine the alignment to extract maximal best alignments. Hence, it is a completely new tool and has a different objective as compared to PM. PA unlike PM provides a detailed alignment and insights as to why two sites are similar and provides an idea about which set of residues and atoms contribute to the similarity and which do not.

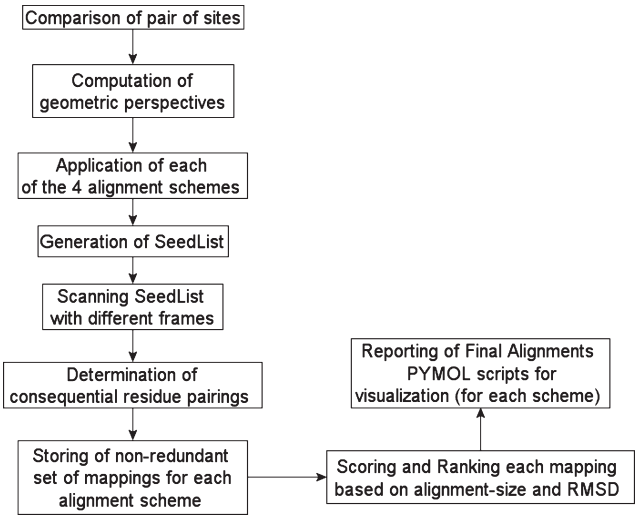


Figure 1. Flowchart depicting overview of the PA algorithm. Various steps involved include construction of scoring matrix between residues, generation of seed mappings, determination of consequential residue pairings, and final reporting of alignments with visualization scripts are all indicated.

METHODS

The PA algorithm has been coded in C, using gcc (GCC) 4.1.2 20070925 and executed on Linux x86 64 2.6.21−1.3194.fc7 using wrapper scripts in PERL and BASH. The least-squares superposition of point sets was carried out using the Kabsch algorithm.[29] PyMol was used for visualizing final alignments.

RESULTS AND DISCUSSION

The overall strategy is shown in Figure 1. First, geometric perspectives for each residue in a given site are computed and compared with the residues in the second site to generate seed mappings. Alignments are obtained by evaluation of mappings in all frames. Each mapping by itself is derived through a greedy approach that incrementally combines residue pairings.

Representation of a Binding Site. In the structure of a protein−ligand complex, residues having one or more atoms in a zone of 4 Å around any atom of the ligand were taken as the binding site. Complete residues of such atoms are taken. For the purpose of obtaining distance lists and geometry perspectives, only C_α atoms of the residues are considered. For alignment purposes, however, each site has been represented using four different schemes: (i) that considers the backbone atoms (N, C_α, C, O) and the centroid of the side chain ($C_{centroid}$ or CNTR); (ii) that considers only the backbone (N, C_α, C, O) atoms; (iii) that considers only the side chain atoms (C_β) and $C_{centroid}$; and (iv) that considers only $C_{centroid}$, as representative atoms of each residue. The terms used in the algorithm and the rest of the manuscript are described in a table in the Supporting Information.

Computing Geometric Perspectives. Within each site, for each residue, distances with all other residues (using only C_α atoms) are computed, sorted based on their distance values in descending order, and stored as sorted distance lists. These lists serve as a set of geometric perspectives (GPs) for a given site, each one corresponding to that obtained with one residue in the site as the anchor. The GPs between two binding sites are then compared in an all versus all manner to construct a matrix referred to as geometric perspective score (GPS) matrix. A pair of GPs is compared simply by counting the number of common distance elements in the corresponding lists. For these two distance elements, whose values differ by less than 0.5 Å, they are considered to be matching and hence counted as common elements. We have systematically varied the threshold limit for considering distance elements and reported it in the previous article describing PM.[9] Based on that, we find that a value of 0.5 Å has a reasonable trade-off between sensitivity and accuracy. The GPS matrix will hold the number of common elements between every pair of residues between two sites, as seen through their individual GPs. The higher the score, the more similar the perspectives, or in other words, the more similar is the location of the residues in context of the location of all other residues in the site.

To give an example, a binding site containing four residues will have four perspectives in it, whereas a second site with three residues will have three perspectives in it, and when compared, they will generate a GPS matrix of 4×3. In order to capture similarity between pairs of residues from the two sites, BLOSUM[30] substitution scores are used. A more appropriate substitution matrix for binding site superposition might be useful. However matrices of that kind are not as yet available. Keeping this in view, a provision has been made for a user-specific matrix to be used in place of the default BLOSUM 62 matrix.

The PA algorithm enumerates alignments based on detection of seed mappings corresponding to fewer pairing of residues between sites, followed by detection of consequential residue pairings. A seed mapping relates to substructural feature of a binding site. Using this framework, use of BLOSUM substitution scores serve to focus the search space. Thus the values in the GPS matrix are multiplied with the substitution scores in BLOSUM62 for the pair of residues, to obtain a GPSxBL matrix.

Seed Alignment. The next step for obtaining an alignment is to identify correspondences between residues in the two sites. The elements in the GPSxBL matrix are sorted in a descending order of their scores into a linear array, which retains the residue pair information (Figure 2). An example is illustrated in Figure 2, which shows that in a 4×3 matrix of the two sites, the highest value of 18 between first residue (1) of site 1 and the second residue (B) of site 2 is written as (18(1,B)). The residue pairs in this list, referred to as a 'SeedList' are considered as potential 'pairings', of which only a small fraction will remain in the final alignment.

Generation and Mappings. The SeedList thus generated has values with the highest at the left most, which can serve as a seed to get the alignments. Starting from the left most residue pair for alignment, the mapping generation procedure attempts to incrementally add subsequent pairings in the list and retains them if they are within the defined threshold of similarity listed in Table S2, Supporting Information. The set of all pairings that can be simultaneously present will comprise a mapping in that frame. There can be multiple mappings in one frame, since alternate residue pairings may be possible. The same exercise is repeated with the second best pairing as the starting point, which will give another mapping in the second frame. For each addition of a pairing, a least-squares superposition using the Kabsch algorithm is carried out, and root-mean-square deviations (RMSD) values of the atoms are computed, which are used for comparing against thresholds. This amounts to a greedy approach of finding mappings based on the most promising (highest score) pairing of residues. This type of scanning is carried out until the whole list is scanned. A mathematical formulation of generation of seed mapping in PA is depicted in Figure 3.

The seed mapping generated at this stage will contain correspondences present in the list that can all simultaneously exist. It is also ensured that there is no double pairing for a residue in either of the sites. If a residue of the first site is already paired with one in the second site, then both the residues are not considered for pairing with other residues. This imposes a strict one-to-one pairing of residues in the two sites.

Within each frame, backtracking is not carried out, to avoid the high computational cost. This however leads to a problem of missing out on certain alternative sets of pairings. The problem is overcome by: (i) considering consequential alignments within each frame for the best mapping and (ii) considering all possible frames for generating alignments.

Consequential mappings were obtained for the best mapping in each frame, by identifying any residues in the second site that were located within a defined threshold (Table 2) of any residues of the first site, in addition to the pairings obtained in the seed mapping. Including such additional correspondences (pairings) one at a time, a least-squares superposition of the second site with respect to the first is again carried out, and the corresponding rotation and translation matrices for the largest set that satisfied the defined RMSD thresholds are retained as the mapping for that frame.

In order to determine the consequential pairs, a 3D grid data structure is employed with each grid cell having information about points that map on to it. Pairs of corresponding atoms are determined by probing neighboring grid cells for given range of distance about an atom. The mappings from all possible frames are stored in a database, for further processing. The entire seed alignment and mapping procedure is repeated for all four alignment schemes. It is possible to have redundant mappings between two frames. Such redundancies are checked, and the smaller of the two mappings removed before storing in the database.

Obtaining Atom−Atom Mapping. For each atom of the superposed site (second), a corresponding closest atom of the first site (fixed) is determined by considering all member atoms of the grid cells surrounding the cell the atom maps to. The radius within which the closest atoms are found is a variable parameter, default values indicated in Table S2, Supporting Information. From those atoms in the vicinity, only those that match in their chemical types are considered. Further, the atoms are chosen such that they satisfy a strict one-to-one correspondence between atoms of the first and the second sites, to ensure unambiguous alignment. In cases where multiple correspondences with matching position and type are possible for a given atom, ambiguities are resolved by simply taking the first atom in the order of occurrence in the input. Only the set of atoms that satisfy all criteria mentioned above is considered for final superposition. It is possible to have redundant mappings between two frames. Such redundancies are checked, and the smaller of the two mappings removed before storing in the database. The entire

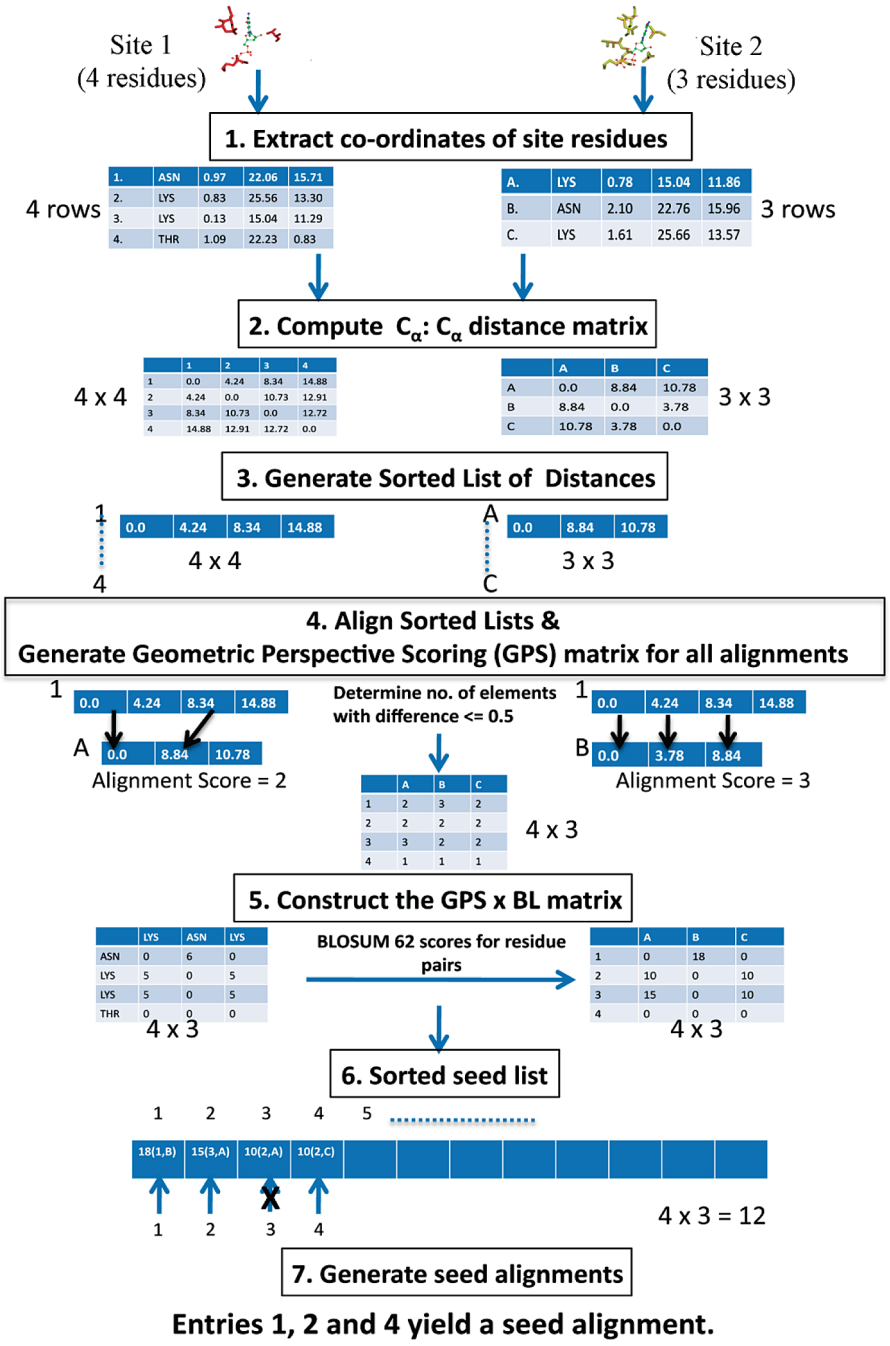


Figure 2. An illustration of generation of seed alignment for a pair of sites of four and three residues. Step 1 denotes extraction of the coordinates. Step 2 denotes computation of all pair distances between residues (C_α atoms) within each site. Step 3 denotes generation of sorted list of distances as GPs. In step 4, each pair of distance lists or the GPs between residues are compared to produce a GPS matrix computed by comparison of each pair of GPs between the two sites. For a pair of distance lists, the number of matching elements (shown by arrows) is determined. A pair of distances match only if the difference is $\leq$0.5 Å. In step 5, BLOSUM62 substitution scores between corresponding pairs of residues are multiplied with the corresponding elements in the GPS to give GPSxBL. In step 6, the elements of the GPSxBL are sorted in descending order retaining the residue pairs they correspond resulting in the SeedList. In step 7, the seed alignment generation algorithm starts from the left most (left most arrow pointing to SeedList) and assigns residues only if one-to-one pairing exists. The third element is crossed since the residue 'A' of site 2 is already paired. Only the first frame is shown here for clarity.

seed alignment and the mapping procedure is repeated for all four alignment schemes. A mathematical abstraction of generation of atom-wise correspondences and determining consequential residue pairing is depicted in Figure 3. For each atom in site 2, corresponding atoms in site 1 are identified by examining the spatial neighborhood of that atom.

Parameters Involved. The RMSD threshold for a mapping to be considered as matched depends on the alignment scheme used. We have chosen a default threshold of 2.5 Å implying that any alignment with higher RMSD would be rejected. For consequential pairings, a new pair of residues is only added to the existing set of residues only if the RMSD is within bounds.

dx.doi.org/10.1021/ci200132z |J. Chem. Inf. Model. 2011, 51, 1725–1736

Representation of sites

Two binding sites are represented as sets of residues

$S = \{R_1 ... R_m\}$ where R_i is i^{th} residue of first site

Each residue defines a partitioning of the set of atoms, A

$R = \{a \in A\} \subset A$

$R_i \cap R_j = \emptyset (\forall i \neq j \in |S|)$

Where $|S|$ denotes cardinality of the set, S

Similarly second site is represented by $S' = \{R'_1 ... R'_n\}$ on set of atoms, A'

Generation of 'SeedList'

Chemical similarities are denoted by a function $BL : S \times S' \rightarrow N$

Geometric similarities (GPS) are denoted by $GPS : S \times S' \rightarrow N$

A combination scoring scheme is defined $GPS \times BL_{ij} \rightarrow GPS_{ij} * BL_{ij}$

A linearization of $GPS \times BL$ is performed

A one-to-one function is defined $L : [1...m] \times [1...n] \rightarrow [1...m * n]$

SeedList is created by obtaining values from $GPS \times BL$

$SeedList^V_{L(i,j)} \leftarrow GPS \times BL_{ij}$ for storing the values

$SeedList^P_{L(i,j)} \leftarrow (i, j)$ for storing the residue pairs

SeedList is sorted such that $(\forall p \leq q) SeedList^V_p \geq SeedList^V_q$

Generation of seed mappings

A mapping is defined as residuewise correspondences between the two sites

A one-to-one function, for a mapping $M : [1...m] \rightarrow [1...n]$

Seed mapping or alignment B is derived by traversal of SeedList

$B \leftarrow \{(p, q)\} \subset SeedList^P$

Detection of consequential residue pairings

Determine matching atoms between the two sites

$\rho : A \times A' \rightarrow 0, 1$ where 1 indicates pair of atoms mapped, and 0, otherwise

A pair of atoms are mapped if their chemical types match and are spatially proximal

Between every pair of residues between sites, scores S_{ij} are defined

$S_{ij} = \{(\forall a \in R_i \subset A, a' \in R'_j \subset A') \rho(a, a') = 1\}$

Residue pairing are generated based on maximal number of matching atoms between them

A one-to-one function is defined $M : [1...m] \rightarrow [1...n]$

A maximal mapping is derived by $Max \leftarrow \underset{M}{argmax} \sum_{i \in M_{domain}, j \in M_{range}} S_{i,j}$

Figure 3. Mathematical formulation of generation of seed and consequential mappings in PA algorithm.

The type and the number of points considered for alignment of a pair of residues depend on the alignment scheme used.

Final Alignment. The large number of mappings stored in the database, each representing a candidate alignment are evaluated based on a variant of the Q score metric defined in ref 28 as

$$Q = \frac{N_{align}^2}{(1 + (RMSD/R_0)^2) N_1 N_2)}$$

In our case, since the different mappings for a given pair of sites will all have the same number of atoms in the two sites across mappings, the terms N_1 and N_2, are not insightful, hence the new score, $Q^* = N^2/\text{RMSD}$ is used here, where, N is the number of matching atoms between sites. It must be noted that, while the Q score is useful for comparing alignments across different site pairs, the Q^* score is useful for comparing different mappings for the same site pair.

In order to ensure that there is no redundancy among different mappings generated for a pair of sites, a mapping is only stored in the database if it is different from the already stored mappings belonging to the same alignment schema. The best mapping among all the frames for each alignment scheme as well as the best mapping among all schemes are output as final alignments. A PyMol script is generated for each reported mapping, to enable visualization.

Rationale Behind the Design of the Algorithm. A key feature of the algorithm is the computation and the use of GPs for abstracting geometric characteristics of a given binding site. GPs capture the spatial distribution of all other residues as viewed from the context of a given anchor residue. In principle, one perspective itself contains the description of the entire site. However to be useful for comparing two sites, the anchor residue will have to be chosen such that it is characteristic and sufficiently discriminatory for that site, which is often difficult in practice. Here, perspectives are therefore computed using each residue in the site as an anchor residue. All perspectives between the two sites are then compared to extract the best pairing of residues (see Figure 2). Since the operation of computing perspectives requires very little time, calculating multiple GPs for each site does not significantly alter the running time of the algorithm. In order to capture biological relevance between the pair of residues being compared between the sites, the values obtained by comparing the GPs are augmented with the substitution scores for the pair of residues in the BLOSUM62 substitution matrix. The implementation of PA provides the user flexibility in choosing another scoring scheme in place of Blosum62, if desired.

Each cell in the GPSxBL matrix contains a score that reflects the extent of similarity between i^{th} perspective of site 1 and j^{th} perspective of site 2. In other words, these scores capture the similarity between the two sites as viewed through different residues as anchors. Not all residues in a given site need to (or generally will) match with those in the second. Hence only some of the GPs are likely to be more informative than the others for the purposes of comparison between sites. This score gives an overall idea of the extent of similarity between the two sites but does not explicitly contain information required for obtaining residue-wise correspondences. In order to obtain an alignment, an exact one-to-one correspondence is needed. GPSxBL contains information that is implicit but not readily useful for inferring residue-wise correspondences.

In computing alignments, evaluating all possible mappings of residues and determining the best set between a pair of sites is a computationally intractable (NP-Complete) task. We use the scores in the GPSxBL matrix to obtain initial mappings, (referred to as seed mappings), by a simple greedy approach of picking up the top scoring pair first and incrementally adding the next top scoring pair(s) to identify all those pairs which are 'in phase' with the first pair. Two pairings can said to be 'in phase' only if they can both be contained simultaneously in a given mapping. The top scoring pair of GPs between the sites will automatically identify the perspective that will have the maximum number of residues positioned in similar spatial arrangements in the two sites. Hence it is chosen first, and the anchor residues of the individual perspectives of the pairs are taken as the first pair or the basis to generate a mapping. The pairs are retained only if their addition does not increase the RMSD beyond the chosen threshold (default of 2.5 Å), as judged by a least-squares superposition using all pairs until that point and including the new pair. Similarly, taking the second top pair in the GPSxBL matrix will lead to a second mapping and so on. The mappings are then evaluated based on the Q^* score metric, which considers the size of the alignment as well as the RMSD. Incorporating consequential pairings to each map helps in identifying any residue pair that may be spatially located in the two sites, albeit with slightly less stringent distance criteria. The same procedure of incrementally adding pairings used for consequential pairs as well ensures that including them does not cross the RMSD bounds. Typically a considerable improvement over the seed alignments is seen both in terms of the number of residue pairings and the least RMSD among them.

From the top-ranked mapping, the corresponding set of atoms in the residues between the two sites is obtained, and the least-squares superposition carried out, resulting in the finer refinement of the alignment, which is output as the final alignment.

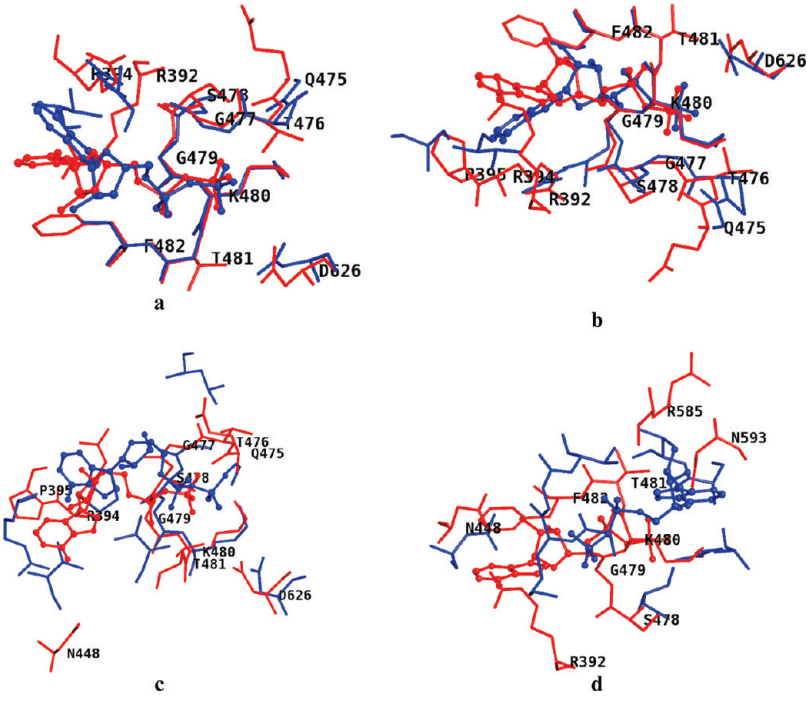


Figure 4. PA alignment of 3KAR_ADP and 1UKZ_ADP for four schemes (a−d). Only aligned residues are shown for clarity. Red represents 3KAR, and blue represents 1UKZ. Superpostions a−d correspond to alignment schemes 1−4, respectively. Ligands are shown in ball and stick representation, where binding site residues are shown in stick representation in this and all the subsequent figures. A naming convention for binding sites used here is PDBID_LIG, where PDBID indicates the protein's PDB code and LIG indicates the three letter code bound ligand. All figures are generated using PYMOL.

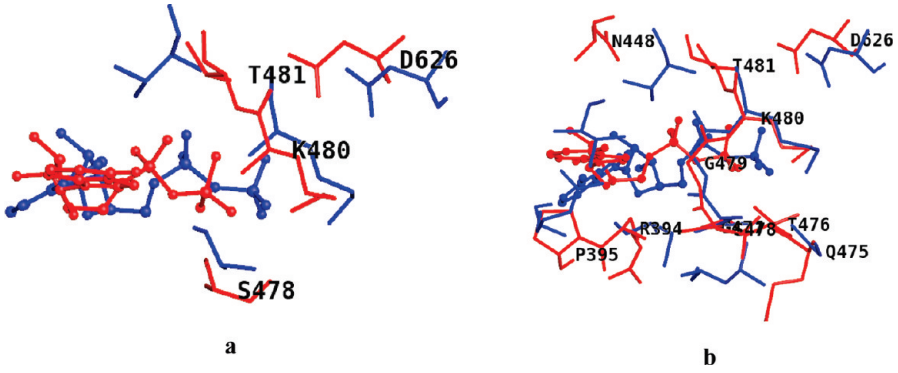
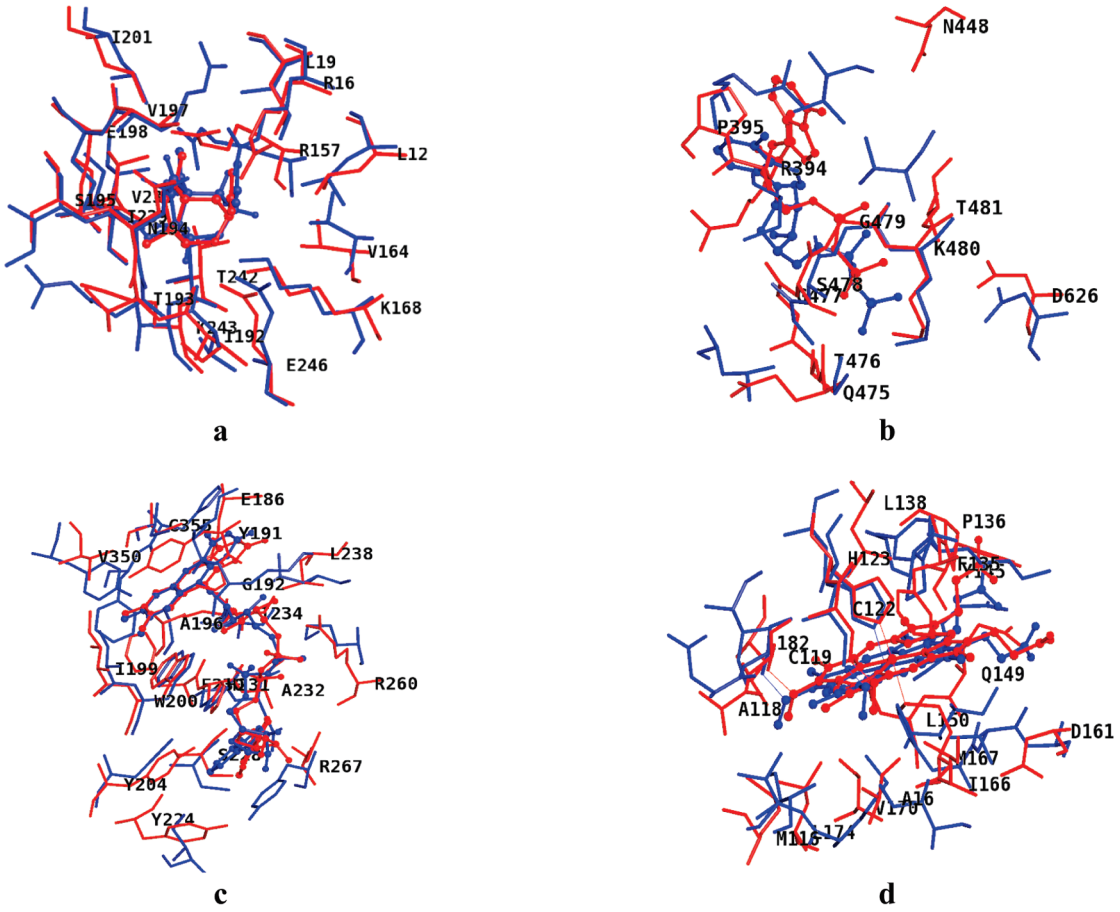


Figure 5. PA alignment of 3KAR_ADP (red) and 1UKZ_ADP (blue). The seed alignment initially obtained is shown (a), whereas final alignment of the same pair of sites is shown in (b).

Four different schemes are chosen for the alignment as detailed earlier in order to consider different scenarios leading to similarity between sites. The first scheme considers all the backbone atoms as well as the direction of the side chain and hence is a rigorous metric, useful for identifying the best alignment between highly similar sites. The second scheme considers only the backbone atoms and is useful for aligning sites that generally emerge from the same folds but with substitutions in residues at the site. The third and the fourth schemes consider only the side chain positions; the fourth considering only the centroid of the side chain, while the third considering both the centroid and the C_α position. These are useful for identifying and aligning binding sites that bear similarity purely by the position of their side chains and do not have any similarities in their main chains. Typically such pairs would be from different structural folds or even different classes and are difficult to align. They are nevertheless important since they may have similar recognition capabilities and hence similarities in function as well. Centroids for side chains are more useful than considering all atoms of the side chain, given the higher flexibility (and perhaps less accuracy) in the side chain atoms as compared to the main chain.

Performance Evaluation and Benchmarking. The performance of the algorithm, the need for considering different alignment schemes, and the usefulness of consequential alignments are illustrated with examples: (a) PA evaluates alignments for each of the four schemes. Each scheme represents a different kind of superposition of a pair of sites. As an example, a pair of ADP binding sites is examined in two proteins PDB identifiers,



Example	RLS	ALS	RRS	ARS	Seed-CR	SeedRMSD	Alignment scheme	#CR	#Atom	#RMSD
6A	18	147	19	159	14	0.66	2	18	130	0.84
6B	16	130	17	120	3	0.39	1	11	52	0.95
6C	26	211	27	247	7	1.44	1	18	95	1.28
6D	36	302	29	212	10	2.29	3	22	93	1.62

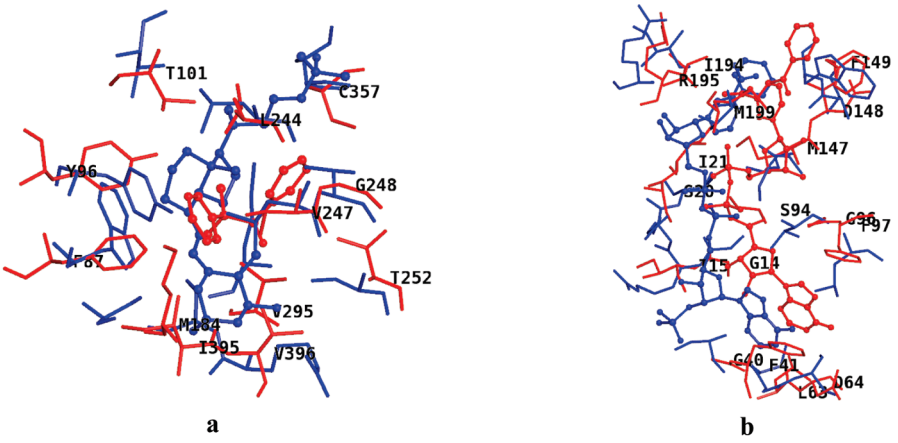
Figure 6. Alignment of pairs of binding sites from proteins having differences at the level of organism or fold or chemical nature of the ligand. Ligand is rendered in ball-and-stick model, whereas site residues are in stick model using PYMOL. (a) Chorismate mutases from *S. cerevisiae* 4CSM (red) and *E. coli* 1ECM (blue). (b) ADP binding sites of a pair of proteins differing at the family level (SCOP identifiers shown within brackets following PDB identifier) motor domain of kinesin like protein, 3KAR (c.37.1.9) and uridylate kinase, 1UKZ (SCOP c.37.1.1). (c) Binding sites of flavin adenine dinucleotide (FAD) in flavoenzyme Ero1p 1RP4 (red) a.227.1.1 and Erv2p 1JRA (blue) a.24.15.1 and (d) heme (HEC) binding sites of cytochromes in *P. stutzeri* (1M6Z) and *R. viridis* (1DXR). The table shows PA scores for each of the pairs of sites a−d. The header columns, RLS and RRS, indicate number of residues in first (L denoting left) and second (R denoting right) sites; ALS and ARS indicated number atoms in the two sites; Seed-CR indicates number of aligned residues found by seed alignment; Seed-RMSD denotes RMSD of the aligned points after seed alignment; Alignment-scheme is indicated (1−4); #CR denotes total number of aligned residues after evaluating consequential pairings; and #Atom, total number of aligned atoms, and #RMSD denotes the RMSD of aligned atoms after final alignment.

3KAR (motor domain of kinesin-like protein) and 1UKZ (uridylate kinase) as illustrated in Figure 4. Each of the schemes identifies slightly different correspondences, of which, schemes 1 and 2 produced the best superposition of the sites. The number of residues and atoms in the two binding sites is 16, 130, and 17 120, respectively.

The number of common residues in seed alignment, seed RMSD in Å, total number of residues aligned, number of atoms aligned, and their RMSD are given: (a) (3, 0.39, 11, 52, 0.95); (b) (6, 1.1, 12, 52, 0.96); (c) (4, 2.0, 3, 11, 46, 1.63); and (d) (3, 2.18, 9, 33, 1.71) clearly showing that the first scheme to be the best of all, for this case.

(b) PA proceeds in two stages: generation of seed alignment followed by determination of consequential residue pairings. Typically the seed alignment is much smaller than the overall residue pairings generated by determining spatial proximities of the residues between the sites. Improvement in performance by considering consequential alignments in PA is illustrated with an example, as shown in Figure 5. The size of the seed alignment is 3 residues which aligns with an RMSD of 0.38 Å of aligned points.

dx.doi.org/10.1021/ci200132z |*J. Chem. Inf. Model.* 2011, 51, 1725–1736



Example	RLS	ALS	RRS	ARS	Seed-CR	Seed RMSD	Alignment-scheme	#CR	#Atom	# RMSD
7A	15	119	28	246	6	2.21	2	11	50	1.75
7B	35	272	33	255	6	2.27	3	17	71	1.75

Figure 7. (a) Alignment of binding sites of cytochrome P450 (1PHG) and vitamin D nuclear receptor (1DB1). The ligands present in the two structures are shown in ball-and-stick representation: red: 1PHG_MYT and blue: 1DB1_VDX. (b) Binding sites of enoyl reductase 1ZID (red) and dihydrofolate reductase 2CIG (blue). Tabulation indicates PA scores for both the pairs of sites. The definition of column headers is same as used for the tabulation in Figure 6.

However, it resulted in a consequential alignment of 11 residues with 52 atoms and an RMSD of 0.95 Å, by the alignment scheme 1. The example also serves to justify the approach of single scanning of the SeedList without backtracking, as any new residue pairings are seen to be identified at the consequential pairings stage.

(c) PA has been tested on several examples: (i) First it is tested for its ability to pick up similarities between a pair of binding sites of the same protein from different species, since such pairs serve as examples of highly likely similarities and hence as positive controls of the study. Typically these will have the same SCOP[31] identifiers at all four levels. (ii) Second, binding sites for the same ligand but with different SCOP identifiers at the family or the superfamily levels are considered. (iii) Third, known examples of binding sites for the same ligand that are from completely different folds, but have been previously reported to be similar, are considered. (iv) Fourth, the algorithm is further validated by considering benchmarking datasets reported by authors of other site alignment algorithms. In some of the examples in this category, the SCOP identifiers in the pair differ at all four levels. As an example of the expected similarities, binding sites of chorismate mutases from *Saccharomyces cerevisiae* and *Escherichia coli* are taken. PA aligns them well and outputs all the expected correspondences between residues in the two sites (Figure 6a). The PA scores are shown in the table inside Figure 6.

An example of a pair of binding sites of the second category is a pair of ADP binding sites from motor domain of kinesin-like protein, 3KAR (SCOP c.37.1.9), and uridylate kinase, 1UKZ (SCOP c.37.1.1), differing in the family level of their SCOP identifiers. The two sites align well as shown in Figure 6b.

An example of a pair of binding sites in the third category is a pair of flavin adenine dinucleotide (FAD) binding sites in flavoenzyme Ero1p 1RP4 with SCOP code a.227.1.1 and Erv2p 1JRA with SCOP code a.24.15.1; the reflecting overall fold level dissimilarity between the proteins is shown in Figure 6c.

Alignment of a pair of heme (HEC) binding sites of cytochromes in *Pseudomonas stutzeri* (1M6Z) and *Rhacophorus viridis* (1DXR) belonging to different folds is shown in Figure 6d. The SCOP identifiers are a.3.1.4 and a.138.1.2, respectively. The PA score details are shown inside tabulation shown inside Figure 6.

As an extreme example of detecting site level similarities from drug action point of view, metyrapone was chosen from the report of Minai and co-workers.[27] The drug is known to act on cytochrome P450 (1PHG) and is predicted to bind to vitamin D nuclear receptor (1DB1) by the authors. The overall folds of the proteins are different. PA is able to align the two sites well as shown in Figure 7A, where the alignment showed conservation of important residues.

Another example of drug cross reactivity has been detected for the front-line drug, isoniazid, used in the treatment of *Mycobacterium tuberculosis*. It is pro-drug, and after activation, is well-known to target the inhA-encoded enoyl-ACP reductase (1ZID) inhibiting mycolic acid biosynthesis in the pathogen. Recently, a surprising observation of binding of isoniazid adduct to *M. tuberculosis* dihydrofolate reductase has been reported[32] (PDB: 2CIG). The alignment of pair of binding sites is illustrated in Figure 7b. PA scores for both the pairs of sites are indicated in the tabulation inside Figure 7.

Most recent methods in the literature are chosen, and datasets used in their validation have been tested here as well. A set of 43 pairs of binding sites used for validation in various algorithms[9,12,15−19,28] has been chosen. The corresponding proteins of the binding sites belong to diverse fold families. The results of comparison are tabulated in Table 1. The pairs of sites are available for interactive visualization at http://proline.physics.iisc.ernet.in/pocketalign/, enabling selection of dataset based on different types of sites and ligands.

Advantages of PA over many other existing methods are that: (i) Alignments are explored with four different schemes catering to different ways of considering an amino acid residue. Each alignment is scored and ranked, and the best alignment of all the four schemes is then programmatically identified. (ii) Alignment is carried out in two phases: first to obtain a seed alignment and then to explore consequential alignments after that. Consideration of these two aspects is seen to significantly improve the quality of alignment both in terms of best superposition as well as the number of atoms and residues aligned. As a result, the

Table 1. Table Showing Examples of 53 Pairs of Sites Chosen for Testing PA

PDB1	LIG1	PDB2	LIG2	#R1	#A1	#R2	#A2	Seed Alignment			Total Alignment			SCOP	Superposition quality		Dataset
								#CR	RMSD	T	#CR	#A	RMSD		PA	Remark	
1DHJ	MTX	4DFR	MTX	19	156	20	165	18	0.2	1	19	153	0.35	Same	G		Yeturu and Chandra, 2008 (PocketMatch)
1A4G	ZMR	1NSC	SIA	20	190	21	202	19	0.21	1	19	182	0.23	Same	G		
1SDU	MK1	1SDT	MK1	32	229	32	229	26	0.36	1	32	229	0.44	Same	G		
1B42	SAH	2VP3	SAH	25	192	24	182	24	0.19	1	24	182	0.19	Same	G		
1GJC	I30	1V2Q	ANH	5	52	5	52	5	0.1	1	5	52	0.1	Same	G		
1KV5	PGA	2JGQ	PO4	16	107	14	91	8	0.33	1	12	75	0.42	Same	G		
1BZC	TPI	1GFY	COL	17	137	16	124	13	0.36	1	15	111	0.45	Same	G		
1DJX	I3P	1DJY	I2P	16	138	16	143	15	0.18	1	15	130	0.24	Same	G		
1ZID	ZID	2CIG	IDG	35	272	33	255	6	2.27	3	17	71	1.75	Same	G		
1V07	HEM	1HBI	HEM	22	200	25	223	9	1.13	1	19	146	1.27	Family	G		
1FWK	ADP	2iYV	ADP	15	102	14	103	3	2.33	3	8	31	1.59	Different	G		
1IN4	ADP	1XJQ	ADP	17	138	16	131	3	2.4	4	9	38	1.65	Fold	G		
9LDT	NAD-5	1SOW	NAD-4	34	245	35	257	16	1.07	1	33	205	0.93	-	G	88/(132,138); 0.53	Gold and Jackson, 2006 (SitesBase)
9LDT	NAD-7	1EE2	NAD-7	34	245	35	253	10	1.77	1	27	132	1.16	Family	G	Low Sequence identity 27%;	
9LDT	NAD-5	1MOK	FAD-4	35	257	44	329	5	2.06	1	18	80	1.35	Fold	G	40/(132,170); 0.60 Nicotinamide and Flavin moieties	
9LDT	NAD-7	3MCT	SAH-1	34	246	24	187	4	1.54	1	13	67	1.41	Different	G	False Positive	
1RM8	BAT-5	1BKC	INN-8	16	132	19	150	8	0.47	1	15	98	0.76	Family	G	53/75; 0.53; H246/H132; 250/136; E247/133	
1RM8	BAT-5	1G2A	BB2-5	16	133	21	160	5	1.49	1	13	74	1.1	Fold	G	Not considered false positive; 34/(95,75) Similar architecture; Ligand aligns closely	
1M6Z	HEC-2	1DXR	HEC-9	36	302	29	212	7	2.21	1	18	86	1.67	Fold	G		
1M6Z	HEC-2	1E2Z	HEC-1	34	296	32	246	7	2.1	1	18	69	1.47	Different	M		
1M6Z	HEC-5	1LGA	HEM-7	33	279	30	239	2	0.44	1	19	65	1.43	Different	M	Heme ligands align; Histidines on opposite side of heme	
1RP4	FAD	1JRA	FAD	26	211	27	247	7	1.43	1	18	95	1.28	Fold	G	49/(122); 0.52; Align well W200/W19; H231/H86; C355/C57	
4CSM	TSA	1ECM	TSA	18	147	19	159	14	0.66	1	18	130	0.83	Family	G		Schimitt et al., 2006 (CavBase)
3KAR	ADP	1UKZ	ADP	16	130	17	120	3	0.39	1	11	52	0.95	Family	G	Phosphate portions of ATP	
1CDK	ANP	1AOG	FAD	28	220	49	343	3	1.77	1	18	68	1.71	Fold	M		
1CYD	NAP-4	2ADM	SAM	36	252	23	181	2	2.4	1	14	52	1.7	Fold	L	Adenine region	
1CYD	NAP-2	1FOH	FAD-1	36	251	39	281	5	2.28	1	18	73	1.48	Fold	G		
1CYD	NAP-4	1FDS	EST-1	35	245	18	144	4	2.4	3	11	49	1.84	Same	M	Co-factor region of carbonyl reductase matching with unoccupied region of steroid site	
1AYL	ATP	1ELQ	PLP	23	179	18	146	5	1.73	3	12	42	1.43	Fold	M		Kinoshita & Nakumara,2003
1C0P	FAD	1HYU	FAD	47	340	43	303	9	1.48	1	36	162	1.38	Fold	G		Minai et al.,2008
1F03	HEM	1MUP	TZL	24	239	13	111	4	2.31	2	12	50	1.57	Different	M		
1LM8	HYP	1VK5	EDO	12	120	10	90	4	2.11	1	8	48	1.5	Different	G		
1UR2	XYP	2NLR	BGC	6	70	6	69	4	1.31	4	4	22	1.48	Different	G		
1EHI	ADP	1G3U	TMP	22	185	17	167	4	2.31	1	10	45	1.85	Fold	M		
1A52	EST	1FDT	EST	18	146	21	173	5	2.2	3	12	54	1.73	Different	G		
1J37	MCO	1UTZ	PF3	14	134	22	174	2	2.11	3	7	34	1.56	Family	G		
1CQP	803	1DKF	OLA	18	160	24	194	5	1.9	3	12	61	1.7	Different	G		
1PHG	MYT	1DB1	VDX	15	119	28	246	5	2.48	1	12	48	1.72	Fold	G		
1PTH	SAL	1DFL	ITU	11	86	15	74	3	2.24	2	7	20	1.64	Fold	M		
4DFR	MTX	1TLL	NAP	33	250	30	230	5	2.48	1	16	65	1.76	Different	M	ique score : 43 and RMSD 2.	Najmonovich et al. 2008 (Isocleft)
4DFR	MTX	1GY3	ATP	33	250	21	164	4	2.5	3	16	58	1.74	Different	L	ique score : 33 and RMSD 1.	
1A4G	ZMR	1JG4	SAM	29	253	31	217	4	2.09	1	16	64	1.69	Fold	L	ique score : 38 and RMSD 2.	
1A4G	ZMR	1JDV	AND	29	253	13	111	4	2.35	1	10	45	1.7	Different	L	ique score : 34 and RMSD 2.	
1ADD	ZN	1BMC	ZN	6	52	6	52	4	1.74	3	5	26	1.48	Different	G		Konc and Janezic 2010 (ProBiS)
1ALK	ZN	1AMP	ZN	5	42	5	42	5	1.13	4	5	36	1.3	Different	G		
1ECE	BGC	2DNJ	DNA	6	53	6	53	2	1.8	1	6	40	1.6	Different	G		
1POW	FAD	1INP	MG	3	25	3	33	3	0.46	3	3	21	1.04	Different	G		
1RIB	FEO	1VHH	ZN	6	53	6	63	5	1.6	4	6	31	1.53	Different	G		
1ALK	ZN	1FJM	MN	6	55	6	55	2	1.42	1	3	26	1.53	Different	L		
1AYL	ANP	1BMF	ANP	5	53	5	41	4	1.76	1	5	31	1.38	Different	G		
1PHR	SO4	1VHR	EPE	4	31	4	31	4	0.7	2	4	30	0.89	Different	G		
1EUR	N/A	1QBA	N/A	6	47	6	97	4	0.18	2	6	44	0.47	Different	G		
2KAU	NI	2MHR	FEO	6	58	6	107	4	1.3	1	5	28	1.32	Different	M		

LEGEND

#R1 = Number of residues in Site-1 **#A1 =** Number of atoms in Site-1 **Similarly #R2, #A2 in Site-2**

#CR = Number of residues aligned **RMSD =** Root mean squared deviation **T =** The alignment scheme of top ranked mapping

SCOP = Different level of similarities in protein structure **Remark =** Remarks on same pair by other methods

PA = Pocket Align alignment quality upon manual inspection is given **G:**Good, **M:**Medium, **L:**Low

performance of PA is superior in many cases as compared to many of the other tools tested here as demonstrated in Table 1.

SENSITIVITY TESTING

PA has been tested for sensitivity with respect to random perturbations of atomic coordinates of the same structure. An ensemble of 10 randomly perturbed structures at each bin of RMSD ranging from 1 Å to 5 Å in steps of 1 Å is generated amounting to a total of 50 perturbed structures. PA has been run for each of the perturbed structures against its unperturbed binding site. For this purpose an example of a NAD (ligand) as bound to lactate dehydrogenase (PDB: 9LDT) was taken. The site has 35 residues and a total of 256 atoms. The mathematical

dx.doi.org/10.1021/ci200132z |J. Chem. Inf. Model. 2011, 51, 1725–1736

Table 2. PA Scores for Pairs of Sites between Ligand Bound and Unbound Structures

Sl. no	PDB1 (complex)	PDB2 (unbound)	LIG	NumRes (complex)	NumAtom (complex)	CR	common atom	RMSD
1	1BID	3TMS	UMP	11	90	11	88	0.62
2	1CDO	8ADH	NAD	21	158	19	125	0.91
3	1DWD	1HXF	MID	18	142	17	136	0.5
4	1FBP	2FBP	AMP	14	131	14	129	0.97
5	1FBP	2FBP	F6P	11	116	10	99	0.91
6	1GCA	1GCG	GAL	14	165	14	165	0.37
7	1HYT	1NPC	BZS	9	76	8	72	0.54
8	1INC	1ESA	ICL	14	105	14	99	0.32
9	1RBP	1BRQ	RTL	12	103	10	76	0.62
10	1ROB	8RAT	C2P	7	62	7	62	0.54
11	1STP	1SWB	BTN	14	119	14	117	0.41
12	1ULB	1ULA	GUN	12	85	8	56	0.98
13	2IFB	1IFB	PLM	12	117	10	101	0.43
14	3PTB	3PTN	BEN	12	80	11	73	0.38
15	2YPI	1YPI	PGA	10	66	8	52	0.6
16	4DFR	5DFR	MTX	10	84	10	84	0.49
17	4PHV	3PHV	VAC	11	82	11	81	0.92
18	5CNA	2CTV	MMA	8	60	8	56	0.62
19	7CPA	5CPA	FVF	14	130	14	127	0.5
20	1A6W	1A6U	NIP	9	105	9	105	0.55
21	1ACJ	1QIF	THA	9	86	9	86	0.14
22	1APU	3APP	STA	8	59	7	49	0.87
23	1BLH	1DJB	FOS	11	76	11	72	0.2
24	1BYB	1BYA	GLC	8	68	7	62	0.62
25	1HFC	1CGE	HAP	17	177	15	131	0.52
26	1IDA	1HSI	QND	8	60	4	57	1.1
27	1IMB	1IME	LIP	12	88	12	88	0.25
28	1IVD	1NNA	ST1	10	138	10	136	0.79
29	1MRG	1AHC	AND	11	95	11	93	0.5
30	1MTW	2TGA	DX9	14	100	14	95	0.6
31	1OKM	4CA2	SAB	13	115	11	98	0.31
32	1PDZ	1PDY	PGA	8	66	6	47	0.5
33	1PHD	1PHC	HEM	24	181	23	170	0.18
34	1PSO	1PSN	STA	9	61	8	54	0.38
35	1QPE	3LCK	PP2	12	97	12	97	0.26
36	1QPE	3LCK	PTR	10	81	10	80	0.2
37	1RNE	1BBS	C60	23	166	20	140	0.57
38	1SNC	1STN	PTP	10	92	8	75	0.24
39	2CTC	2CTB	LOF	7	57	7	56	0.57
40	2H4N	2CBA	AZM	10	93	10	90	0.4
41	2PK4	1KRN	ACA	7	75	7	75	0.37
42	2SIM	2SIL	DAN	13	126	13	126	0.19
43	2TMN	1L3F	PHO	7	65	7	65	0.42
44	3GCH	1CHG	CIN	9	67	8	54	1.05
45	6RSA	7RAT	UVC	9	79	9	78	0.46

procedure behind random perturbation is outlined in Figure S1, Supporting Information. The Q^* scores are extracted from all the PA alignment scores of the perturbed structures and plotted against the RMSDs of the perturbation (Figure S2, Supporting Information). The plot demonstrates the expected behavior that, as the extent of perturbation in terms of RMSD increases, the number of matching atoms reflected in terms of the Q^* scores decreases.

Bound and Unbound Structures. The algorithm has further been evaluated for its ability to pick up similarities between sites in a set of 45 pairs of bound and unbound protein ligand complexes reported in Huang et.al.[33] The unbound structure is superposed onto the complex using local version of DALI[34] in order define ligand environment in the unbound structure. The amino acids of the superposed unbound structure around 4 Å of the ligand of the complex structure are considered for defining

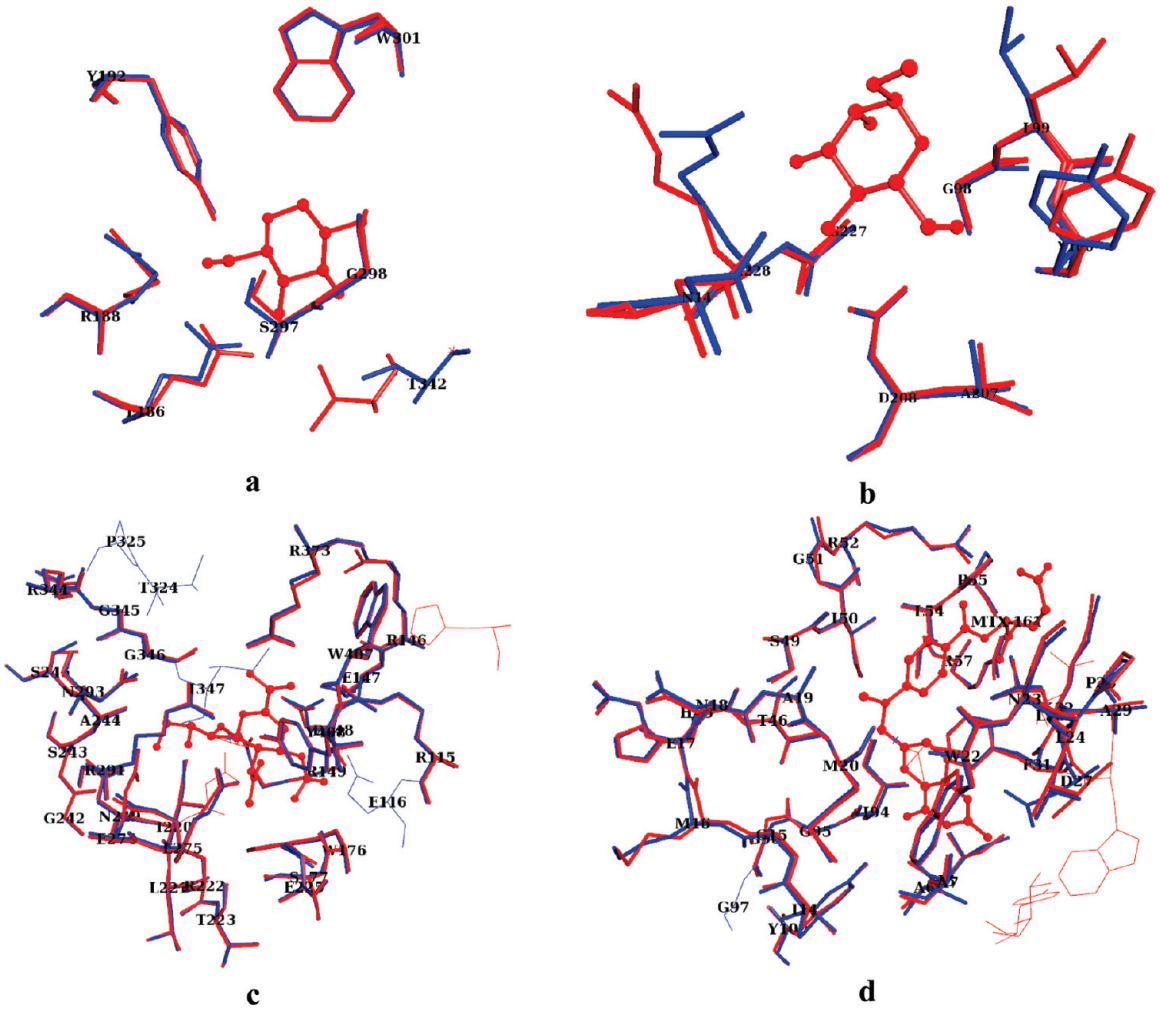


Figure 8. Illustration of pairs of sites belonging to apo and holo structures (a and b) and predicted pockets (c and d). (a) Glucose binding sites in 1BYA and 1BYB. (b) Mannose binding sites in 2CTV and 5CNA. Alignment of predicted sites for (c) sialic acid binding sites in 1A4G and 1NSC and (d) folic acid binding site in 1DHJ and 4DFR. The ligand is shown in ball-and-stick model in red color. The aligned residues are shown as thick sticks whereas unaligned residues in thin lines. Blue and red colored sites correspond to the left and right; PDB IDs, respectively, indicated in the legend.

binding site.The binding site residues in the original complex structure are also chosen similar around the bound ligand. The two binding sites are then aligned using PA, and the corresponding scores are reported in Table 2. The scores indicate that PA are robust enough to pick up similarities between sites of the bound and unbound forms. Examples of glucose binding sites in 1BYA and 1BYB (Figure 8a) and mannose binding sites in 2CTV and 5CNA (Figure 8b) are illustrated in Figure 8. In both cases, the PA is seen to perform well. The superposition results are made available on the Web site of PA (and also at http://proline.physics.iisc.ernet.in/pocketalign/apo-holo.html).

In addition, it has been verified that performance of PA does not depend much upon the method used for defining a binding site. This is confirmed by the comparison of pairs of predicted sites of sialic acid binding pockets of (1A4G and 1NSC) (Figure 8c) and separately the folic acid binding pockets of (1DHJ and 4DFR) (Figure 8D). The alignment of predicted sites in 1A4G and 1NSC resulted in an alignment of 27 residues (237 atoms) with an RMSD of 0.2 Å as opposed to the crystallographic counterparts of the sites with 19 residues (182 atoms) with an RMSD of 0.23 Å. For the pair, 1DHJ and 4DFR, the alignment of predicted sites resulted in correspondences for 32 residues (243 atoms) with an RMSD of 0.3, while the crystallographic counterparts of the sites with 19 residues (153 atoms) resulted in an RMSD of 0.35. The robustness of PA in handling variations in pocket definitions is thus demonstrated.

CONCLUSIONS

A new algorithm has been developed and validated for obtaining accurate alignments of binding sites in protein structures. The algorithm considers whole residues as input for comparison and shows significant improvement over others in terms of coverage of atoms. The concept of geometric perspectives for capturing the geometry is seen to be useful. The perspectives serve as a new representation for abstracting a binding site and a methodology for enumeration of possible alignments between a pair of binding sites are presented. A new heuristic of scanning a sorted list of pairings based on a score that takes into account both the chemical nature and the shape of the site. This also avoids the costly backtracking algorithm. Four different scoring schemes to cover different types of possibilities of atomic overlap in a pair of sites are used.

dx.doi.org/10.1021/ci200132z |J. Chem. Inf. Model. 2011, 51, 1725–1736

Finally, useful Pymol scripts are generated for easy visualization. In addition to being useful for gaining functional insights, analysis of binding sites is also expected to be of significant use for structure-based drug design.

■ ASSOCIATED CONTENT

Supporting Information. Definition of various terms used in PA description is available in Table S1. Parameters defined for superposition of sites in available in Table S2. More information on sensitivity analysis is available in Figures S1 and S2. This information is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: nchandra@serc.iisc.ernet.in. Telephone: +91-80-22932892.

■ ACKNOWLEDGMENT

Support for the Centre of Excellence in Bioinformatics by Department of biotechnology (DBT), Government of India and facilities at the Supercomputer Education and Research Centre of this institute are gratefully acknowledged. Support from the DBT computational genomics initiative is also acknowledged.

■ REFERENCES

(1) Thornton, J. M.; Todd, A. E.; Milburn, D.; Borkakoti, N.; Orengo, C. A. From structure to function: approaches and limitations. *Nat. Struct. Biol.* **2000**, *7* (Suppl), 991–994.

(2) Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Protein clefts in molecular recognition and function. *Protein Sci.* **1996**, *5*, 2438–2452.

(3) Russell, R. B.; Sasieni, P. D.; Sternberg, M. J. Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **1998**, *282*, 903–918.

(4) Dodson, G.; Wlodawer, A. Catalytic triads and their relatives. *Trends Biochem. Sci.* **1998**, *23*, 347–352.

(5) Carter, P.; Wells, J. A. Dissecting the catalytic triad of a serine protease. *Nature* **1988**, *332*, 564–568.

(6) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(7) Raman, K.; Yeturu, K.; Chandra, N. targetTB: a target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. *BMC Syst. Biol.* **2008**, *2*, 109.

(8) 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.

(9) Yeturu, K.; Chandra, N. PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinf.* **2008**, *9*, 543.

(10) Stockwell, G. R.; Thornton, J. M. Conformational diversity of ligands bound to proteins. *J. Mol. Biol.* **2006**, *356*, 928–944.

(11) Ramachandraiah, G.; Chandra, N. R. Sequence and structural determinants of mannose recognition. *Proteins* **2000**, *39*, 358–364.

(12) Stark, A.; Russell, R. B. Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res.* **2003**, *31*, 3341–3344.

(13) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387–406.

(14) Morris, R. J.; Najmanovich, R. J.; Kahraman, A.; Thornton, J. M. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics* **2005**, *21*, 2347–2355.

(15) Kleywegt, G. J. Recognition of spatial motifs in protein structures. *J. Mol. Biol.* **1999**, *285*, 1887–1897.

(16) Gold, N. D.; Jackson, R. M. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J. Mol. Biol.* **2006**, *355*, 1112–1124.

(17) Konc, J.; Janezic, D. Protein-protein binding-sites prediction by protein surface structure conservation. *J. Chem. Inf. Model.* **2007**, *47*, 940–944.

(18) Konc, J.; Janezic, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26*, 1160–1168.

(19) Konc, J.; Janezic, D., ProBiS: a web server for detection of structurally similar protein binding sites. *Nucleic Acids Res.* **2010**, *38*, (Web Server issue), W436-440.

(20) Najmanovich, R.; Kurbatova, N.; Thornton, J. Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics* **2008**, *24*, i105–11.

(21) Brakoulias, A.; Jackson, R. M. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins* **2004**, *56*, 250–260.

(22) Artymiuk, P. J.; Poirrette, A. R.; Grindley, H. M.; Rice, D. W.; Willett, P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.* **1994**, *243*, 327–344.

(23) Wallace, A. C.; Borkakoti, N.; Thornton, J. M. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **1997**, *6*, 2308–2323.

(24) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607–633.

(25) Kinoshita, K.; Nakamura, H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.* **2003**, *12*, 1589–1595.

(26) Gold, N. D.; Jackson, R. M., SitesBase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Res.* **2006**, *34*, (Database issue), D231-234.

(27) Minai, R.; Matsuo, Y.; Onuki, H.; Hirota, H. Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. *Proteins* **2008**, *72*, 367–381.

(28) Yeturu, K.; Utriainen, T.; Kemp, G. J.; Chandra, N. An automated framework for understanding structural variations in the binding grooves of MHC class II molecules. *BMC Bioinf.* **2010**, *11* (Suppl 1), S55.

(29) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A* **1976**, *5*, 1.

(30) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10915–10919.

(31) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536–540.

(32) Argyrou, A.; Vetting, M. W.; Aladegbami, B.; Blanchard, J. S. Mycobacterium tuberculosis dihydrofolate reductase is a target for isoniazid. *Nat. Struct. Mol. Biol.* **2006**, *13*, 408–413.

(33) Huang, B.; Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, *6*, 19.

(34) Holm, L.; Park, J. DaliLite workbench for protein structure comparison. *Bioinformatics* **2000**, *16*, 566–567.