JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Conserved Core Substructures in the Overlay of Protein−Ligand Complexes

Barry C. Finzel,*[,†] Ramprasad Akavaram,[†] Aravind Ragipindi,[†] Jeffrey R. Van Voorst,[†] Matthew Cahn,[‡] Malcolm E. Davis,[§] Matt E. Pokross,[§] Steven Sheriff,[§] and Eric T. Baldwin[§]

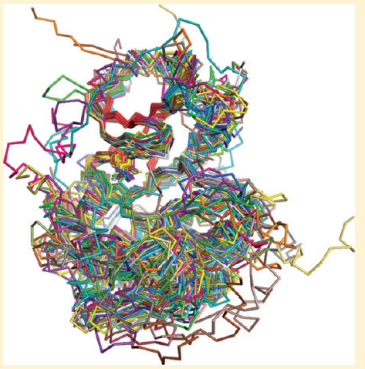
[†]Department of Medicinal Chemistry, University of Minnesota College of Pharmacy, Minneapolis, Minnesota 55455, United States

[‡]BioPharma Information Technologies, Bristol-Myers Squibb Company, Princeton, New Jersey 08543, United States

[§]Research & Development, Chemical and Protein Technologies, Molecular Sciences and Candidate Optimization, Bristol-Myers Squibb Company, Princeton, New Jersey 08543, United States

Supporting Information

ABSTRACT: The method of conserved core substructure matching (CSM) for the overlay of protein−ligand complexes is described. The method relies upon distance geometry to align structurally similar substructures without regard to sequence similarity onto substructures from a reference protein empirically selected to include key determinants of binding site location and geometry. The error in ligand position is reduced in reoriented ensembles generated with CSM when compared to other overlay methods. Since CSM can only succeed when the selected core substructure is geometrically conserved, misalignments only rarely occur. The method may be applied to reliably overlay large numbers of protein−ligand complexes in a way that optimizes ligand position at a specific binding site or subsite or to align structures from large and diverse protein families where the conserved binding site is localized to only a small portion of either protein. Core substructures may be complex and must be chosen with care. We have created a database of empirically selected core substructures to demonstrate the utility of CSM alignment of ligand binding sites in important drug targets. A Web-based interface can be used to apply CSM to align large collections of protein−ligand complexes for use in drug design using these substructures or to evaluate the use of alternative core substructures that may then be shared with the larger user community. Examples show the benefit of CSM in the practice of structure-based drug design.



INTRODUCTION

Structure-based drug discovery often benefits from the comparative visualization of the three-dimensional structures of multiple protein−ligand complexes. In fragment-based discovery,[1,2] larger more potent chemical entities may be designed by judicious combination of two or more smaller molecules shown experimentally to occupy adjacent or overlapping protein subsites. The larger molecules derive increased affinity by preserving key interactions made by each smaller molecule, and the accurate overlay of different experimental complexes is needed to inform the design of a chemical linker that joins the two fragments with unstrained geometry. In scaffold hopping, multiple complexes with diverse inhibitors that occupy the same binding pocket can be overlaid to inspire potent new chemical scaffolds that incorporate desirable features or functional groups differently.[3] Inhibitors optimized against one protein target have been increasingly used as leads for the development of more selective or specific agents against a different but related target, a process called target hopping. This approach has been widely used in the field of kinase inhibition.[4,5]

The benefits of structure comparison may be realized across a variety of activities ranging from detailed and comprehensive computational analyses of structure ensembles[6] to a casual discussion between scientists in front of a laptop. Experience has shown that good ideas are likely to come from any project contributor, so it is important that structural data be accessible to all. Since structural models can arise from a variety of different sources, including X-ray crystallography, NMR spectroscopy, or molecular modeling studies, the models are not readily comparable without first placing the atomic coordinates on a common positional frame of reference. A plethora of different structure superposition tools exist, but those that are both broadly accessible and convenient for use by collaborators of all backgrounds are not necessarily those that produce the most useful superposition for drug design. Moreover, a casual survey of peers exposes a general lack of understanding of the strengths and limitations of different methods; structure alignment procedures are as likely to be chosen for their convenience as for their applicability to the superposition problem at hand. The quality of a superposition is important, however. When joining two fragments with known binding modes, only a small relative shift in one fragment may alter design decisions over how many carbons a synthetic linker should have or the selection of preferred atomic hybridization. Very small shifts can affect predictions of hydrogen-bond strength.

To support a dynamic collaborative environment conducive to structure-based design, we have found it practical to create and

Received: December 6, 2010
Published: July 08, 2011

maintain libraries of prealigned structures relevant within the scope of each project that are accessible to all drug discovery team members. Several important attributes characterize the aligned structures that comprise these libraries: (1) Aligned structures optimally emphasize structural similarities and differences near the ligand binding site. (2) The same positional reference frame should be used throughout the lifespan of the project. Since new data will likely become available as each project matures, it should be possible to add new structures to the library without repositioning all existing structures. (3) It must be possible to overlay nonliganded apo structures, proprietary, and unpublished structures (those not deposited in public structure databases) and any structures with localized binding site structural homology, regardless of sequence differences in the site or elsewhere. (4) The ability to support the overlay of sites that lies at the interface between different polypeptide chains. Since relevant structures can number in the thousands for some projects, an efficient and reliable means to construct this library is essential. The methods and software described here fulfill each of these required characteristics and provide a means to accomplish high-throughput structure alignment.

The problem of how to best overlay multiple protein structures has been approached in a variety of different ways. Numerous methods have been described for pairwise structure alignment, in which one structure is optimally aligned on another, and multiple structure alignment, where collections of structures are aligned on a single reference or consensus structure. Procedures also have been created to identify and align binding sites. We consider each approach, in turn, but make no attempt to provide a comprehensive review.

Pairwise superposition algorithms build upon the seminal procedure of Kabsch[7] for the optimal superposition of two sets of vectors. Because this approach is very sensitive to the influence of a few large outliers, widely used algorithms such as those embedded in popular molecular graphics software (PyMOL, http://www.pymol.org/; Maestro, http://www.schrodinger.com/) use pairwise sequence alignment to first identify individual residues to be included in a trial superposition and then invoke an iterative procedure to successively remove outliers until convergence is achieved. Since structural similarity may exist despite sequence divergence, these sequence-based approaches can sometimes fail altogether. More complex algorithms exploit either multisequence alignment,[8,9] secondary structure matching,[9,10] or substructure matching[11] to enable the overlay of structures with more sequence diversity. Some procedures strive to gauge both localized and global alignment quality,[12] and many of these algorithms include provisions for the user specification of a domain or substructure to provide a more localized alignment, but default parameterizations regard the solution that emphasizes the largest structural similarity as the 'best' alignment.

Multiple structure alignment methods[10,13−20] expand alignment capabilities to systems where pairwise sequence comparisons may fail. Methods differ in the way that conserved structural features common to all structures being aligned are identified. Each structure is aligned onto a single reference structure or a consensus mean structure, and heuristics are applied to calculate the best rigid-body transformation for each structure. Most procedures strive to identify key conserved amino acid residues automatically using multiple sequence alignment,[13,14,19,20] secondary structure,[10,16] or similarity in distance geometry.[15,17,18] Since these methods were created to specifically identify similarity in structures that would not be obvious from sequence comparison alone, these algorithms are parameterized to be inclusive. In a comparison, Hill and Reilly[21] show that none of the automated procedures for conserved substructure selection perform as well as human identification. Our own experience supports this conclusion. As we show below, global structure alignment does not provide the best localized alignment of binding sites or ligands. None of the sequence or secondary structure-based methods employ a strategy specifically aimed at best consensus alignment in the local vicinity of a ligand binding site.

The overlay of similar binding sites has also received considerable attention, and multiple databases relating similar binding sites exist.[22−25] Site alignment tools generally represent localized atomic structure as a collection of atoms, residues, or chemical attributes in three-dimensional space and store these attributes in a relational database. Computational procedures are then used to identify sites from other structures with similar attributes. In CavBase,[25] for example, "cavities" are represented by a three-dimensional distribution of "pseudocenters" that convey the geometric distribution of chemical properties. Each site is stored in relation to bound ligands, and similar sites (those with similar pseudocenter distributions) are catalogued for rapid access. Such methods universally benefit from the fact that the alignment is always localized around the binding site. Site-based alignment tools can be used to identify similar sites in otherwise dissimilar proteins and have potential applicability in the identification of possible bioisosteres for drug design.[24,26]

Nevertheless, site databases have limitations. Such databases require constant curation and can be out-of-date or accessible only by subscription. In many cases, a site is not stored in the database if it does not have a ligand bound in it, so unoccupied sites cannot be aligned. Appropriate site descriptors are chosen based on their proximity to the ligand, but pockets may be truncated in cases where the bound ligand occupies only a subsite. Moreover, subsite query specifications can be complex. While site-based methods allow for the alignment of sites that might not be aligned at all using other approaches, the quality of alignment may not be better than that which can be achieved with local substructure alignment, particularly in cases where structural homology is known to exist.

After reviewing available structure alignment methods, we found no existing tool specifically suited our needs for library preparation. Methods that seek to find the best alignment of whole structures instead of localized sites or subdomains will fail to produce the best binding site overlay if structural diversity exists outside that site. Methods that superimpose structures on a consensus of all structures will not retain the same orientation as newer structures are added. In our hands, site-based alignments too often fail to find similarity where similar sites clearly exist.

Instead, we have elected to heed the analysis of Hill and Reilly[21] and restrict superimposed atoms to a small number of conformationally invariant residues selected manually. This affords the opportunity to empirically tailor the alignment strategy to a particular binding site or subsite. Others have taken a similar approach.[6] By making sure the residue selection includes only highly conserved structural features that help to shape the character of the site, the same substructural core may be used to align not only complexes with the same protein and different ligands but also distantly related family members. Since an empirical examination is required, a selection of diverse family members is first overlaid by some other method. A careful study

of family members may be required to identify the appropriate residues representing the core substructure. While several databases exist to annotate[27] or identify and quantify similarity in binding sites[28−30] that may be useful as aids in the selection of conserved binding site residues, the choice of structural features to align is ultimately empirical.

Even after an appropriate substructure selection is made, automating the alignment of relevant structures on a large scale is complicated. To align specific substructures from two proteins with existing software, the specific atoms or residues to be aligned must be explicitly named, a daunting task when possibly thousands of related complexes are involved. Conserved core substructures may be too small to be unambiguously identified by sequence alignment. Arbitrary residue numbering schemes used in protein structures often make it difficult to identify the corresponding residues when different proteins are involved. Any practical application of local substructure alignment on a large scale requires new software.

An approach and software for overlaying proteins based on core substructure matching (CSM) is described here. The procedure uses distance geometry matching to locate an "exact" match to an empirically selected core substructure identified in a reference structure to obviate the need for manual residue selection and to avoid a dependence on potentially unreliable sequence alignment. A Web-based interface to computational services is provided that can be used to generate ensembles of structures to assess the applicability of different core substructures for alignment of a given ensemble of structures. Since a core substructure is specified only as a set of distinct residues from a single reference structure, a database of empirically selected substructures has been created that comprise a collection of overlay methods useful to reproducibly align any homologue structures also containing a similar conserved core. It will be shown that CSM accomplishes protein−ligand complex superposition with smaller ligand position variance than other commonly used software for protein structure alignment and that the method can be used to align entire families of protein−ligand complexes automatically without loss of accuracy.

■ ALGORITHMS

CSM. CSM is implemented using distance geometry and the flexible search algorithm of Finzel.[31,32] This distance geometry implementation is preferred over others because it can be used to locate complex substructures comprised of multiple polypeptide segments, enforcing both the geometry of each segment and the positional relationship between disconnected segments. Residues from a single reference structure that comprise the conserved core are reduced to a list of distances between all α-carbon atoms (the "target" distance geometry). All inter-α-carbon distances representing the structure to be overlaid (the "subject" distance geometry) are also computed. The algorithm efficiently searches through the subject distance geometry for values that match the target within a tolerance. A separate tolerance is applied to distance geometry contributions arising from within bonded segments and those relating different segments. Since matching distance geometry implies matching substructure geometry,[33] the algorithm identifies a 1:1 correspondence between residues in the target and subject structures without regard to similarity in the respective amino acid sequences.

The CSM uniquely identifies portions of the two structures that will contribute to optimal alignment. When the target substructure is sufficiently complex, only one match per ligand binding site will occur. The software allows for the specification of amino acid sequence constraints on returned hits, but these are seldom needed; the core geometry alone is sufficient to create a single unique alignment. When the procedure fails to find a match, no poor alignment of structures results. This is an important benefit of our approach. A simple count of structures with matches can reveal how widely a selected core is conserved.

This resulting 1:1 alignment of structurally conserved residues is then used to generate the best transformation that brings corresponding subject atoms into superposition on the target. While any atoms common to both target and subject substructures might be included, it is typical to use only backbone atoms and β-carbons. The algorithm will return more than one hit when subject structures contain multiple binding sites in the crystallographic asymmetric unit. Matching alignments are then ranked by root-mean-square difference (RMSD) of the aligned atoms. The deviations that may sometimes be seen in ligand binding of each of these might be considered a valuable experimental observation that requires study. The procedure has been implemented using the Python v2.7 programming language and makes use of open source BioPython.PDB modules for coordinate file parsing and manipulation.[34]

Database Implementations. We utilize relational databases to manage a collection of empirically derived core substructures for use with the CSM method to align structures. CSM has been integrated into two different platforms described in more detail below, but both capitalize on the relatively simple nature of a core substructure that enables it to be used in conjunction with CSM software as an "overlay method". The databases store the atomic coordinates of a reference structure, a selection of amino acid residues comprising the core, the tolerances to be applied in distance geometry, and the allowed amino acid residue constraints (if any). CSM can retrieve this information from the database, apply the method to align any provided subject PDB file without any other input, and produce aligned subject atomic coordinates in the same PDB format as input.

■ RESULTS AND DISCUSSION

Evaluating Overlay Methods. An objective and quantitative evaluation of different protein overlay methods employed in drug design is challenging, because of the subjective and the context-dependent nature of the "right" solution. Generally, one wants an overlay method to always align the components of structure important to ligand binding. The most useful quantitative measure of alignment quality might be the RMSD in the positions of the ligand molecules following a protein-based superposition, but such a metric can apply only in the comparison of identical or highly similar ligands, where there is no doubt about the appropriate 1:1 correspondence of ligand atoms. Examples of sets of dissimilar proteins bound to identical or similar ligands are rare, however, and can seldom be used to inform the best overlay of targets for drug design. Even in a set of complexes involving very similar ligand analogs, very small differences in ligand chemical composition may produce disproportionate and unexpected changes in ligand binding that obscure smaller differences arising from different overlay methodologies.

One set of structures stands out as particularly well-suited for use in overlay method evaluation: structures of different kinase catalytic domains with the natural product staurosporine.[35]

Table 1. Kinase:Staurosporine Complex Basis Set

PDB ID	kinase[a]	resolution
1AQ1	CDK2	2.0
1BYG	CSK	2.4
1NVR	CHEK1	1.8
1NXK	MAPKAPK2	2.7
1OKY	PDK1	2.3
1Q3D	GSK3B	2.2
1QPD	LCK	2.0
1SM2	ITK	2.3
1U59	ZAP70	2.3
1WVY	DAPK1	2.8
1XBC	SYK	2.0
1XJD	PKCt	2.0
1YHS	PIM1	2.15
2BUJ	STK16	2.6
2CLQ	MK3K5	2.3
2DQ7	FYN	2.8
2GCD	TAO2	2.55
2HW7	MNK2	2.7
2NRY	IRAK4	2.15
2Z7R	RSK1	2.0
3BKB	FES	1.8
3CKX	MST3	2.7
3FME	MAP2K6	2.3
1STC	PKACa	2.3

[a] Kinase manning code.[36]

Staurosporine is broadly active as a pan-kinase inhibitor and has frequently been used to stabilize kinase catalytic domain conformation and facilitate crystallization, so crystal structures exist with a diverse set of kinase catalytic domains sampling all seven major branches of the kinome (Table 1).[36] The inhibitor is a conformationally constrained molecule (MW 466.5 Da) that has been found to occupy roughly the same binding position in at least 24 different kinase complexes in the Protein Data Bank (http://www.rcsb.org/pdb; accessed September 7, 2009). The kinase domains are representative of exactly the sort of family of related proteins that would be valuable to overlay for structure-based design.

To evaluate the effectiveness of different overlay methods, one of these structures was selected to serve as a reference structure (the cAMP-dependent protein kinase of PDB ID 1STC), and 23 other complexes were superimposed onto it using different procedures. Following each trial overlay, an RMSD was computed using all 35 non-hydrogen ligand atoms with respect to the corresponding atoms in the reference structure. Since the ligand itself was never included in the superposed atoms, ligand RMSD values provide an unbiased metric of overlay quality (Figure 1). To establish a baseline for the best achievable RMSD, each complex was also superimposed on the reference structure using only the ligand atoms (Figure 1A). The resulting RMSD (0.3 ± 0.1 Å) is consistent with the expected range of errors for structures in this resolution range (1.8−2.8 Å). In no other case was the ligand included in the overlay process.

Sequence-Based Overlay. The "super" function implemented in PyMOL[37] is representative of procedures that use pairwise sequence alignment to identify residues from the two protein structures to overlay. We find that our peers frequently rely on this or similar procedures because they are very easy to use, mostly automated, and are highly likely to succeed in common cases where significant sequence similarity exists between the subject proteins. One undesirable property of these algorithms is that they always result in a transformation, even if the outcome is poor or nonsensical. Another shortcoming is the lack of transparency regarding which atoms or residues were used to achieve each overlay. Different pairwise overlays may result in the inclusion of different residues. PyMOL was able to successfully overlay all 23 kinase structures onto the reference structure with an overall ligand RMSD of 1.1 ± 0.3 Å. Individual ligand RMSDs ranged from 0.5−1.8 Å. The resulting ensemble of overlaid structures is illustrated in Figure 1B.

Multiple-Structure-Based Alignment. MAMMOTH-Mult[14] was selected as a representative of more complex algorithms. It uses a progressive iterative multiple-structure alignment to identify a consensus conserved core, then aligns all proteins onto it, and is implemented with a Web interface (http://ub.cbm.uam.es/mammoth/mult/; accessed September 10, 2009). In applying this software to all 24 staurosporine complexes, the procedure identified a consensus core of 109 residues. RMSDs in Figure 1C were computed by comparing the positions of each reoriented ligand to the positions of the reoriented reference structure (1STC). While this algorithm was very capable with regard to consensus structure selection, it was difficult to employ in a practical design setting, because the resulting consensus structure adopts a different orientation whenever structures are added or subtracted from the set. Perhaps more importantly, this method did not produce a tighter visual ensemble of overlaid ligands when compared to PyMOL. The mean ligand RMSD for the MAMMOTH-Mult ensemble of structures is also not improved (mean 1.2 ± 0.3 Å).

Site Attribute Alignment. An evaluation of binding site attribute-based alignment methods was conducted using both SitesBase[22] and SiteEngine,[23] which employ completely different methods to enable the superposition of sites unrelated by protein sequence or secondary structure similarity. SitesBase[22] employs a procedure for determining the best alignment between conserved binding site atoms surrounding the site. Only sites with bound ligands are included in the SitesBase database. A convenient Web interface returns all other sites with similarity to the 1STC staurosporine site in a single query (http://www.modelling.leeds.ac.uk/sb/; accessed January 15, 2011). SitesBase performed very well compared to sequence- and multiple-structure-based alignment methods for those sites that are tabulated in the database (mean RMSD 0.6 ± 0.2 Å; data not shown). Unfortunately, it does not appear that the database has been updated since 2005; one-half of our staurosporine basis set could not be included in the analysis. The software also failed to identify similarity between staurosporine binding sites in 1STC and 1NRY, even though both structures are in the database with bound staurosporine.

SiteEngine[23] utilizes low-resolution surface matching to rapidly identify binding sites similar to a reference structure. Each of the 23 staurosporine complexes were submitted to the convenient Web interface (http://bioinfo3d.cs.tau.ac.il/SiteEngine/; accessed February 1, 2011) to be aligned with the 1STC reference complex. Resulting ligand RMSD values are presented in Figure 1D. The service performed nearly as well as SitesBase (mean RMSD 0.7 ± 0.3 Å), but in two specific cases (1NXK and 2CLQ), the site with highest similarity to 1STC was not the

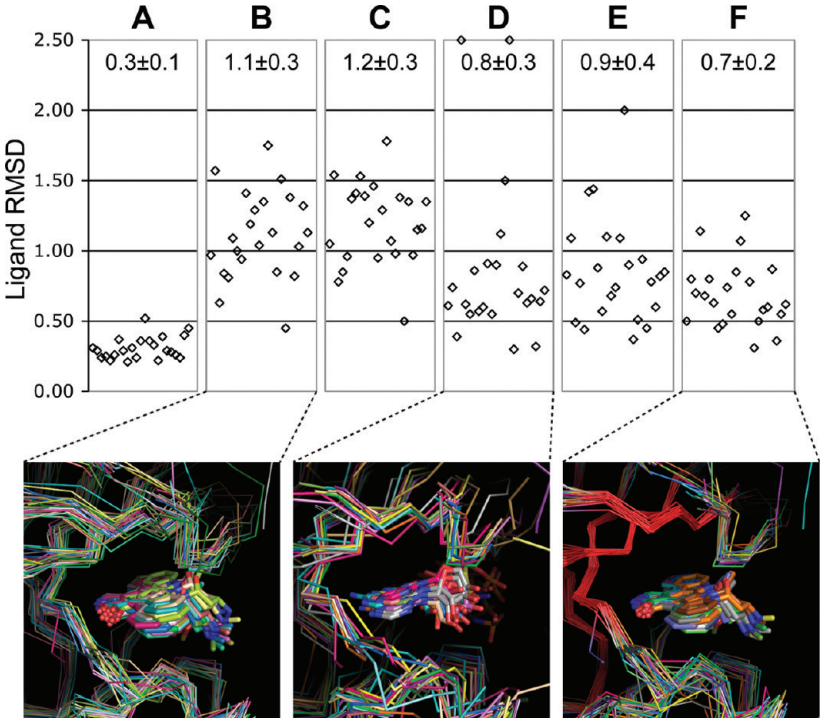


Figure 1. The RMSD in ligand positions compared to the same atoms in complex 1STC resulting from overlay of 24 kinase−staurosporine complexes using different methods. The mean of all ligand RMSDs for each method is across the top. (A) Superposition driven by direct overlay of only staurosporine atoms. (B) PyMOL overlay using "super PDB ID, 1STC" overlay.[37] Ensemble visualized at bottom. (C) MAMMOTH-Mult multiple structure alignment.[14] (D) SiteEngine overlay of binding sites.[23] RMSD values rendered at the upper limit of the graph (2.5 Å), but the actual RMSD for these ligands was much larger (24−26 Å). These are not included in the displayed mean and standard error. Ensemble visualized at bottom. (E) CSM-based overlay using 3-segment "hinge" conserved substructure core of 10 residues (1STC catalytic domain residues 120:122, 170:173, 181:183). (F) CSM-based overlay using "best" 4-segment conserved substructure core of 25 residues in N-terminal domain β-strands (1STC residues 56:60 68:73 102:109 118:122). Ensemble visualized at bottom. Residues included in the core are shown in red. Any core substructure-based alignment (D−F) provides advantages over methods that overlay the entire structure (B and C). Panel F represents the "best" core as described in the text and clearly a better ensemble.

staurosporine binding site, resulting in grossly misaligned structures. These two examples are excluded from the mean RMSD calculation.

From our experience, we conclude that these site-attribute-based alignment methods work very well when they work, but they fail to produce any reasonable alignment more often than other methods. In aligning 1STC and p38a structure 3HEC, for example, SiteEngine flips the site over and places the ligand upside down, while reporting an alignment score not significantly unlike that of other correctly oriented site alignments. The absence of a definitive metric that conveys success or failure of the alignment attempt complicates the use of this or similar methods in comprehensive library assembly.

Conserved Substructure Matching. Twenty-five residues were selected from four different N-terminal domain antiparallel β-strands and from the interdomain hinge of the kinase reference structure 1STC to represent a "best" conserved substructure core for evaluation of the CSM method (residues 56:61, 68:73, 102:109, 117:122). The process used to select this core will be discussed below. The method easily and unambiguously identified the corresponding core substructure in each of the 24 kinase domains and overlaid them to give an overall ligand RMSD of 0.7 ± 0.2 Å. This represents a significant reduction in ligand position variance over sequence-based methods that can be easily recognized visually in the ensemble displayed in Figure 1F. While all overlay methods tested result in overlaid ligand ensembles with considerable variance in the ligand position within the plane of the staurosporine macrocycle, the CSM-aligned structures are more tightly clustered within that plane. Sequence-based alignments that include portions of the C-terminal domain will result in ligand positions that vary as the C-terminal domain helical bundle moves up or down relative to the N-terminal domain, which happens often in kinase structures.[38] By choosing a conserved substructure core that specifically includes the hinge that makes hydrogen bonds to the ligands but excludes the C-terminal residues that vary in position, a better overlay could be achieved.

Cross Validation. Multiple kinase PDB structures also exist with the nonhydrolyzable ATP analog, AMP−PNP (phosphoaminophosphonic acid-adenylate ester). Like staurosporine, this pan-kinase inhibitor is often used to facilitate crystallization. Not surprisingly, few kinase domains have been cocrystallized with both staurosporine and AMP−PNP; therefore, the AMP−PNP collection provides a means to cross-validate the overlay method evaluation. Nineteen unique AMP−PNP complexes were overlaid onto 1STC using both PyMOL sequence-based superposition, SiteEngine site alignment, and CSM with the optimized four-stranded core (1CDK, 1CM8, 1J1B, 1JNK, 1O61, 1YXT, 1ZY5, 2ACX, 2C6D, 2OU7, 2PML, 2R5T, 2V55, 2ZV8, 3COK, 3DAK, 3EHA, 3HKO, and 3HX4). All complexes were successfully overlaid by PyMol or CSM methods, but SiteEngine failed to achieve any reasonable alignment for three (2V55, 3COK, and

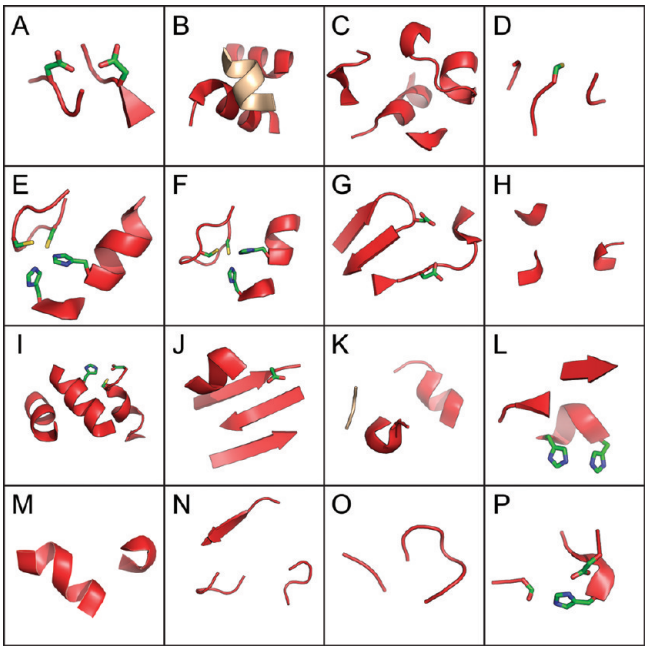


Figure 2. A sampling of core substructures illustrating the structural diversity in different drug targets. Amino acid side chains are shown for strictly conserved amino acids. Protein substructures required to exist as part of a second polypeptide are colored tan. (A) Aspartyl proteases; (B) Bcl-2; (C) β-lactamases; (D) caspases; (E) C_2H_2 Zn-fingers; (F) CX_4C-Zn-fingers; (G) DNA polymerase I; (H) dipeptidyl peptidase-IV; (I) farnesyl transferases; (J) GHKL-ATPases; (K) HMG-CoA reductases; (L) matrix metalloproteinases; (M) nuclear homone receptor ligand-binding domains; (N) NMDA-receptor-1 glycine binding domain; (O) protein tyrosine phosphatases; (P) blood coagulation serine proteases. Sampling derived from CSM overlay methods available within the DrugSite database.

2DAK). The AMP−PNP complex with CDK2 (1CDK) was arbitrarily selected as a standard, and ligand RMSDs for all other structures were computed over the 10 adenosine heterocycle atoms with respect to 1CDK. Using PyMOL, resulting ligand RMSDs ranged from 0.6 to 1.9 Å (mean; 1.0 ± 0.4 Å). With SiteEngine, ligand RMSD values ranged from 0.4 to 1.3 Å (mean 0.8 ± 0.3 Å, excluding those that failed altogether), while the CSM method produced an ensemble with ligand RMSDs ranging from 0.3 to 1.5 Å (mean; 0.8 ± 0.3 Å). These results are summarized in Figure S1, Supporting Information. CSM was able to produce a set of overlays with a smaller distribution of ligand positions with no gross misalignment or failure.

Selecting a Conserved Substructure Core. Selecting a conserved core substructure to drive superposition may be straightforward when the objective is to overlay very similar structures. The first step is to choose a single structure to provide the source for core substructure atomic coordinates. This choice may be completely arbitrary or may be driven by a desire to overlay newly determined structures on a reference frame already in use. The second step requires a choice of residues that will comprise a conserved core. To achieve the best overlay at the ligand binding site, residues that form that site should be given preference but with an eye toward excluding any residues that might have alternate positions when ligands are bound. Since only α-carbon positions contribute to the distance geometry recognized by CSM, side chain flexibility is inconsequential. Generally, residues that are part of the scaffold of surrounding secondary structure are preferred, but any residues may be used, so long as they are conformationally invariant among all the structures that will ultimately be overlaid. The substructure may include disconnected polypeptide segments. Core substructures, both large and small, have been successfully employed (Figure 2). A larger core will serve to diminish the influence of isolated structural differences on the resulting overlays as will a core with residues distributed evenly through three-dimensional space (not nearly coplanar, for example). The selected substructure should represent a unique conformational constellation of Cα geometry within this structure so it can be unambiguously recognized using distance geometry, but it need not be unique to the proteome.

A genuine challenge arises when overlaying diverse structures within a large protein family, because a thorough grasp of the conformational diversity of the family needs to be acquired to make good choices. CSM will fail if the position of any α-carbon in the subject structure is not consistent with the core geometry. The choice of core is always empirical, driven by the specific area of focus in the binding site. No single "correct" choice can be made unambiguously, but choices should provide a more relevant alignment in the context of user need. Moreover, the core residue selection will invariably evolve as more diverse structures are considered. Since the same reference structure may be used as the target for superposition during this evolution, an abrupt change in core will be of little consequence to users of the structural data; only the process for overlay evolves, not the reference frame.

The kinase core substructure selection we have found most generally applicable has undergone significantly more vetting than is typical, but a description of how this core was chosen may illustrate some of the considerations that contributed to the choice.

Selecting a Kinase Core Substructure. For a first trial kinase core, residues were selected in three segments of polypeptide chain immediately surrounding the ATP binding site. This core included 1STC residues M120:Y122 of the hinge and two short strands that lie just under the ligands that pack against the hinge (E170:L173 and Q181:T183). No residues from the N-terminal β-hairpin that lies over the top of bound ATP were included, because this loop is often reoriented or disordered. A match to this small substructural core was found in all the staurosporine complexes of Table 1. The mean RMSD in staurosporine ligand atoms following CSM overlay with this core was lower than had been achieved with any sequence-based alignments (0.9 ± 0.4 Å; Figure 1E), but not all ligand overlays were improved. Some had RMSD as high as 2.0 Å. Larger deviations were attributed to real local structural differences that had untoward influence on the computed rotation matrix because so few atoms were used in the alignment. Small conserved core selections seem to lead to larger ligand position variance, so a larger substructural core was identified that would result in more robust transformations. The consensus structure alignment produced by MAMMOTH-Mult was examined to facilitate selection of an optimal core. As described previously, this procedure identified 109 residues structurally conserved in all the staurosporine-bound structures. Seventy-five of these residues lie in one of seven conserved elements of secondary structure, β2 (1STC residues 56:61), β3 (68:73), β4 (101:109), and β5 (117:122) of the N-terminal domain antiparallel β-sheet, and α_C(142:159), α_D(162:174), and α_F (217:233) of the C-terminal domain. (An eighth segment, including the conserved DFG sequence was not considered

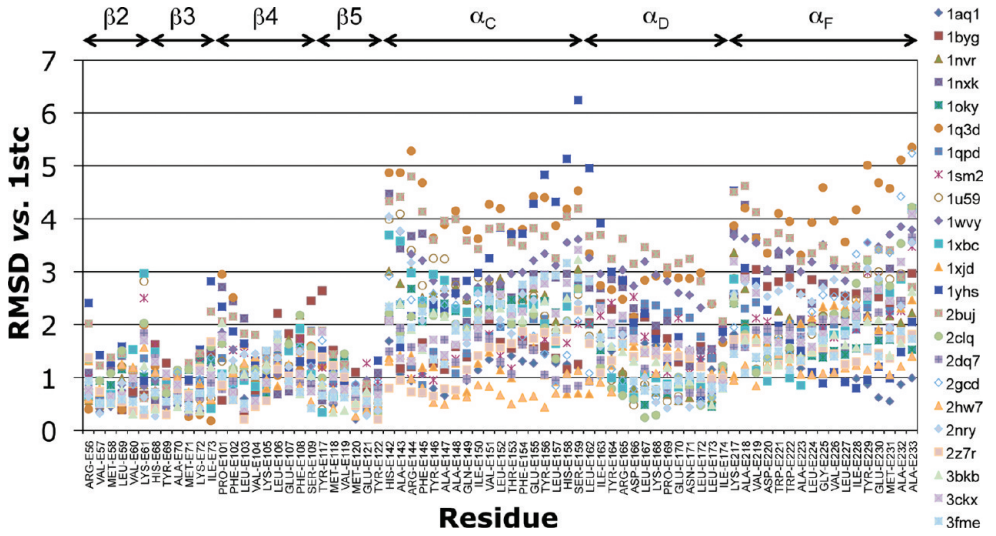


Figure 3. Difference in overlaid staurosporine−kinase complexes by residue. For each of the 75 residues on the horizontal axis, 23 points are shown depicting the RMSD in backbone atom (N, Cα, C, O) positions following pairwise superposition of all these aligned residues from other staurosporine−kinase complexes (Table 1) onto reference structure 1STC. The location of each residue within kinase catalytic domain secondary structure is shown at top.

because of well-known conformational changes to this segment induced by ligand binding, This is an important empirically derived exclusion that arises from a comprehensive understanding of kinome structure and dynamics and stands as a striking example of the importance of human selection of core substructure elements). Structure overlays guided by forced alignment of the 75 residues in this expanded substructural core produced an ensemble of structures with individual ligand RMSDs only marginally improved from those obtained using PyMOL or MAMMOTH (1.0 ± 0.3 Å; data not shown).

In order to better understand how differences are distributed throughout these structures, a detailed comparison of these 24 aligned structures was undertaken. The results are presented in Figure 3, where the RMSD of backbone atoms (N, Cα, C, O) for each of the 23 other structures with respect to the 1STC reference structure is plotted for each of the 75 residues in the expanded substructural core. The trend is quite clear; much better consensus arises from the segments involved in the N-terminal domain antiparallel β-sheet, while the C-terminal domain helices exhibit much greater variability.

Armed with this understanding, the use of structural cores comprised of only the N-terminal domain segments was explored. After evaluating a number of similar alternatives, the 4-stranded "best" core (above) was selected for improved performance as measured by reduced mean ligand RMSD of 0.7 ± 0.2 Å (Figure 1F) and broad applicability to all kinases that have been tested (see below). The choice might seem counterintuitive because these residues make little direct contact with the ligand. Nevertheless, the binding site is at the interface between two domains, and clearly, the position of the N-terminal domain does more to define the position of the hinge and, therefore the inhibitor position, than the C-terminal domain.

The kinase example illustrates an important benefit of CSM over more pairwise structure alignment procedures intended for the overlay of entire structures. Ligand binding sites very often occur at the interface between domains. By specifically guiding the structural alignment to focus on preserving the alignment in one domain that anchors inhibitor binding, it is much easier to visualize and understand conformational changes in the protein target. The kinase example also serves to illustrate that this particular core substructure has been selected specifically to align ligands that bind at the ATP-binding site occupied by staurosporine. Other core substructures may be chosen to achieve optimal overlay at other allosteric ligand binding sites.

Databases of CSM Overlay Methods. As the above kinase example shows, the process for choosing an appropriate core substructure to accomplish alignment of a large family of diverse protein structures with CSM can be involved. We have found it convenient to manage a collection of empirically derived core substructures within relational databases to facilitate the alignment of large numbers of structures. Two different database platforms have been created for this purpose:

BMS-PDB. A proprietary database has been created at Bristol-Myers Squibb to store internally derived crystallographic data describing protein−ligand complexes. Each structure is linked to a protein target and (possibly) a family of related protein targets. One of the objectives of this database is to provide BMS scientists with ready-made "oriented" structures to simplify comparative visualization of different complexes and promote understanding of structure−activity relationships (SAR). Deposited structures are reduced to a computationally relevant biological unit and are oriented on a common positional frame of reference using CSM. Depositors select an overlay method (i.e., conserved substructure) appropriate to the protein and/or inhibitor binding site from a list of stored overlay methods to which they may contribute. By relating specific protein sequence constructs to each overlay method, the choice can be automated, and oriented structures generated automatically by back-end computational services.

The BMS-PDB also allows medicinal chemists uncomplicated access to structures from the RCSB Protein Data Bank[39] that might supplement internal data. Desired structures may be identified by sequence similarity to a BMS target protein and then imported from the RCSB-PDB and oriented by automated back-end services using a defined overlay method and CSM. Due to the concerted effort to load and overlay large numbers of

dx.doi.org/10.1021/ci100475y |J. Chem. Inf. Model. 2011, 51, 1931–1941

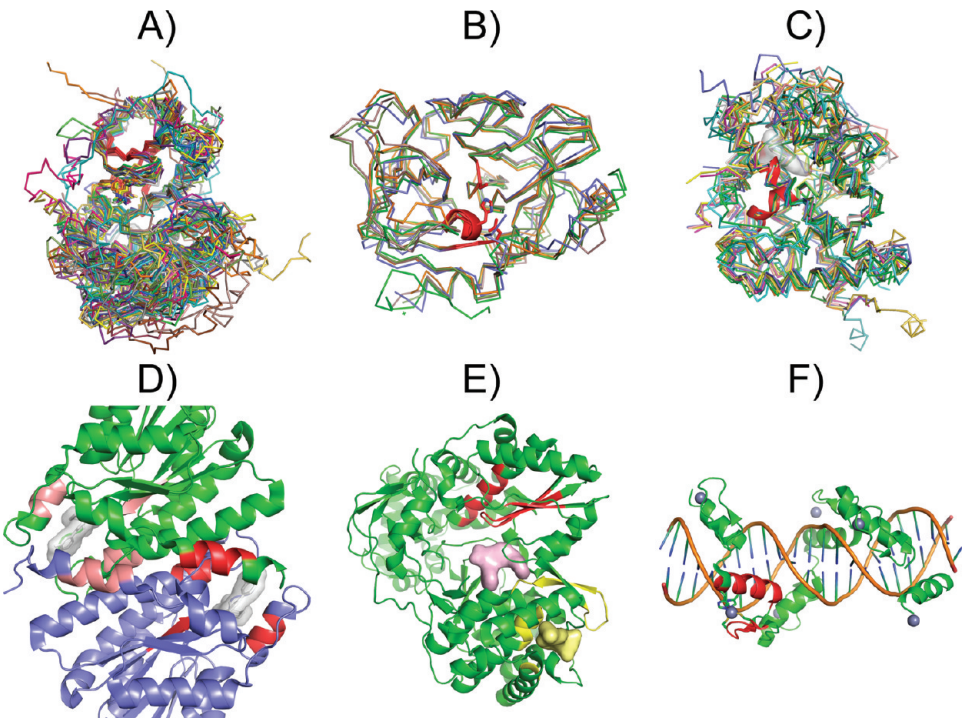


Figure 4. Sample of core substructures and overlay results. (A) Kinases: shown are 24 staurosporine complexes (Table 1) overlaid using the four-segment core (red). (B) A selection of blood coagulation factor structures aligned using a conserved active site core (red ribbon). Shown are factor VIIa (PDB ID 2AEI), factor Xa (3M36), factor XIa (1ZSL), and kallikreins KLK-1 (1SPJ) and KLK-B1 (2ANY). (C) Representative nuclear hormone receptor ligand binding domains overlaid using a core consisting of two helical segments from the fold interior (red ribbon). Structures shown include androgen receptor (2AMB), glucocorticoid receptor (3BQD), liver X receptors α (3IPS) and β (1P8D), peroxizome proliferator activated receptor α (2NPA) and δ (2ZNQ), xenobiotic receptor PXR (3HVL), retinoic acid receptors β (1XAB), γ (1FCZ), and ROR-α (1N83), and thyroid hormone receptor (3HCF). (D) The structure of dimeric 11β-HSD binding sites is formed from different components of the homodimer. Appropriate conserved core substructure is selected from each (red). E) Multiple ligand binding sites on hepatitis C virus polymerase. Inhibitor complexes are overlaid using an appropriate surrounding core (red or yellow). (F) Each DNA recognition site (e.g., 2IL3) or homologues can be overlaid using a conserved C_2H_2 Zn-finger motif (red).

structures from the RCSB-PDB for use in the BMS-PDB, considerable experience has been gained using CSM for this purpose, and a collection of conserved substructural cores useful for aligning drug target families have been amassed.

DrugSite Database. The DrugSite (https://drugsite.msi.umn.edu/; accessed June 15, 2011) is a Web-based interface to CSM algorithms, and a MySQL relational database specifically created to provide a platform for sharing empirically derived overlay methods and usefully reoriented structures. Contributors may test the applicability of different core substructures by identifying residues comprising a trial core substructure within a reference structure and then invoke computational services to align any collection of structures to it using CSM procedures. Distance geometry is computed for subject RCSB-PDB structures and stored in the database to accelerate overlay trials. Aligned ligands (with or without reoriented protein structures) may be conveniently downloaded for more thorough analysis. Overlay methods deemed useful may be published within the database for sharing with other users, so they may be used to reorient any structure sharing the substructure geometry. A graphical summary of example core substructures currently included in the database is provided in Figure 2.

In keeping with the DrugSite mission as an informational resource, a unique computational service is provided that will search any structure for each registered core substructure, thereby returning a list of overlay methods that might be applicable to a structure. Users need not have knowledge of overlay methods to apply them. For many families of well-studied drug targets, overlay methods exist that make the use of CSM methods as convenient as any other approach. It is anticipated that the number of these will grow and help to establish standard frames of reference for many drug targets as more users contribute to this knowledge base. To utilize the alignment service, users need only provide a PDB formatted coordinate file or list of PDB IDs, and the service computes aligned structures that can be downloaded.

Example Applications. The BMS-PDB includes structural data for 180 distinct protein targets of potential interest in drug discovery and design. For each of them, an appropriate core substructure has been identified to allow protein−ligand complexes to be overlaid automatically upon deposition using CSM methods. To enable convenient target hopping, core substructures are often chosen to represent an entire family of proteins so oriented structures may be viewed within the context of a single common frame of reference by users without computational chemistry or bioinformatics expertise. Some protein alignment challenges that were easily addressed with conserved substructure matching are shown in Figure 4.

Kinases. In addition to proprietary structures, the BMS-PDB includes 1158 kinase structures downloaded from the PDB representing 109 distinct kinase catalytic domains.[36] Almost all of these structures have been oriented onto a single common reference frame using CSM and the "best" core substructure of

four segments described above. A sampling of these alignments is shown in Figure 4A. Only 19 structures (1.6%) failed to overlay automatically using this core, which presents a significant advantage over site attribute-based alignment methods. For each of the structures in this subset, others with the same amino acid sequence were aligned; the failures are individual structures scattered throughout the structural kinome that belong to no single kinase classification. Failure of CSM will occur whenever the core substructure geometry (represented by α-carbon distance geometry) is not conserved. Each of these structures either lack specific residues of the conserved core due to disorder or have structural diversity not permitted by the tolerances imposed in distance geometry matching. For example, two structures of Pyk2 described by Han et al.[40] (3FZO and 3FZS) fail to overlay with the larger core because of unanticipated diversity in the β2 position not observed in other Pyk2 structures.[41,42] When such structural diversity is encountered, an alternative core substructure, such as the smaller "hinge core" described earlier (see Figure 1E caption) may be used for CSM alignment instead. A kinase domain structure that cannot be aligned with one or the other of these two core selections has not been encountered.

The core selections are perfectly tolerant of ligand-induced conformational changes known to occur in the kinase family. Aligned coordinates may be readily mined for examples of specific protein conformational variants, such as those representing DGF-in or DGF-out states or structures that are inactive due to a C-helix shift.[43] Since ligand atomic coordinates are reliably anchored to a common reference point, the overlaid collection is particularly useful as a source of bound ligands for computational virtual screening employing fragment deconstruction and reassembly, such as that performed by BREED.[6]

Blood Coagulation Factors. A single conformational core has facilitated the collective superposition of diverse blood coagulation factors including Factor IIa (thrombin), factor VIIa, factor Xa, factor XIa, and kallikrein A and B. The superposition of these related trypsin-like serine proteases is complicated by the fact that proteolytic maturation results in the separation of some, but not all, of these proteins into two protein chains. Since CSM proceeds without regard to connectivity of different peptide fragments and sequence relationships, it handles either subclass with ease. The conserved core used includes only a few highly conserved residues that contain the trypsin-like serine protease catalytic triad and a constraint requiring exact sequence matching at the conserved serine, histidine, and aspartic acid positions. The sequence constraint may not be required but is included to help guard against possible frame shifts in the structural alignment that can occur when distance geometry matching is applied to very short structural segments. CSM software has been used to align all 280 structures imported from the PDB belonging to this class (Figure 4B).

Nuclear Hormone Receptors. The largely α-helical NHR agonist binding domains possesses strong topological homology as a class, but specific helix positions may vary from one protein to another. Ligand binding may induce further changes. To align all members, a simple core substructure comprised of portions of conserved helices 6 and 11 was chosen.[44] The BMS-PDB includes 266 structures from the RCSB-PDB representing 14 different NHR ligand-binding domains. All but three of these can be aligned with the chosen substructural core (Figure 4C). CSM is also well suited for bringing a specific portion of the NHR domains structures into alignment so that the shift of other protein substructures may be quantified in relation to ligand binding. An expanded core comprised of helices beneath the hormone binding site in thyroid hormone receptor, for example, can be utilized to compare different conformational states and the shift of helix 12 that occurs in response to binding of different agonists and antagonists.[44] CSM is particularly well-suited for addressing very specific substructure superposition needs.

Dimeric Drug Targets. Drug targets containing noncrystallographic symmetry and/or binding sites that exist at the interface between two protein chains may be a particular challenge to alignment methods that rely on sequence matching. When the crystallographic data provide two or more independent views of the ligand binding site, more than one chain can be aligned. Sequence-based methods will return only one best match. CSM works well to identify all possible structural matches based solely on geometry. Drug targets, like HMG CoA reductase[45] or 11-β-hydroxysteroid dehydrogenase,[46] that have multiple binding sites at the interface between two protein chains, are easily handled (Figure 4D).

Targets with Alternate Allosteric Binding Sites. An one-size-fits-all approach to structure overlay does not allow for the optimized alignment of structures at remote allosteric binding sites. The RNA-dependent RNA polymerase of the hepatitis C virus (NS5B) is an excellent example (Figure 4E). Different inhibitor classes are known to bind in the primer grip subsite[47,48] or to other allosteric sites separated by over 20 Å from the primer grip subsite.[49] Since the protein is large and has three domains that are known to shift during the catalytic cycle, it is not surprising that secondary structure or sequence-based overlay techniques do a poor job of aligning ligands at remote binding sites. CSM coupled with different core substructure selections may be used to obtain an overlay for each structure that is most appropriate to each binding site (Figure 4E).

Overlaying Repeating Motifs. The CSM technique is also well suited for the superposition of small recurring structural motifs that occur in many protein structures, such as the CX_4C-H_2 Zn-finger (Figure 4F). Superposition of individual protein domains or motifs is often complicated in practice by the need to specify individual residue alignments. By defining a core substructure, CSM can easily identify and overlay all occurrences of the motif in one or multiple structures.

SUMMARY

Confronted with a need to assemble large libraries of pre-aligned protein−ligand complex structures to aid in the understanding of structure−activity relationships and structure-aided drug design, but finding no existing software to accurately and conveniently assemble such libraries, we have created a software system that facilitates the automated alignment of empirically selected core substructures in related proteins. We have amassed a collection of conserved substructures suitable for alignment of important drug receptors and have created both proprietary and publicly accessible databases and web services that can be used to share overlay methods with colleagues. The software can be trivial to apply, once conserved core substructure and reference structure have been selected. The software can also be used to prepare ensembles of overlaid structures for comparison of different core substructure choices.

Core substructure matching has been found to be a convenient and scalable method for the overlay of large numbers of related protein structures of interest in drug design with significantly better accuracy than achieved by sequence-based alignment

services embedded in many popular molecular graphics packages. CSM is more robust than site attribute-based alignment tools that sometimes fail to find sufficient similarity or create misalignments of related structures. The method permits us to meet important library design objectives to allow selection of a single known structural reference frame on which all related structures can be aligned, the overlay of apo structure binding sites and sites that lie at protein domain interfaces. Identifying an appropriate conserved structural core may be challenging, but the end reward is a collection of overlaid complexes that is more accurate and relevant than those produced by more holistic overlay approaches.

Every protein−ligand system is unique, and one's understanding of each may benefit from an expert, knowledge-driven approach to complex superposition. CSM may provide a level of control in superposition that can be exploited to address very specific questions regarding ligand-induced conformational change. An important virtue of CSM is that it never gives rise to an unexpectedly bad overlay. To succeed, the assumption that the core substructure is conserved between the two structures must be true. This makes the method particularly useful for assembling large numbers of overlaid structure that can be subjected to computational mining or analyses. CSM could also serve as an excellent tool for automated identification of substructures to be aligned as the first step in seeding multiple structure alignment tools, such as PYMSS,[21] although we have not attempted to apply it in this application.

In sharing applicable overlay methods through the DrugSite database (https://drugsite.msi.umn.edu/), a platform has been created where others may benefit from specialized expertise or make their own contributions to developing best practices for overlaying particular protein−ligand systems. This Web site provides simplified access to CSM-based procedures and overlaid atomic coordinates of proteins that facilitate easy viewing and collaboration for structure-based design.

■ ASSOCIATED CONTENT

Supporting Information. Figure S1 is provided. This information is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: finze007@umn.edu.

■ ACKNOWLEDGMENT

Funding for this work (B.C.F.) is provided by the Minnesota Department of Employment and Economic Development (SPAP-06-00140P-FY07). The Authors wish to thank Todd Geders for assistance in the preparation of this manuscript. The support of the Minnesota Supercomputer Institute is also gratefully acknowledged.

■ REFERENCES

(1) Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discovery* **2007**, *6*, 211–219.

(2) Ciulli, A.; Abell, C. Fragment-based approaches to enzyme inhibition. *Curr. Opin. Biotechnol.* **2007**, *18*, 489–496.

(3) Mauser, H.; Guba, W. Recent developments in de novo design and scaffold hopping. *Curr. Opin. Drug. Discovery Dev.* **2008**, *11*, 365–374.

(4) Velcicky, J.; Feifel, R.; Hawtin, S.; Heng, R.; Huppertz, C.; Koch, G.; Kroemer, M.; Moebitz, H.; Revesz, L.; Scheufler, C.; Schlapbach, A. Novel 3-aminopyrazole inhibitors of MK-2 discovered by scaffold hopping strategy. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 1293–1297.

(5) Gopalsamy, A.; Shi, M.; Hu, Y.; Lee, F.; Feldberg, L.; Frommer, E.; Kim, S.; Collins, K.; Wojciechowicz, D.; Mallon, R. B-Raf kinase inhibitors: hit enrichment through scaffold hopping. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 2431–2434.

(6) Pierce, A. C.; Rao, G.; Bemis, G. W. BREED: Generating novel inhibitors through hybridization of known ligands. Application to CDK2, p38, and HIV protease. *J. Med. Chem.* **2004**, *47*, 2768–2775.

(7) Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1978**, *A34*, 827–828.

(8) Madhusudhan, M. S.; Webb, B. M.; Marti-Renom, M. A.; Eswar, N.; Sali, A. Alignment of multiple protein structures based on sequence and structure features. *Protein Eng., Des. Sel.* **2009**, *22*, 569–574.

(9) Maiti, R.; Van Domselaar, G. H.; Zhang, H.; Wishart, D. S. SuperPose: a simple server for sophisticated structural superposition. *Nucleic Acids Res.* **2004**, *32* (Web Server issue), W590–W594.

(10) Krissinel, E.; Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 2256–2268.

(11) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739–747.

(12) Zemla, A. LGA A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **2003**, *31*, 3370–3374.

(13) Oldfield, T. J. CAALIGN: a program for pairwise and multiple protein-structure alignment. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2007**, *63*, 514–525.

(14) Lupyan, D.; Leo-Macias, A.; Ortiz, A. R. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics* **2005**, *21*, 3255–3263.

(15) Shatsky, M.; Nussinov, R.; Wolfson, H. J. A method for simultaneous alignment of multiple protein structures. *Proteins* **2004**, *56*, 143–156.

(16) Dror, O.; Benyamini, H.; Nussinov, R.; Wolfson, H. MASS: multiple structural alignment by secondary structures. *Bioinformatics* **2003**, *19* (Suppl 1), i95–i104.

(17) Gerstein, M.; Levitt, M. Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci.* **1998**, *7*, 445–456.

(18) Taylor, W. R.; Flores, T. P.; Orengo, C. A. Multiple protein structure alignment. *Protein Sci.* **1994**, *3*, 1858–1870.

(19) Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779–815.

(20) Russell, R. B.; Barton, G. J. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* **1992**, *14*, 309–323.

(21) Hill, A. D.; Reilly, P. J. Comparing programs for rigid-body multiple structural superposition of proteins. *Proteins* **2006**, *64*, 219–226.

(22) 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–D234.

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

(24) Jambon, M.; Imberty, A.; Deléage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52*, 137–145.

(25) 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.

(26) Günther, J.; Bergner, A.; Hendlich, M.; Klebe, G. Utilising structural knowledge in drug design strategies: applications using Relibase. *J. Mol. Biol.* **2003**, *326*, 621–636.

(27) Porter, C. T.; Bartlett, G. J.; Thornton, J. M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **2004**, *32* (Database issue), D129–D133.

(28) Nebel, J.-C.; Herzyk, P.; Gilbert, D. R. Automatic generation of 3D motifs for classification of protein binding sites. *BMC Bioinformatics* **2007**, *8*, 321.

(29) Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**, *71*, 1755–1778.

(30) Dundas, J.; Adamian, L.; Liang, J. Structural Signatures of Enzyme Binding Pockets from Order-Independent Surface Alignment: A Study of Metalloendopeptidase and NAD Binding Proteins. *J. Mol. Biol.* **2011**, *406*, 713–729.

(31) Finzel, B. Mastering the LORE of protein structure. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1995**, *51*, 450–457.

(32) Finzel, B. LORE: exploiting database of known structures. *Methods Enzymol.* **1997**, *277*, 230–242.

(33) Jones, T. A.; Thirup, S. Using known substructures in protein model building and crystallography. *EMBO J.* **1986**, *5*, 819–822.

(34) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.

(35) Tamaoki, T.; Nomoto, H.; Takahashi, I.; Kato, Y.; Morimoto, M.; Tomita, F.; Staurosporine, A Potent Inhibitor of Phospholid/Ca++ Dependent Protein Kinase. *Biochem. Biophys. Res. Commun.* **1986**, *135*, 397–402.

(36) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912–1934.

(37) Delano, W. *Introduction to PyMol*; Schrödinger, LLC: Cambridge, MA; http://www.pymol.org/. Accessed January 15, 2010.

(38) Kannan, N.; Haste, N.; Taylor, S. S.; Neuwald, A. F. The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1272–1277.

(39) 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.

(40) Han, S.; Mistry, A.; Chang, J. S.; Cunningham, D.; Griffor, M.; Bonnette, P. C.; Wang, H.; Chrunyk, B. A.; Aspnes, G. E.; Walker, D. P.; Brosius, A. D.; Buckbinder, L. Structural characterization of proline-rich tyrosine kinase 2 (PYK2) reveals a unique (DFG-out) conformation and enables inhibitor design. *J. Biol. Chem.* **2009**, *284*, 13193–13201.

(41) Walker, D. P.; Bi, F. C.; Kalgutkar, A. S.; Bauman, J. N.; Zhao, S. X.; Soglia, J. R.; Aspnes, G. E.; Kung, D. W.; Klug-McLeod, J.; Zawistoski, M. P.; McGlynn, M. A.; Oliver, R.; Dunn, M.; Li, J.-C.; Richter, D. T.; Cooper, B. A.; Kath, J. C.; Hulford, C. A.; Autry, C. L.; Luzzio, M. J.; Ung, E. J.; Roberts, W. G.; Bonnette, P. C.; Buckbinder, L.; Mistry, A.; Griffor, M. C.; Han, S.; Guzman-Perez, A. Trifluoromethylpyrimidine-based inhibitors of proline-rich tyrosine kinase 2 (PYK2): structure-activity relationships and strategies for the elimination of reactive metabolite formation. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 6071–6077.

(42) Walker, D. P.; Zawistoski, M. P.; McGlynn, M. A.; Li, J.-C.; Kung, D. W.; Bonnette, P. C.; Baumann, A.; Buckbinder, L.; Houser, J. A.; Boer, J.; Mistry, A.; Han, S.; Xing, L.; Guzman-Perez, A. Sulfoximine-substituted trifluoromethylpyrimidine analogs as inhibitors of proline-rich tyrosine kinase 2 (PYK2) show reduced hERG activity. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 3253–3258.

(43) Jacobs, M. D.; Caron, P. R.; Hare, B. J. Classifying protein kinase structures guides use of ligand-selectivity profiles to predict inactive conformations: structure of lck/imatinib complex. *Proteins* **2008**, *70*, 1451–1460.

(44) Togashi, M.; Borngraeber, S.; Sandler, B.; Fletterick, R. J.; Webb, P.; Baxter, J. D. Conformational adaptation of nuclear receptor ligand binding domains to agonists: potential for novel approaches to ligand design. *J. Steroid Biochem. Mol. Biol.* **2005**, *93*, 127–137.

(45) Pfefferkorn, J. A.; Choi, C.; Song, Y.; Trivedi, B. K.; Larsen, S. D.; Askew, V.; Dillon, L.; Hanselman, J. C.; Lin, Z.; Lu, G.; Robertson, A.; Sekerke, C.; Auerbach, B.; Pavlovsky, A.; Harris, M. S.; Bainbridge, G.; Caspers, N. Design and synthesis of novel, conformationally restricted HMG-CoA reductase inhibitors. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 4531–4537.

(46) Wang, H.; Ruan, Z.; Li, J. J.; Simpkins, L. M.; Smirk, R. A.; Wu, S. C.; Hutchins, R. D.; Nirschl, D. S.; Van Kirk, K.; Cooper, C. B.; Sutton, J. C.; Ma, Z.; Golla, R.; Seethala, R.; Salyan, M. E. K.; Nayeem, A.; Krystek, S. R.; Sheriff, S.; Camac, D. M.; Morin, P. E.; Carpenter, B.; Robl, J. A.; Zahler, R.; Gordon, D. A.; Hamann, L. G. Pyridine amides as potent and selective inhibitors of 11beta-hydroxysteroid dehydrogenase type 1. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3168–3172.

(47) Pfefferkorn, J.; Greene, M.; Nugent, R.; Gross, R.; Mitchell, M.; Finzel, B.; Harris, M.; Wells, P.; Shelly, J.; Anstadt, R.; Kilkuskie, R.; Kopta, L.; Schwende, F. Inhibitors of HCV NS5B polymerase. Part 1: Evaluation of the southern region of (2Z)-2-(benzoylamino)-3-(5-phenyl-2-furyl)acrylic acid. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2481–2486.

(48) Ruebsam, F.; Webber, S. E.; Tran, M. T.; Tran, C. V.; Murphy, D. E.; Zhao, J.; Dragovich, P. S.; Kim, S. H.; Li, L.-S.; Zhou, Y.; Han, Q.; Kissinger, C. R.; Showalter, R. E.; Lardy, M.; Shah, A. M.; Tsan, M.; Patel, R.; Lebrun, L. A.; Kamran, R.; Sergeeva, M. V.; Bartkowski, D. M.; Nolan, T. G.; Norris, D. A.; Kirkovsky, L. Pyrrolo[1,2-b]pyridazin-2-ones as potent inhibitors of HCV NS5B polymerase. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3616–3621.

(49) Li, H.; Tatlock, J.; Linton, A.; Gonzalez, J.; Borchardt, A.; Dragovich, P.; Jewell, T.; Prins, T.; Zhou, R.; Blazel, J.; Parge, H.; Love, R.; Hickey, M.; Doan, C.; Shi, S.; Duggal, R.; Lewis, C.; Fuhrman, S. Identification and structure-based optimization of novel dihydropyrones as potent HCV RNA polymerase inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4834–4838.