JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

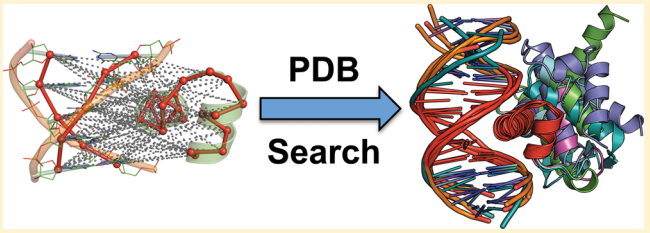
Searching for Likeness in a Database of Macromolecular Complexes

Jeffrey R. Van Voorst and Barry C. Finzel*

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

Supporting Information

ABSTRACT: A software tool and workflow based on distance geometry is presented that can be used to search for local similarity in substructures in a comprehensive database of experimentally derived macromolecular structure. The method does not rely on fold annotation, specific secondary structure assignments, or sequence homology and may be used to locate compound substructures of multiple segments spanning different macromolecules that share a queried backbone geometry. This generalized substructure searching capability is intended to allow users to play an active part in exploring the role specific substructures play in larger protein domains, quaternary assemblies of proteins, and macromolecular complexes of proteins and polynucleotides. The user may select any portion or portions of an existing structure or complex to serve as a template for searching, and other structures that share the same structural features are identified, retrieved and overlaid to emphasize substructural likeness. Matching structures may be compared using a variety of integrated tools including molecular graphics for structure visualization and matching substructure sequence logos. A number of examples are provided that illustrate how generalized substructure searching may be used to understand both the similarity, and individuality of specific macromolecular structures. Web-based access to our substructure searching services is freely available at https://drugsite.msi.umn.edu.



INTRODUCTION

Several years ago, Pabo and Nekludova[1] undertook a comprehensive analysis of the geometry of then available structures of protein−DNA complexes in an attempt to explain the complex relationship between the amino acid sequences of DNA recognition elements and the nucleotide sequences they recognize. Using a coordinate frame anchored at the center of a DNA duplex major groove, they examined differences in the way that the recognition helix is positioned in different complexes. Their coordinate system was subsequently expanded and updated by others who used it to quantify similarity and to cluster protein−DNA interactions.[2] In both studies, examples of similar helix placement could be found even in cases where the rest of the protein fold was quite different. Conversely, examples were found where the same motif binds DNA in geometrically distinct ways (e.g., the structurally conserved helix-turn-helix motif of homeodomains and other transcription factors[1]). An identification of similar interaction subclasses emerged wherein the defining geometric feature was the orientation of the recognition helix in the major groove. This helix defines the framework or scaffold upon which specific amino acid sidechains that interact with specific nucleobases are placed. It becomes easier to compare and understand the relationship between different contacts when only those that utilize the same framework geometry are considered.[1]

Protein−DNA complexes represent a special case in macromolecular assembly, in that the DNA duplex major groove provides a readily identifiable focal point of the interaction, but a similar framework description likely applies equally well to any protein−protein complex. In several papers, Keskin has discussed the presence of recurrent substructures in similar frameworks in protein−protein complexes[3,4] and notes that these substructures may originate from proteins with varying degrees of similarity in fold.[3] Presumably, each of these frameworks also relies on patterns of compatible sequence substitutions, so that only certain combinations of sequence and fold can lead to an effective interaction. If we could easily and clearly identify exactly that subset of all experimentally derived macromolecular complexes that employ the same combination of substructural motifs with the same relational geometry, it would be easier to recognize sequence dependencies when they exist.

Structures that are similar to this level of detail described above possess what we will call a backbone conformational *likeness*: a structural similarity that ensures a one-to-one correspondence of aligned amino acid residues. Such likeness may exist in entire folds, but it may extend to only a local substructural motif. Likeness can also exist across an intermolecular interface, such as in the case of the similar protein DNA interactions[1] or locally similar protein−protein complexes.[3] It is important when comparing structures to distinguish *likeness* from *similarity*. Numerous software tools exist to help identify similarity in protein sequences,[5,6] functional constellations of residues,[7−9] motifs,[10−12] or folds,[13−15] or to align homologous protein structures.[16−18] While some of these methods incorporate geometry into how

Received: April 28, 2013
Published: September 18, 2013

   dx.doi.org/10.1021/ci4002537 | J. Chem. Inf. Model. 2013, 53, 2634−2647

similarity is assessed, they are also parametrized to overlook geometrical nuance so that more distant topological relationships may be identified; that is, they do not impose a strict requirement for backbone likeness. In many applications, this flexibility is desirable, but there are instances where small differences in structure, such as those conferred by an insertion or deletion or the shift of one element of secondary structure in relation to others represent precisely the difference that confers a specificity of function. There are tools that seek to extend the notion of similarity to protein−protein complexes, building upon successful domain fold classifications of SCOP[19] to classify fold−fold interactions,[20,21] and create subclassifications by fold face type (e.g., SCOPPI[22]). These classifications provide annotations of interaction similarity ("who is like whom?"), but they do not attempt to capture the fine details of interaction molecular structure and do not provide the means for the direct manipulation of atomic coordinates. In addition, these methods cannot handle RNA or DNA conformation at all. Structure-based alignment tools like CE,[18] TM-align,[23] or FATCAT[24] produce only pairwise structure alignments, are slow for use in comprehensive searching, and cannot be used to find likeness in intermolecular complexes.

In this report, we describe a tool for generalized substructure searching that can be used to identify and overlay multi-component substructural motifs in experimental protein structures and macromolecular complexes that share true likeness. We look for identity (within a user-adjustable tolerance) in the relative positions of backbone atoms and do not rely on any explicit consideration of protein sequence, secondary structure assignment, or fold annotation. In finding substructural likeness, as we will define it, geometrically dissimilar structures, such as those protein−DNA interactions which employ an alternate orientation for the recognition helix, are easily excluded so that one-to-one mapping of the amino acid sequences of the remaining like substructures is meaningful. The software described here allows us to ask the question "What *other structures* have this geometry?" and to retrieve and align all other examples for direct conformational comparison to the target geometry. This question is frequently asked by molecular scientists as they try to understand the affinity of a particular transcription factor for an RNA sequence, or to speculate on the likely stability of a nanoparticle design, or if they want to explore structural similarity in the interaction of a biotherapeutic agent (an antibody or growth factor) with other proteins that target the same protein receptor.

Our searching services utilize Distance Geometry Matching (DGM) to identify likeness in protein substructures. DGM relies on the fact that the distances between alpha-carbons in a polypeptide chain can be used like fingerprints to efficiently identify conformationally similar substructures.[25] DGM has long been successfully exploited to identify short peptide segments with particular conformational characteristics for use in electron density fitting,[26,27] to validate and correct protein models in crystallographic refinement,[28,29] and for use as tools in homology modeling or threading.[30−33] For most of these applications, however, only the distances within short polypeptides or between near neighbors are used, because the distance geometry needed to represent short peptides is of easily manageable size. Distances between more far-flung residues in a protein chain have been used effectively for fold classification,[34,35] but geometry is typically first reduced to fewer distances representing contiguous residues or secondary structure. Distances between a small number of catalytic residues have also been used to locate functional groups in proteins that might be consistent with a particular catalytic function.[8,36,37]

We have recently described methods that invoke DGM to superimpose diverse protein families by aligning locally conserved core substructures.[38] The method has been shown to be exquisitely sensitive to small differences in structure, and it is very capable of finding similar substructural motifs in a library of other structures, provided the distance geometry needed to characterize subject structures is computed beforehand. With this algorithm in hand, we have established a searchable database representing all structures in the Protein Data Bank[39](PDB). The worldwide PDB has grown to include nearly 90 000 individual protein structures and more than 40 000 examples of proteins molecules in contact with one another. This repository of data represents an incredible resource for understanding both molecular structure and the foundations of intermolecular interactions, if it can be mined effectively and efficiently.

The size and molecular diversity of the PDB presents significant challenges to the implementation of comprehensive substructure searching using distance geometry matching. If we want to extract like substructures from intermolecular complexes or quaternary assemblies we must store and manage the distances that relate different chains to each other. Moreover, the PDB now includes nearly 2500 RNA or DNA structures, and over 4000 protein−oligonucleotide complexes that constitute different challenges in the identification of similar structures.

Here, we describe our implementation of methods and a searchable database that enable comprehensive substructure searching of the entire PDB for any arbitrary substructure comprised of polypeptide and/or polynucleotide segments of an existing structure or model. Several possible applications of substructure searching are illustrated with specific examples. While we anticipate future updates to software and hardware that will result in further performance enhancements, we report the existence of a web-based interface that provides universal access to our database and DGM services and describe the current performance attributes in this report.

METHODS

Search Service Capabilities. The computational services that we have implemented may be used to search a centrally located structural library for those library structures that include the complex substructure (target geometry) identified by the user. Ensembles of structures with like substructures may be retrieved and superimposed onto the user's substructure by using the matched substructures. A target geometry is defined by a selection of specific residues from a given PDB structure or a provided protein and/or polynucleotide model in PDB coordinate format. The intent is to provide a means to identify structures that are similar in a particular way and to create an ensemble of overlaid atomic coordinates that may be directly visualized in a way that emphasizes this similarity. Sequence logos[40,41] may also be prepared from ensembles of substructures that are similar. The Web service also embeds 3DCOFFEE for rapidly computing multiple sequence alignments that constrain the coincidence of sequence segments corresponding to structural matches.[42]

Distance Geometry Search Algorithm. The search algorithm based on distance geometry is conceptually quite simple and closely resembles and extends the approach for

protein substructure searching implemented in *LORE*.[43] Distances between all possible pairs of alpha-carbons in a structure to be searched (a "library structure") are precomputed and stored as a triangular matrix ($\mathbf{D}_{ij}$). The amino acid sequence of the library structure is also stored, so that it may be used to impose user selected explicit sequence constraints on returned hits. The distance geometry representing the target substructure geometry is computed from the alpha-carbon coordinates selected by the user and stored in a smaller triangular matrix $\mathbf{I}_{ij}$ (Figure 1). Matching library structures are

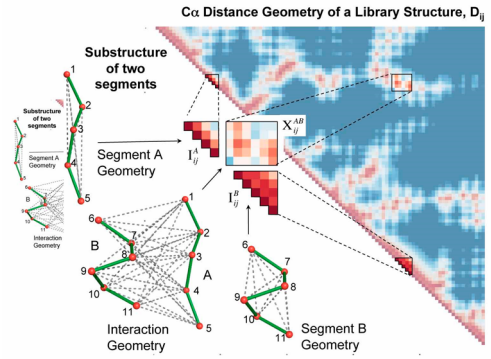


Figure 1. Graphical rendering of the Distance Geometry Matching algorithm. Distance geometry ($\mathbf{D}_{ij}$) in the form of a table of all N^2 Cα−Cα distances in a library structure is precomputed. Only half of these distances are unique due to symmetry (example rendered for illustrative purposes as a triangular heat map). A target substructure potentially consisting of multiple bonded fragments such as the beta-alpha pair shown (segments A and B) is represented by an aggregate fingerprint of distance geometry elements encoding the internal geometry of each segment (I^A and I^B), and separate elements encoding the interaction distance geometry (X^{AB}). The algorithm first seeks a match for each segment by crawling along the near diagonal, looking for elements identical to I^A, then I^B (within a tolerance), as illustrated by Finzel,[77] and then proceeds to cherry-pick known off-diagonal blocks for examination. A complete match will possess identical distance geometry in both near-diagonal and off-diagonal blocks as shown.

identified only when this entire target substructure is found intact in the library structure. Simple target geometries may match different substructures in the same structure, and each match is reported.

A search for an intact target segment of multiple residues is completed using a brute-force walk "down" the diagonal of the library structure one residue (matrix row) at a time while looking for blocks of distances ($\mathbf{D}_{nm}$) that are identical to the target distances ($\mathbf{I}_{ij}$). Matching substructures must have the same constellation of interatomic distances ($\mathbf{D}_{nm} = \mathbf{I}_{ij}$) within the selected tolerance if the segment beginning at residue n of the library structure is to be considered a match. This procedure can easily be extended to search for more complex substructures, such as those comprised of disconnected segments of polypeptide (Figure 1), and therein lies its primary distinction. After finding a first segment using the above procedure (I^A), a second search is conducted for the second segment (I^B), and then others (if necessary). As each candidate segment match is found, the off-diagonal interaction distance geometry X^{AB}, corresponding to the interatomic distances between the candidate segment and all previous segment matches, is also checked to ensure a match (Figure 1).

Tolerances imposed in the identification of matching distance geometry are important parameters in searching. Unlike the RMSD, which represents the mean error in all atoms, the tolerances impose a maximum difference in the position of any one atom, or more precisely, the largest difference in any Cα−Cα distance allowed in matches. Because we may want to allow more flexibility in the relative orientation of different segments, separate tolerances may be imposed for intrasegment distances, and intersegment distances ($^{C\alpha}T_{AA}$, and $^{C\alpha}T_{AB}$, respectively). Typical tolerance values are presented as part of the examples provided below.

Distance Geometry Matching identifies a one-to-one correspondence between all residues in the target and library structures. A user selected subset of the atoms (typically backbone atoms) present in corresponding residues are then used to overlay the matching substructures by determining the optimal least-squares superposition.[44,45] The RMSD generated for each matching substructure provides the final gauge of likeness to the target.

Since the algorithm searches for segments rather than individual residues, it first checks elements furthest from the diagonal (the distance between residue i and residue $i + N$ of an N-residue target segment); these interatomic distances are likely to vary the most. Most nonmatching substructures can be discarded after examining only one value. This is a distinct difference from the approach taken in other applications employing distance geometry matching such as SPASM, where each target residue is treated as an independent entity.[36] Consequently, searching for a particular segment geometry can be very fast; the speed of searching is often limited by the time required to load the library distance geometry matrices into memory (RAM). Moreover, since any single mismatch in distance geometry will cause rejection of a subject substructure, it is faster to search for larger, complex segments of unusual geometry, than shorter segments of common geometry. The method thereby complements software like SPASM[36] or PINTS,[8] that work well as tools to search for small clusters of residues, although our method will return the same results.

Because this is a brute-force search method relying only on distances calculated directly from experimental atomic coordinates, there is no ambiguity regarding whether or not a structure is a match to the target—either the distances are equivalent within the requested tolerance, or they are not. There is no need to impose probabilistic models to account for uncertainty. False positives are not possible because every matching substructure must share the distance geometry likeness at least at the level of the user provided tolerances. Similarly, false negatives are not possible; any "missing" structures are due to them having truly distinct, local, geometric differences.

Polynucleotide Distance Geometry. The utility of Cα−Cα distances in polypeptide substructure classification is well-established, but it is not as obvious which atom or atoms could play a comparable role to encode RNA or DNA geometry. The polynucleotide backbone contains six rotatable bonds per residue, with six atoms connected in a chain along the backbone from 5′ to 3′ (Figure 2). While the phosphorus atom might most centrally represent the phosphate position, C4′ might be regarded as most centrally located on the ribose. C1′ could be most well correlated with the position of the nucleobase, since it anchors an end of the glycosidic bond. It is convenient to have only one position from each residue contributing to distance geometry, but that position could be

dx.doi.org/10.1021/ci4002537 | J. Chem. Inf. Model. 2013, 53, 2634−2647

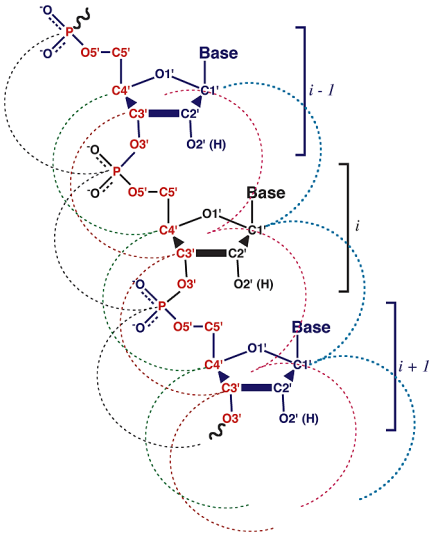


Figure 2. Possible interatomic distances along the polynucleotide backbone considered for use as distance geometry.

easily computed from other atom positions, such as the center of geometry of all ribose ring atoms. The most useful distance geometry will be that which correlates most closely with similarity in the associated substructures, as reflected in the backbone atom RMSD.

In order to choose the most representative and discriminating geometrical marker for polynucleotide distance geometry, we designed computational experiments wherein thousands of different RNA substructure geometries were used as targets in distance geometry matching. Matches identified with alternate distance geometry representations were evaluated by calculating of the corresponding RMSD from all ribonucleotide backbone atoms (P, O5′, C5′, C4′, C3′, O3′, C2′, C1′, and O1′). In separate experiments, the 16S rRNA from the *Haloarcula marismortui* 50S ribosome (chain 0: PDB-id 1vq8)[46] was rendered as distance geometry for searching as a library structure using different candidate atomic positions P, C4′, C3′, C1′, and R′. (We define R′ as the center of geometry of the five ribose ring atoms). This library structure was then searched for substructures similar to each successive 5-nucleotide segment (residues *i*:*i* + 4) extracted from the same 16S RNA model. For segments of length five, this includes 2691 different target substructure geometries. For each target, the number of matches was tabulated (Δd tolerance 1.0 Å), along with the RMSD realized following overlay of the corresponding geometry atom (or atoms), and the RMSD following overlay of all nine backbone atoms. This gave a very large statistical sampling of search and superposition results for comparison. A similar experiment was repeated using target substructures of length eight (8-mers). A summary of results is given in Table 1.

Distance geometry based on C1′ is clearly preferred to describe the structure of polynucleotides. The use of C1′ results in a much higher number of quality hits (those where good distance geometry translates into a good RMSD). We were also encouraged to find that a good match to C1′ distance geometry translates into at least a fair RMSD fit (better than 1.5 Å) more than 96% percent of the time, regardless of target length. Changes in C1′ position often result when the base occupies a nonstandard position. Other atoms further from the base, particularly the phosphate, can be much less sensitive to differences in base position. We find it interesting that the mean backbone RMSD for C1′ derived hits (0.84 Å) is better even than that found for the R′ derived hits (0.95 Å), where R′ represents the aggregate geometry of more than half of all backbone atoms.

The search algorithm and searchable database has extended to include a C1′ distance geometry representation for any RNA or DNA represented on PDB SEQRES records. As with protein chains (see above), each chain is represented by a triangular matrix of intramolecular distances, and, when multiple chains are present, a separate rectangular matrix representing the interaction geometry. For structures containing protein and polynucleotide chains, the interaction geometry between protein and polynucleotide chains is encoded by the distances between protein Cα-atoms and nucleotide C1′ atoms (Figure 3). Separate tolerances ($^{C1'}T_{AA}$, $^{C1'}T_{AB}$, $^{C1'C\alpha}T_{AB}$) can be applied in assessing distance geometry equivalence involving polynucleotides.

Searchable Database. We search a database of precomputed distance geometry derived from individual structures in the Protein Data Bank. Distances between all possible pairs of geometry-defining atoms in all chains (Cα, in proteins) in each

Table 1. Assessment of Atom Choices for Nucleic Acid Distance Geometry[a]

geometry atom[b]	no. matches[c]	no. unique[d]	1-atom RMSD[e]	backbone RMSD[f]	% good[g]	% fair[h]	% mismatch[i]
			5-mers				
C1′	335,702	682	0.56	0.80	79.7	96.2	0.8
C3′	656760	149	0.65	1.13	54.4	76.0	10.7
C4′	648816	240	0.66	1.04	58.2	81.9	6.3
R′	607832	208	0.62	1.03	60.7	82.8	6.9
P	426452	194	0.64	1.29	48.3	67.4	20.7
			8-mers				
C1′	3,950	2101	0.64	0.84	75.6	97.5	0.2
C3′	7,542	1717	0.71	0.98	61.3	90.1	2.5
C4′	6,728	1825	0.72	0.94	67.0	91.3	1.8
R′	8,220	1766	0.70	0.95	66.4	91.6	2.0
P	4,428	1900	0.67	1.20	47.8	73.5	13.1

[a]All target and matching polypeptides from 50S ribosomal subunit structure PDB-id 1vq8. [b]Atom used for distance geometry. R′ is center of geometry of ribose atoms (C5′,C4′ C3′ C2′ C1′ O1′). [c]Count of segments matching all targets by distance geometry (1Å tolerance). [d]Count of target segments that appear unique by distance geometry (1Å tolerance). [e]Mean RMSD for all matches computed over only the geometry atom (or atoms). [f]Mean RMSD of backbone atoms (including P, O5′, C5′, C4′, C3′, O3′, C2′, C1′, O1′). [g]Backbone RMSD < 1.0 Å. [h]Backbone RMSD < 1.5 Å. [i]Backbone RMSD > 2.0 Å

dx.doi.org/10.1021/ci4002537 | J. Chem. Inf. Model. 2013, 53, 2634−2647

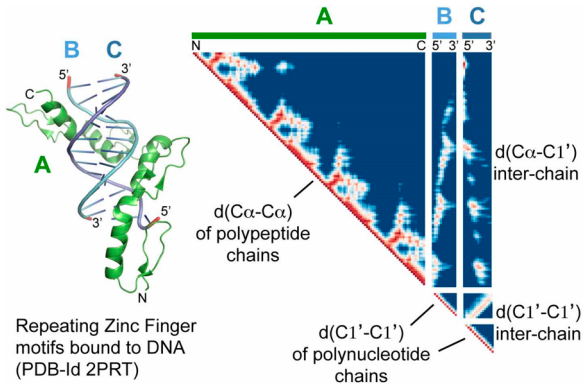


Figure 3. Example of a crystal structure comprised of mixed polypeptide and polynucleotide chains and the corresponding representation as distance geometry. The structure is the single-chain Wilms tumor suppressor zinc-finger domain protein (PDB-id 2prt)[75] that includes four Zn-finger motifs (chain A) clearly identifiable in the distance geometry heat map (red spikes extending off-diagonal) complexed with duplex DNA (chains B, C). Distance geometry is stored in the database as six discreet objects: a large half-matrix of distance geometry $d(C\alpha-C\alpha)$ encoding the conformation of the polypeptide chain, two small half-matrices encoding C1′−C1′ distances in each DNA monomer, a C1′−C1′ off-diagonal block relating the distances (d(C1′−C1′)) between the two DNA strands (including the antidiagonal feature characteristic of closely paired nucleotides), and two off-diagonal blocks of $d(C\alpha-C1')$ distances encoding the distance of each protein residue to each nucleotide residue.

library structure asymmetric unit are computed and stored separately using a custom binary file format that reduces the computational and I/O load of a search. A MySQL relational database is used to store the names of searchable structures (including privately accessible user-contributed structures), lists of residues, chain identifiers and corresponding amino acid sequences (as annotated by the SEQRES records in the PDB file), and chain−chain interaction metadata that can be rapidly extracted to direct and control substructure searching. A more detailed description of the database is provided in the Supporting Information Computational Methods section.

Distance geometry between all residues in all chains of some PDB structures can to be large, particularly when large oligomers with many chains in the asymmetric unit are involved. To reduce the storage needed to represent structures, we store the distance geometry in blocks as illustrated in Figure 3, with separated intramolecular and intermolecular components. The unique triangular matrix of distances representing each polymer chain is computed. Each such matrix is comparable to the large triangular block representing a "library structure" in Figure 1. Distances describing the interaction geometry between one chain and another (rectangular matrices like those in Figure 2) are also computed, but the entire matrix representing the interaction geometry is stored in the database only if some resulting distances show that the chains are in contact (at least two Cα atoms closer than 8 Å; C1′−C1′ distances closer than 14 Å). The discarded distance geometry relates chains that are so far apart that the relationship between them is unlikely to ever be of any interest anyway. By this practice, we reduce storage and I/O by as much as 80% for many structures but still have access to a complete distance geometry representation when needed for queries involving multiple molecules.

Searchable Subsets. Our searchable database currently includes the asymmetric units of over 88 000 PDB structures. Included in this collection is 62 000 distinct protein sequences, 5500 unique DNA and RNA sequences, 194 000 protein chains in contact, and 28 000 examples of protein−DNA interactions. Software exists to update this collection at regular intervals as the PDB is updated. While the database includes all structures, in many cases it may not be necessary to search them all. The PDB is well-known to contain an uneven sampling of protein structure. The same structure often appears repeatedly in the database, differing only by crystal form, a single amino acid substitution, or in being cocrystallized with a different substrate analog or inhibitor. When working to identify appropriate target substructures, it is sometimes helpful to search only a few specific structures known to be of interest. It is important therefore to provide a mechanism to permit the search of database subsets that are likely to contain what is sought, but without the burden of excessive multiplicity. Our search method is designed so that any SQL query that yields a set of structure names and chains can be used as a searchable subset. Therefore, it is trivial to search only protein structures, only structures with RNA/DNA, or only complexes with both. One such subset ("Benchmark 2012" representing all structures in the database prior to December 18, 2012) is used for performance benchmarking (Table 1). It includes about 10% of all PDB structures. One can easily search only single polypeptide chains or only intermolecular complexes. Searches can be confined to a user-provided subset of PDB-ids, a predefined set of 4287 unique folds with less than 25% sequence identity (PDB-select25)[47] or a larger subset of nonredundant proteins (26,890 structures) unique at 90% sequence identity (PDB-select90).[47] Because we still find that searching very large structures consumes a disproportionately large fraction of search resources, there is a subset of "Non-Big" structures that include fewer than 10 chains. While it has not yet been implemented, we envision a future ability to descend outward from hits obtained in searching a small subset of user-selected PDB-ids (e.g., unique folds), to include all other structures with sequence similarity (better than some specified Blast E-cutoff) to chains found to match the target.

Target Attributes. To define the kinds of substructures that are sought in searching, we require a specification of the target substructure: a selection of residues from a PDB formatted coordinate file or library (PDB) structure. The target substructure (or target) may include any number of segments of bonded residues of any length and may include segments from different macromolecules in the structure of a complex. The target identifies both the molecular pattern of connected residues we seek and the geometry that relates these residues in space. The molecular pattern specifies the first residue and the length (number of residues) of each connected segment of the target substructure and may be used to constrain different segments to exist as part of the same polypeptide chain. Matching substructures must retain this pattern exactly. Each target segment may appear anywhere in a library structure, although a constraint may be specified to require that any target segment follow the previous segment (i.e., be "joined") in the sequence of the chain. The number of residues in this connection is not restrained. When one segment must precede another in the same polypeptide chain, or if two segments must exist in different chains, these restraints are incorporated as heuristics to conduct the search more

dx.doi.org/10.1021/ci4002537 | J. Chem. Inf. Model. 2013, 53, 2634−2647

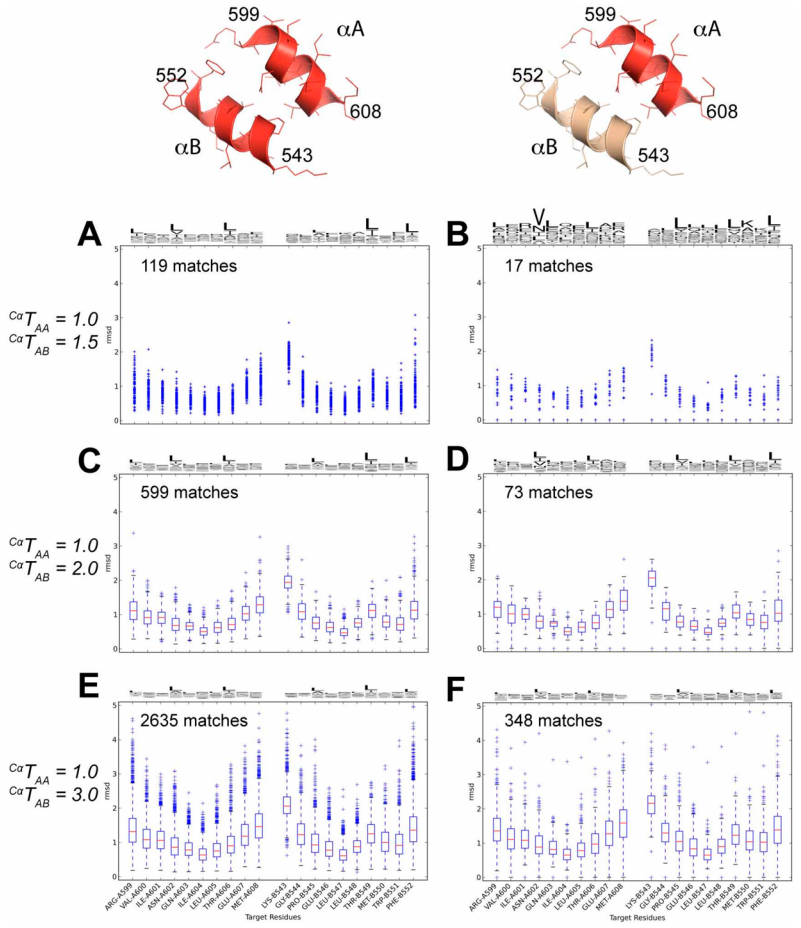


Figure 4. Helix−helix search results. Target characteristics are defined in the text. Plotted in the RMS deviation (Å) for each residue of each helix−helix pair following superposition of all backbone atoms. (left) Results from searches where both strands are a part of the same polypetide chain. (right) Results obtained where one helix (aA) is extracted from one chain and aB from another. Distance geometry tolerances applied are shown far left. The lower box plots illustrate the median RMSD for all matches (red). The box encloses 50% of all RMSD values, and the bars extend to 1.5 times the spread of the box. Outliers are plotted individually. Each graph includes the self-match, since the target structure (3cf0) is part of the searchable subset of library structures.

efficiently and return only matches that retain these characteristics.

The other important component of the target defines how these residues are positioned relative to each other in space. Distance geometry calculated from the coordinates of the geometry atom (Cα or C1′) from target residues defines how these residues are positioned relative to each other in space. As with our original implementation of difference geometry searching,[43] we associate a single boolean value (a "geometry mask") with each target residue so that the geometry of masked residues can be ignored, without removing the connectivity constraint imposed by the molecular pattern of residues. This masking makes it convenient to find, for example, loops of a fixed length and fixed end-points, without restraining the geometry. Geometry masking permits the introduction of geometrical flexibility and provides a way to allow more diverse substructures to be found. Additional flexibility can be introduced by declaring a target with multiple segments joined by loops of any length.

Software Implementation. A searchable database representing the distance geometry of the PDB occupies 64 GB of memory, and the uncompressed PDB files themselves require an additional 66 GB of memory. At the present, a data set requiring about 130 GB of disk space is too large to realistically distribute to individual users on mobile devices and too difficult to keep updated in multiple locations. The substructure searching software we have implemented therefore facilitates searches of a centralized database through a web service and a web application. The search method that has been described previously[38] has been updated to increase computational efficiency; it now exists as a combination of Python (v2.6+) and C++ program modules. The DGM algorithm has been reimplemented in a more modular manner to better facilitate search speed optimization, to separate computational code from user-interface code, and to permit future distribution of search load to multiple servers and parallel processors. A key-value store (Redis; http://redis.io) serves as a network-accessible shared-memory space to store temporary data; the more permanent data is stored in a relational database (MySQL).

The primary human−computer interface to the substructure searching methods is a Python Web application that uses the Flask Web framework (http://flask.pocoo.org/). The Web application code runs as its own network process, and it provides Web pages to define target substructures, conduct substructure searches, view search results, download the search results, and download structural hits that are aligned to the target substructure. The substructure search results page

dx.doi.org/10.1021/ci4002537 | J. Chem. Inf. Model. 2013, 53, 2634−2647

Table 2. Helix−Helix Search Results

search as shown in Figure 3	${}^{C\alpha}T_{AA}$ (Å)	${}^{C\alpha}T_{AB}$ (Å)	molecular pattern	matching substructures	unique sequence matches[b]	RMSD range (Å)[c]	search time (s)[d]
A	1.0	1.5	AA	286	119	0.62−1.49	12
B	1.0	1.5	AB	41	17	0.72−1.07	6
C	1.0	2.0	AA	2020	599	0.62−1.73	24
D	1.0	2.0	AB	228	73	0.68−1.43	10
not shown	1.0	2.5	AA	4330	1241	0.62−2.08	28
not shown	1.0	2.5	AB	542	161	0.68−2.20	16
not shown	1.0	3.0	AA	5000[e]	2020	0.62−2.73	35
not shown	1.0	3.0	AB	892	275	0.68−3.86	14
E	1.5	3.0	AA	5000[e]	2635	0.62−2.98	40
F	1.5	3.0	AB	1305	348	0.68−3.86	17

[a]All results obtained with target substructure from PDB id 3cf0, residues A599−A608 and B543−B552. The searchable library subset "Benchmark 2012" included 9418 structures with one chain (pattern AA) and 5314 structures with two or more chains patterns (AB). A total of 22 851 chains were searched for AA matches. 19 143 pairs of interacting chains were searched for the AB matches. [b]Only the one match with lowest RMSD is counted for each unique amino acid sequence combination. [c]Backbone RMSD computed from aligned residues only (N, CA, CB, C, O). Provided range excludes the self-match [d]Wall-clock time crudely based on web-interface response times and the return of unique sequence matches. [e]Matches returned by the user interface for any search are limited to 5000.

periodically updates the status of in-progress searches, and users can request to update the list of current search results. After a search has been completed, users may retrieve and download matching substructures or intact transformed PDB with identified substructures overlaid on the target. Providing an easy to use web application interface to the substructure searching programs and the ability to download selected structural hits and records gives users significant power.

RESULTS AND DISCUSSION

Algorithm Performance. The Web-based user interface provides access to searching services, and a searchable database populated with the entire PDB has allowed us to gain practical experience with the methods presented here. Search times vary dramatically depending on the nature of the query, and the PDB subset being searched, and (because of operating system caching) the recent search history. In general, queries involving large irregular substructures complete much faster than those involving small regular substructures. Most searches of all structures in our database can be completed in less than 10 min.

Examples of search times are provided. A search of the 4287 unique single-domain structures of the PDB Select 25 subset for the 6 helical segments conforming to the molecular pattern of the a prototypical ankyrin repeat cluster (see below) completes in about 5 s. A search of all PDB structures (about 85 000) for this same target takes just over 2 min and returns 134 unique hits. Searching of all protein−polynucleotide complex structures (about 3250) for the helix−DNA intermolecular pattern illustrated in Figure 6B (below), returns 62 unique hits in about 10 s.

Simple Example of a Substructure Search. A principal impetus for expanding the searchable database to include chain−chain contact geometry was to be able to answer the question "Are there other *complexes* that look like this?". Two complexes might be similar if interacting partners utilize a similar backbone scaffold geometry similarly positioned in space.

To illustrate capabilities of substructure searching and to evaluate search algorithm performance, we consider a simple example of two antiparallel helices in contact (Figure 4). Many different kinds of contacts fit this general description, but we have chosen one specific contact—two ten-residue segments extracted from the monomer−monomer interface in AAA-ATPase as a test case.[48] While this example involves two common elements of secondary structure, this need not be the case. The algorithm is equally applicable to any conformation, and to substructures involving more segments (as will be shown). This example was chosen because the interaction motif is relatively common. Searches return many complex structures involving the same scaffold geometry.

The target geometry for this example consists of residues 599−608 of chain A and residues 543−552 of chain B of PDB structure 3cf0 (Figure 4). Searches may be constrained to find only other matches where the two segments are in the same chain (i.e., intramolecular matches) or to find only matches where the two segments are found in different chains (intermolecular matches), by defining the target molecular pattern as either "AA" or "AB", respectively. We compare results from both kinds of searches. The interface also permits a search for all matches, regardless of chain of origin. No sequence filtering or residue masking was imposed in this example.

The target definition requires specification of tolerances for intrasegment distances, and intersegment distances (${}^{C\alpha}T_{AA}$, and ${}^{C\alpha}T_{AB}$, respectively), which were varied in different searches. After identifying segments with matching distance geometry, backbone atomic coordinates of atoms in these residues are aligned to the corresponding target atoms, and the overall RMSD is computed to rank-order and filter matches. For this example, no matches were excluded based on RMSD; the distance geometry provided the only match selection criteria. Matches were filtered to select only the best match of each unique amino acid sequence (as judged by RMSD), to remove redundant matches. (This filtering is a feature of the user interface).

Results and algorithm performance of several searches of the database (accessed April 25, 2013) are summarized in Table 2 and Figure 4, where results of intramolecular (pattern AA) and intermolecular searches (pattern AB) using progressively higher tolerances are compared. Plotted is the RMSD in the position of backbone atoms (N, CA, CB, C, O) for each residue of each match following superposition on the target residues. Also shown, is the sequence logo representing the ensemble of nonredundant matches. A more detailed algorithm performance assessment and dependence on physical memory and parallelization has been presented elsewhere.[49]

These results clearly show that the algorithm performs efficiently to identify similarly oriented helices irrespective of protein molecule of origin. Matches identified in separate cohorts (AA vs AB) retain similar characteristics, even in the amphipathic sequence preference for leucine in the helix pair interior, whether the matches originate in one molecule, or from the interaction of two. The results also illustrate the effect of increasing tolerances on both the number of matches identified, and the quality. The internal segment tolerance (${}^{C\alpha}T_{AA}$) of either 1.0 or 1.5 Å imposed over 10 connected residues imposes a strong constraint on the geometry of either segment to be helical. It does not constrain the geometry at the ends of each segment very effectively, so there is a sharp rise in residue RMSD at segment termini. The intersegment tolerance (${}^{C\alpha}T_{AB}$) constrains the two helices to be oriented similarly with respect to each other. When ${}^{C\alpha}T_{AB}$ is small (1.5−2.0), the resulting overall RMSD of matched substructures is also small, reflecting truly similar helix pairings. As it is increased, however, there is a variety of possible ways that more diverse structures can be accepted as matches, including examples that might reflect either an erosion of the helical geometry (typically at the ends), or a change in the angle relating the two helices. The matches with largest RMSD arise from rotation of one helix around an axis. In this case, high residue RMSDs are evenly distributed across the entire residue range; all residues fit badly. (In the case of the worst-fitting "match" in the AB collection, with overall RMSD 3.8 Å, a realignment of the matching substructure using only helical segment αA shows that αB is rotated ∼100° around αA from the target αB position. We would not typically consider this a geometrical match and routinely reject all hits with poor RMSD (>2.0 Å). The use of a ${}^{C\alpha}T_{AB}$ tolerance greater than 3 Å is never recommended, because misalignment to the wrong Cα can result.

Example Applications. Just as it is usually helpful to examine an ensemble of different protein sequences in order to identify important sequence features, it is also important to consider ensembles of related protein structures when looking to identify important structural features. Methods which contribute to the identification of structures that embed similar folds like DALI,[13,14] VAST,[15] or SCOP[19] can be extremely valuable to help identify similar molecules for comparison, but not all meaningful similarity extends to entire folds. While there are powerful tools for overlaying similar protein structures,[16−18] users must know that the similarity exists to use them. Moreover, none of these tools specifically allows for the identification and manipulation of similar structures that span opposing molecules involved in macromolecular complexes. A primary strength of our approach to substructure searching lies in its generality. There need be no prior recognition that a particular geometry is unique or common. There is no need to annotate it, name it, or describe it. With only atomic coordinates from a single example of a particular substructure or substructure assembly, it is possible to identify other structures that include the same geometry and, then, to overlay the similar structures in a way that emphasizes this likeness.

Likeness can exist in protein structures at many levels. It can extend to entire folds or be localized to only a small region of a structure. Likeness can be trivial, or very important. Interestingly, one of the best known examples of important likeness—that of the position of the catalytic triad in serine proteases—extends to only three conserved amino acid residues. We would posit that there is no one best way to explore likeness in macromolecular structures. The questions to be addressed in the comparison of any two structures are too varied. It is, however, up to the user of searching services to identify candidate substructures that are meaningful in the context of their structures of interest. By exploring the presence or absence of likeness in many different ways through repeated geometry-based searching, we can learn a great deal about how proteins function, what makes different proteins unique, and how they have evolved.

In this discussion we present three specific examples that illustrate how "complex substructure searching" (that is, searching for multiple substructures that share a common geometry of interaction with each other) can be used to understand the essential features of macromolecular structure. With the first, we use our method to identify otherwise unrelated proteins that contain a serine protease catalytic triad. It is meant as a simple confirmation that the method works as expected and as a means to compare our results to those observed with others using comparable methods. The second example involving ankyrins is an application involving repetitive protein substructures that exist within the *same protein fold*. It serves to illustrate how sensitive backbone distance geometry can be for the detection of the presence or absence of specific structural features, and how searches of the entire PDB can be helpful in making generalizations regarding the importance of these features to biological function. The third example involving motifs of protein−DNA interaction capitalizes on the unique ability of our software to perform geometrical searching of a mixed database of distance geometry encoding *protein, polynucleotide, and intermolecular interaction geometries*.

Serine Protease Catalytic Triad. Substructure searching tools utilizing distance geometry matching have been previously described.[8,36] SPASM[36] can be used to search a preconfigured library of structures for any arbitrary constellation of amino acid residues positioned in space in a manner that resembles a target constellation. The library consists of the positions of the alpha-carbon and the centroid of side chain atoms, and a query may be constructed to require correspondence of one, the other, or both of these geometrical. PINTS[8] provides a similar search capability, but it may be used to conduct only pairwise searches unless the user seeks to find similarity in a constellation of previously annotated residues of functional significance. Since these existing tools both invoke the example of a serine protease catalytic triad as an example of a substructure that can be located using distance-geometry-based searching, we explored this as a potential application of our searching service.

A target was selected comprised of the three catalytic residues of Trypsin-like blood coagulation factor VIIa (PDB id 2ec9; His-57, Asp-102, Ser-195). Alpha-carbon coordinates define the distance between these one-residue "segments", and a sequence filter was imposed to select only substructures exactly matching these three amino acids. A search of the unique folds in our database (4294 structures) with tolerance (${}^{C\alpha}T_{AB}$ = 1.2 Å) identifies nine structures in this sequence-diverse collection that can align to all atoms in these three residues with an RMSD of less than 1.0 Å. The search takes about 8 s. A search of the subset of ∼24 000 nonredundant PDB structures (PDB_Select90 subset) identifies 115 proteins in less than one minute. All of these are annotated as serine proteases. With the same target and tolerances, we also identify several proteins that are subtilisin-like proteases, but to find more subtilisin-like proteases that have a different orientation of the backbone atoms in these catalytic residues, it is better to use as a target coordinates from one of the subtilisin structures.

dx.doi.org/10.1021/ci4002537 | J. Chem. Inf. Model. 2013, 53, 2634−2647

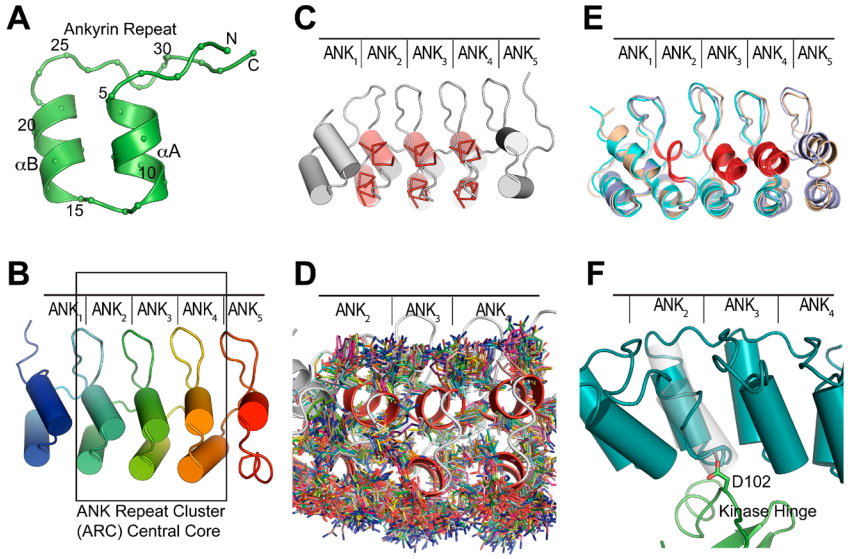


Figure 5. Substructure-based manipulation of ankyrin repeats. (A) Prototypical Ankyrin Repeat of 33-residues. Coordinates shown are residues 502−534 from ANK_{16} of AnkyrinR (PDB id 1n11).[57] (B) Representative ANK repeat cluster (ARC) domain incorporates from 4 to 5 ANK repeats. Shown is a designed ANK repeat protein (DARPin) prepared as a caspase inhibitor (PDB id 2p2c).[76] The box encloses the regular repeats of the ARC central core. (C) Six repeat cluster core helical segments from ankyrinR (PDB id 1n11) used to represent the ARC central core in PDB searching are highlighted in red (residues 440−446 and 451−458 from ANK_{14}, residues 473−479 and 484−491 from ANK_{15}, and residues 506−512 and 517−524 from ANK_{16}). (D) Best structural alignment resulting from 100 unique sequences with structural likeness to the target of C. All substructures align with an RMSD of backbone atoms < 1.0 Å. Only side chain atoms are shown, as the backbone is constrained to be equivalent. (E) p16INK4 kinase inhibitor structures aligned using the three segments shown in red from ANK_{2-4}. Structures aligned include 1bi7 (used as the target; cyan), 1bi8 (violet), and 1g3n (sand). All homologues include an atypical ANK repeat in place of ANK_2. (F) Structure of the complex of CDK6 and p16INK4A.[60] The partially transparent display of a typical ANK repeat geometry (white) is overlaid on ANK_2 to illustrate where it would lie in relation to the atypical ANK2 of the INK4 proteins. The helix would extend to overlap the side chain of CDK6 residue 102.

Ankyrins. A varied collection of proteins include ankyrin (ANK) repeats.[50−52] These ∼33-residue structural motifs include an α helix (A), a two-residue turn leading into an antiparallel helix (B) that stacks beside the first, and loops that extend outward at right angles before and after the helices to give the repeating motif an L-shape (Figure 5A). Successive repeats assemble as localized 4-α-helical bundles with helices A and B of one repeat packed beside helices A and B of the previous repeat. ANK repeats often form clusters of 4−5 repeats (termed "Ankyrin Repeat Clusters" or ARCs) with specific protein recognition functions[53] (Figure 5B). As a class, ARCs participate in protein−protein interactions important to cell cycle regulation, transcription regulation, cytoskeletal localization, and development. Although there is a distinctive and recognizable pattern to the sequences of ankyrin repeats (reviewed by Mosavi et al.[54]), some proteins incorporate ANK repeats with extensive structural diversity.[55] Even in these more extreme examples, the structures share elements of considerable likeness.

The ANK sequence motif generally appears only as a repeat in larger structures. The prototypical ANK-repeat geometry is most commonly found in the center of an ARC; the ARC-capping repeats tend to exhibit more sequence and conformational diversity and have different sequence attributes.[56] Proteins that incorporate a large number of ANK repeats, such as the 24-repeat ankyrinR tend to better recapitulate the canonical geometry.[57] This implies that the conformation of each repeat is determined as much by its surroundings, as it is by the sequence of the repeat itself. Our substructure searching methods can be used to identify and compare individual motifs such as these that exist within a larger molecular context of similarity and, thereby, permit identification of consensus structural determinants. Significant deviations from this consensus are likely to impact structural topology and, therefore, biological function.

Approximately 120 protein structures in the PDB include recognizable ANK repeats (by PSI-Blast[5]). A search of the PDB for structures that share complete geometrical likeness to an intact prototypical ANK repeat, represented by the ANK substructure in Figure 5A, identifies only 56 intact repeats that are actually "alike" (as identified with a maximum $\Delta C\alpha$ difference tolerance ${}^{C\alpha}T_{AA}$ of 1.5 Å) through all 33 residues of the repeat. This is a very small fraction of all repeats present in PDB structures. If the search is conducted imposing a restraint on only the residues of the helix-turn-helix motif and not the trailing loop (residues 4−23), nearly 500 examples with RMSD < 1.0 Å are found. The helix-turn-helix motif, even as defined with the precise loop and helix orientations of the ANK prototype, is not unique to ANK repeats, and can be found in dozens of other proteins. A more defining characteristic of the ARC proteins is the highly regular and parallel helical bundles that arise from the stacking of successive repeats.[57] A substructure composed of six short helical segments from three consecutive repeats of ankyrinR (Figure 5C) can be regarded as representative of the more regular ARC central core. A search with this target identifies only ANK-repeat proteins. Even among these proteins there is variation in structure, and the search returns only highly conforming core structures (Figure 5D). The list of aligned structures incorporating 90 unique sequences, along with the aligned sequences of each central ANK repeat, constitutes an excellent data set for an analysis to determine consensus sequence requirements for ARC proteins, such as has been provided previously based on a much less comprehensive subset of

dx.doi.org/10.1021/ci4002537 | J. Chem. Inf. Model. 2013, 53, 2634−2647

structures.[54] A complete list of matching substructures and sequence alignments is provided as Supporting Information (Supplemental Table S1).

It is worth noting that our procedure identifies no other structures that have both topological and geometrical similarity to the anykrins. Within the tolerances set for "likeness" ($^{C\alpha}T_{AA}$ = 1.5 Å, and $^{C\alpha}T_{AB}$ = 3.0 Å), the ankyrin core geometry is unique. Structures such as the TOG domains[58] are also comprised of regular sequence repeats with a similar helix-turn-helix topology (so-called "HEAT" repeats[59]), but the geometry of the assembly of these proteins is quite distinct from those with ANK repeats (and far more irregular). TOG domains share little conformational likeness with the ankyrins.

The INK4 cyclin dependent kinase (CDK) inhibitors also do not conform to this core structure, although they can clearly be classified as ankyrin-family ARCs.[60] A single amino acid deletion occurs in the second ANK repeat of each of these proteins, conferring an irregular nonhelical structure for ANK_2 helix A (Figure 5E). In other respects, the INK4 structures do conform, and the INK4 proteins can be considered to be "like" other ARCs if the first helix is omitted from the ARC central core of Figure 5C. A search using a substructure from p16INK4a (PDB id 1bi7)[60] that includes this irregular helix (residues 48−53) and more normal ANK A-helices from the remainder of the ARC central core (residues 81−88 and 115−122) produces matches to all other known INK4 homologue structures (Figure 5E),[60−62] but to none of the other ANK proteins, confirming that this distinctive structural feature is both conserved within the functional class and unique among ARC domains. When the complexes of INK4 proteins and their CDK partners are overlaid on other ARC proteins, it becomes clear why the deletion in ANK_2 is necessary for INK4s to bind the kinases. The irregular ANK_2 backbone forms specific hydrogen bonds with an aspartic acid carboxylate conserved in the CDK hinge and exposed on the kinase molecular surface. ANK proteins with typical repeat geometry in place of ANK_2 would bump this kinase structural feature, explaining why other ANK proteins do not inhibit cyclin-dependent kinases (Figure 5F).

Transcription Activators. As discussed in the Introduction, it is well-established that proteins that regulate gene transcription often employ similar substructural motifs in interactions with DNA but that even similar folds can bind DNA in different ways.[1,2] Because our database includes a representation of distance geometry of DNA conformation in such complexes, searching services can provide a means to readily identify ensembles of complexes that share a common interaction framework, while excluding those complexes that do not.

As an illustrative example, we use one of several crystal structures of ETS transcription activators classified as winged helix-turn-helix domains that have been determined in complex with double-stranded DNA.[63−66] In these complexes, an α helix (H3) lies across the DNA major groove (Figure 6A and D) and contributes a number of base-pair-specific interactions that stabilize the complexes.[65,67] To identify other unrelated complexes that share this same interaction motif, a search of all protein−nucleic acid complexes in our database was initiated, using a short DNA duplex of seven base pairs and bound helical 11-mer from one such structure (PDB id 1awc)[63] as a target substructure geometry (Figure 6B). In addition to other ETS transcription activators, the query identifies two other protein families that make locally identical interactions with DNA (RMSD < 2.0 Å): the C-terminal domain of

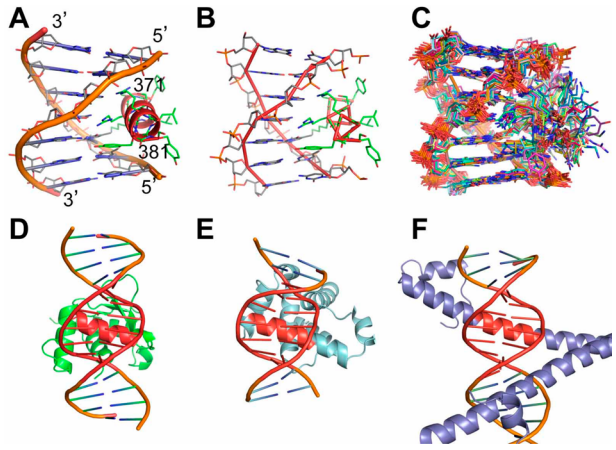


Figure 6. Diverse folds exploit the same helix orientation in the major groove. (A) Duplex DNAand bound helical substructure from the Ets transcriptional activator/DNA complex of Batchelor et al.[63] (PDB id 1awc). The dominant component of the interaction is the helix (residues 371−381) lying across the DNA major groove. (B) Substructure search target geometry comprising C1′ atoms in two complementary strands of DNA (residues D6−D12 and E32−E38) and Cα atoms of the protein substructure (residues 371−381). (C) Ensemble of hits with RMSD < 2.0 Å (protein and DNA backbone atoms) resulting from a search of all protein:nucleic acid complexes in the PDB with likeness to the target. Tolerances used were $^{C\alpha}T_{AA}$ = 1.0; $^{C1'}T_{AA}$ = 2.0; $^{C1'}T_{AB}$ = 4.0; $^{C1'C\alpha}T_{AB}$ = 3.0). (D−F) Sample of diverse proteins that incorporate this interaction motif. Some portions of these complexes are not shown for clarity. (D) Ets/DNA complex (PDB id 1awc) from which the target geometry was drawn. The targeted substructures are red. (E) The SpoA0 transcription factor (PDB id 1lq1).[68] The matching substructure (red) RMSD is 1.6 Å. (F) One of many members of the basic region leucine zipper (bZIP) proteins matching the target geometry (PDB id 2wty). Matching substructure (red) RMSD is 1.2 Å.

sporulation protein A[68] (Figure 6E), and a large family of basic region leucine zipper (bZIP) transcription factors (Figure 6F). While the protein helix in these cases is a component of very different protein folds, there is good local structural likeness in the helix−DNA geometry. The ETS domains have high DNA selectivity for the 5′-GGA(A/T)-3′ binding site. The bZIP family members recognize a variety of DNA sequences.[69] The alignment of structural homologues permits a one-to-one assignment of residues in ETS proteins that contribute to DNA recognition with their counterparts in the other proteins, thereby vastly expanding the structure−activity correspondence in all these protein families. A complete list of the helical segment and DNA sequences matching this query is included as Supporting Information Table S2.

The geometry embodied in this particular interaction is distinctly different from interactions made by classic helix-turn-helix RNA binding motifs,[70] Zn-finger motifs,[71] or homeodomains[72] that also incorporate a topological feature that can be described as a helix in the major groove. Pabo and Nekludova noted that an alignment of helical axes does not necessarily result in a spatial alignment of the residues in the helices.[1] Distance geometry matching provides a sensitive means to identify only those complexes where the unique orientation of the helix in relation to the DNA is truly alike; the query above cannot match the alternative interaction geometries because the relational distance geometry between different macromolecular chains does not match. Different queries can be used to identify complexes that share likeness to

dx.doi.org/10.1021/ci4002537 | J. Chem. Inf. Model. 2013, 53, 2634−2647

each of these alternate binding interactions, and motifs can be compared one to another by overlaying only the appropriate DNA substructures (Supporting Information Figure S2).

Conserved motifs can also bind DNA in different ways. DGM can be used to select only like complexes. The helix-turn-helix (HTH) motif that is part of homeodomains[72] provides an excellent example. While there are variations on the length of the connecting loop and the lengths and relative orientations of the two helices in many HTH motifs, the example illustrated in Figure 7A is typical. A substructure search of nonredundant protein structures (PDB_select90) conducted using the segment shown (residues 96−124 of homeodomain MAT A1; PDB id 1yrn[73]) identifies 216 structures embedding a motif with the same HTH geometry (RMSD of backbone atoms < 1.0 Å). In selecting these substructures, the geometry of only residues 100−119 (red) are constrained ($^{C\alpha}T_{AA}$ = 1.5 Å). The extended motif of 29 residues is manipulated only for comparison of sequences. The sequences of these motifs shows little conservation outside of the small hydrophobic residues prevalent at canonical positions 8 (Ala) and 16 (Val) needed to allow the close approach of the helices, and the preference (but not requirement) for small turn-generating residues at positions 13 (Gly) and 15 (Ser) (Figure 7G). The motif is found in many proteins that are not related homeodomains. Examples culled from structures of protein−DNA complexes are shown in Figure 7B. Some of these motifs form very different complexes with the DNA. Two examples are illustrated in Figures 7C/D and E/F, where the structures are overlaid using the DNA to which they are bound, to emphasize the difference in relative HTH position (red).

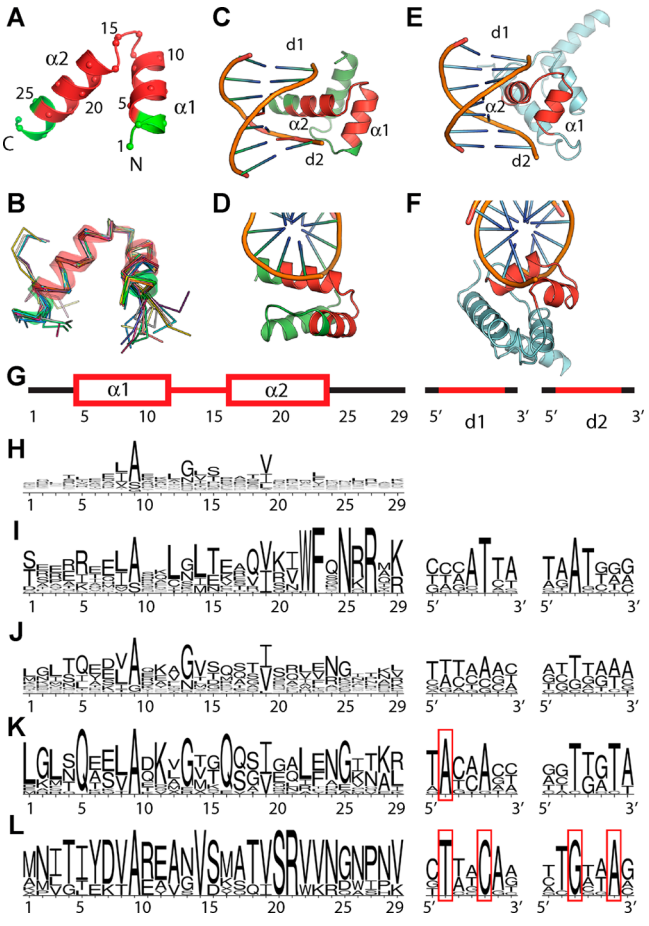


Figure 7. Alternative HTH binding modes. (A) An example HTH motif used as the target geometry for substructure searching. Residues highlighted in red defined the target geometry. (B) 27 substructures with unique sequences selected from the PDB subset of protein−polynucleotide complexes. (C−F) Two geometrically distinct examples HTH/DNA interactions. HTH residues matching the geometry of A are red. (C/D) A typical homeobox transcription factor. (E/F) Lambda repressor-like DNA-binding domains. (G) Residues in red identify geometrically constrained target residues relative to a canonical sequence position. (H) Sequence logo derived from all 216 unique sequences in the PDB that include structures matching the HTH motif of A. (I) Sequence logo from 32 structures matching the interactions like C/D. Five base pairs surrounding the recognition helix were included in the substructure query. (J) Sequence logo from 40 structures matching the interaction like that of E/F. Sequence logos K and L represent a subset of structures in J (17 and 10 structures, respectively) that have been filtered to include only specific nucleobases at the indicated positions (highlighted with red squares) in the bound DNA.

A search of all protein−nucleotide complexes can be conducted to find examples of each type of interaction by including DNA base pairs from one such complex as part of the target geometry. We elected to use a seven base−pair target, but only the five nearest the HTH were used to impose geometry constraints (red residues in Figure 7G). Distance geometry tolerances used were $^{C\alpha}T_{AA}$ = 1.5 Å; $^{C\alpha}T_{AB}$ = 3.0 Å; $^{C1'}T_{AA}$ = 2.0 Å; $^{C1'}T_{AB}$ = 4.0 Å; $^{C1'C\alpha}T_{AB}$ = 3.5 Å. All matches had overall backbone RMSD < 1.5 Å vs the original target. For complexes of the type shown in Figure 7C, the target chosen from 1yrn[73] included protein residues 96−124 and DNA residues C15−C21 and D24−D30. The distinct sequences of 32 matching complexes are shown in the sequence logo of Figure 7I. For complexes of the type shown in Figure 7E, the target came from 1lmb[74] (protein residues 29−57; DNA residues 3−9 and 33−39). The distinct sequences of 40 matching complexes are shown in the sequence logo of Figure 7J.

Only complexes with other annotated homeodomains (pfam PF00046 or PF05920) match the geometry shown in Figure 7C. This geometry is quite distinctive. The sequence homology is strongest at the C-terminal end where the B helix makes extensive contacts with the DNA major groove (Figure 7I). Our query did not constrain the geometry of this C-terminal end of this helix, but all the complexes found have the same geometry anyway. The sequences matching the geometry of interactions like that shown in Figure 7E are more diverse. This is, however, a composite of several more highly conserved domain sequences that can be separated by filtering for particular DNA sequence characteristics. Complexes with an adenosine at the 5′+2 position (the 16 unique complexes of Figure 7K) highly conserve glutamine residues at positions 16 and 5. Gln-16 interacts directly with the Ade base, and the Gln at position 5 serves to hold Gln-16 in place. Most of the proteins in this subset are annotated as pfam HTH_3 (PF01381). The dominant sequence motif of Figure 7L (selected with 5′+2 T and 5′+5 C) are Lac1 family regulatory proteins (PF00356), but it also includes at least one structure (1zs4) annotated as Bacteriophage CII protein (PF05269). Ten structures share this geometry. This contact geometry primarily involves residues in the turn between helices, and side chains emanating from the loop following the end of α2, as reflected in the sequence conservation.

These examples illustrate the striking differences that can exist between different complexes that incorporate a very

dx.doi.org/10.1021/ci4002537 | J. Chem. Inf. Model. 2013, 53, 2634−2647

similar structural motif, and how substructure searching may be used to isolate only those complexes that are comparable. It has been a major source of confusion in the literature that many distinct interaction geometries are often annotated using the same words ("helix-turn-helix", "helix-loop"), when in fact the complexes show no positional equivalence with respect to DNA binding. Distance geometry matching as implemented here provides the means to identify interactions that are truly alike.

CONCLUSIONS

We describe a database and software for querying macromolecular structures to identify other structures that possess backbone geometrical likeness to a complex target substructure. A Web interface to the search tools (https://drugsite.msi.umn.edu/searches) provides a convenient means to look for assemblies of substructures of any length or conformation. The searchable database includes all experimental complexes in the worldwide Protein Data Bank, including all X-ray, NMR, and EM-derived structures for which either alpha-carbon or nucleotide C1′ atomic positions are deposited. Structural likeness may be found whether it exists as part a single folded polypetide, polynucleotide or complex, or quaternary assembly, and atomic coordinates may be overlaid to emphasize this likeness.

Our searchable database currently encodes crystallographic asymmetric units rather than biological units—the molecules or assembly that is presumed to function together in biological systems. Asymmetric units may contain fewer or more molecules than the biological unit. Ideally, we believe that there is value in being able to search the interactions of both, but a biological unit database has not yet been implemented. Such a database might constitute a valuable next step in search service development.

We have presented several specific examples of complex substructure searching using diverse types of substructure targets. The brief examination of ankyrins presented here serves as an example of how layered substructure searching can be used to look at similarity and diversity in a family of proteins in a variety of different ways. These examples only hint at possible applications of our method. Not every observed macromolecular complex or quaternary assembly is built from recurring interaction motifs, but when likeness to another motif does exist, a comparison of like substructure sequences will usually help to define possible constraints on the diversity of structures that can participate in such interactions. These constraints are almost always strengthened when evaluated in the larger context of a protein fold. With our software, any protein−protein or protein−polynucleotide substructures can serve as the subject of a query to determine if the same backbone interaction geometry occurs elsewhere in the structureome and, then, to determine the extent to which other nearby features are also conserved. This is a powerful capability.

In the case of the DNA-binding proteins, there are clearly recurring interaction motifs and distance geometry matching can be used very effectively to identify specific complexes that share a high degree of similarity. The curated database of homeodomain structures[72] (http://research.nhgri.nih.gov/homeodomain/; accessed April 24, 2013), for example, includes proteins that engage in a number of discreet interaction types that can be (and probably should be) subclassified by geometrical likeness. Alignments based solely on observed structural likeness involving both the protein and DNA geometry such as allowed by distance geometry matching, do not depend on the annotation of the structures, so a much greater faith may be placed in the quality and the meaning of resulting sequence segment alignments achieved by distance geometry matching. As more and more complexes are experimentally characterized, we expect the value of our distance geometry database and the ability to search it effectively for like complexes to continue to grow.

ASSOCIATED CONTENT

Supporting Information

Tables S1 and S2, Figures S1 and S2, and more detailed computational methods. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: finze007@umn.edu

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Funding for this work is provided by the Minnesota Department of Employment & Economic Development (SPAP-06-00140P-FY07). The support of the Minnesota Supercomputer Institute is gratefully acknowledged. We thank Emily Wagner for helping the preparation of Supporting Tables S1 and S2.

ABBREVIATIONS

ANK, Ankyrin Repeat; ARC, Ankyrin Repeat Cluster; CDK6, cyclin-dependent kinase-6; DARPin, designed ankyrin repeat protein; DGM, Distance Geometry Matching; ETS, E-Twenty-Six Transcription Factors; HTH, helix-turn-helix; INK4, PDB, Protein Data Bank; RMSD, root mean-square deviation

REFERENCES

(1) Pabo, C. O.; Nekludova, L. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.* **2000**, *301*, 597−624.

(2) Siggers, T. W.; Silkov, A.; Honig, B. Structural alignment of protein−DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol.* **2005**, *345*, 1027−1045.

(3) Keskin, O.; Nussinov, R. Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Eng. Des. Sel.* **2005**, *18*, 11−24.

(4) Tuncbag, N.; Gursoy, A.; Guney, E.; Nussinov, R.; Keskin, O. Architectures and functional coverage of protein-protein interfaces. *J. Mol. Biol.* **2008**, *381*, 785−802.

(5) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.

(6) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(7) Ausiello, G.; Via, A.; Helmer-Citterich, M. Query3d: a new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinf.* **2005**, *6* (Suppl 4), S5.

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

(9) Gherardini, P. F.; Ausiello, G.; Helmer-Citterich, M. Superpose3D: a local structural comparison program that allows for user-defined structure representations. *PLoS ONE* **2010**, *5*, e11988.

(10) Bateman, A.; Birney, E.; Durbin, R.; Eddy, S. R.; Howe, K. L.; Sonnhammer, E. L. The Pfam protein families database. *Nucleic Acids Res.* **2000**, *28*, 263−266.
(11) Rychlewski, L.; Jaroszewski, L.; Li, W.; Godzik, A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **2000**, *9*, 232−241.
(12) Gough, J.; Karplus, K.; Hughey, R.; Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **2001**, *313*, 903−919.
(13) Hulo, N.; Bairoch, A.; Bulliard, V.; Cerutti, L.; Cuche, B. A.; de Castro, E.; Lachaize, C.; Langendijk-Genevaux, P. S.; Sigrist, C. J. A. The 20 years of PROSITE. *Nucleic Acids Res.* **2008**, *36* (Database issue), D245−249.
(14) Holm, L.; Sander, C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **1997**, *25*, 231−234.
(15) Gibrat, J. F.; Madej, T.; Bryant, S. H. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **1996**, *6*, 377−385.
(16) 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), W590−594.
(17) Krissinel, E.; Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.* **2004**, *60*, 2256−2268.
(18) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739−747.
(19) Hubbard, T. J.; Ailey, B.; Brenner, S. E.; Murzin, A. G.; Chothia, C. SCOP: a Structural Classification of Proteins database. *Nucleic Acids Res.* **1999**, *27*, 254−256.
(20) Stein, A.; Céol, A.; Aloy, P. 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* **2011**, *39 (Database issue)*, D718−723.
(21) Davis, F. P.; Sali, A. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* **2005**, *21*, 1901−1907.
(22) Winter, C.; Henschel, A.; Kim, W. K.; Schroeder, M. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.* **2006**, *34* (Database issue), D310−314.
(23) Zhang, Y.; Skolnick, J. The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 1029−1034.
(24) Ye, Y.; Godzik, A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* **2003**, *19* (Suppl 2), ii246−255.
(25) Jones, T. A.; Thirup, S. Using known substructures in protein model building and crystallography. *EMBO J.* **1986**, *5*, 819−822.
(26) Finzel, B. C.; Kimatian, S.; Ohlendorf, D. H.; Wendoloski, J. J.; Levitt, M.; Salemme, F. R. Molecular modeling with substructure libraries derived from known protein structures. In *Crystallographic and Modeling Methods in Molecular Design*; Ealick, S., Bugg, C., Eds; Springer Verlag: New York, NY, 1989; pp 175−189.
(27) McRee, D. E. XtalView/Xfit–A versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **1999**, *125*, 156−165.
(28) Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr, A. Found. Crystallogr.* **1991**, *47*, 110−119.
(29) Laskowski, R.; Macarthur, M.; Moss, D.; Thornton, J. Procheck - A Program To Check The Stereochemical Quality Of Protein Structures. *J. Appl. Crystallogr.* **1993**, *26*, 283−291.
(30) Levitt, M. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **1992**, *226*, 507−533.
(31) Wendoloski, J. J.; Salemme, F. R. PROBIT: a statistical approach to modeling proteins from partial coordinate data using substructure libraries. *J. Mol. Graphics* **1992**, *10*, 124−126.
(32) Sali, A.; Potterton, L.; Yuan, F.; van Vlijmen, H.; Karplus, M. Evaluation of comparative protein modeling by MODELLER. *Proteins* **1995**, *23*, 318−326.
(33) Budowski-Tal, I.; Nov, Y.; Kolodny, R. FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 3481−3486.
(34) Holm, L.; Ouzounis, C.; Sander, C.; Tuparev, G.; Vriend, G. A Database Of Protein-Structure Families With Common Folding Motifs. *Protein Sci.* **1992**, *1*, 1691−1698.
(35) Holm, L.; Kääriäinen, S.; Rosenström, P.; Schenkel, A. Searching protein structure databases with DaliLite v.3. *Bioinformatics* **2008**, *24*, 2780−2781.
(36) Kleywegt, G. J. Recognition of spatial motifs in protein structures. *J. Mol. Biol.* **1999**, *285*, 1887−1897.
(37) Fetrow, J. S.; Skolnick, J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* **1998**, *281*, 949−968.
(38) Finzel, B. C.; Akavaram, R.; Ragipindi, A.; Van Voorst, J. R.; Cahn, M.; Davis, M. E.; Pokross, M. E.; Sheriff, S.; Baldwin, E. T. Conserved core substructures in the overlay of protein-ligand complexes. *J. Chem. Inf. Model.* **2011**, *51*, 1931−1941.
(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) Schneider, T. D.; Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **1990**, *18*, 6097−6100.
(41) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **2004**, *14*, 1188−1190.
(42) O'Sullivan, O.; Suhre, K.; Abergel, C.; Higgins, D. G.; Notredame, C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.* **2004**, *340*, 385−395.
(43) Finzel, B. Mastering the LORE of protein structure. *Acta Crystallogr. D Biol. Crystallogr.* **1995**, *51*, 450−457.
(44) Liu, P.; Agrafiotis, D. K.; Theobald, D. L. Fast determination of the optimal rotational matrix for macromolecular superpositions. *J. Comput. Chem.* **2010**, *31*, 1561−1563.
(45) Kneller, G. R. Comment on "Fast determination of the optimal rotational matrix for macromolecular superpositions". *J. Comput. Chem.* **2011**, 183−186.
(46) Schmeing, T. M.; Huang, K. S.; Kitchen, D. E.; Strobel, S. A.; Steitz, T. A. Structural insights into the roles of water and the 2' hydroxyl of the P site tRNA in the peptidyl transferase reaction. *Mol. Cell* **2005**, *20*, 437−448.
(47) Griep, S.; Hobohm, U. PDBselect 1992−2009 and PDBfilter-select. *Nucleic Acids Res.* **2010**, *38* (Database issue), D318−319.
(48) Davies, J. M.; Brunger, A. T.; Weis, W. I. Improved structures of full-length p97, an AAA ATPase: implications for mechanisms of nucleotide-dependent conformational change. *Structure* **2008**, *16*, 715−726.
(49) Van Voorst, J. R.; Finzel, B. C. Rapid efficient macromolecular substructure searching in a cloud. *Proceedings of the ACM Conference on Bioinformatics, Computational Biology, and Biomedicine*, Orlando, FL, October 7−12, 2012; pp 548−550.
(50) Michaely, P.; Bennett, V. The ANK repeat: a ubiquitous motif involved in macromolecular recognition. *Trends Cell Biol.* **1992**, *2*, 127−129.
(51) Sedgwick, S. G.; Smerdon, S. J. The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends Biochem. Sci.* **1999**, *24*, 311−316.
(52) Li, J.; Mahajan, A.; Tsai, M.-D. Ankyrin repeat: a unique motif mediating protein-protein interactions. *Biochemistry* **2006**, *45*, 15168−15178.
(53) Seimiya, H.; Smith, S. The telomeric poly(ADP-ribose) polymerase, tankyrase 1, contains multiple binding sites for telomeric repeat binding factor 1 (TRF1) and a novel acceptor, 182-kDa

tankyrase-binding protein (TAB182). *J. Biol. Chem.* **2002**, *277*, 14116−14126.
(54) Mosavi, L. K.; Minor, D. L.; Peng, Z.-Y. Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 16029−16034.
(55) Phelps, C. B.; Huang, R. J.; Lishko, P. V.; Wang, R. R.; Gaudet, R. Structural analyses of the ankyrin repeat domain of TRPV6 and related TRPV ion channels. *Biochemistry* **2008**, *47*, 2476−2484.
(56) Tamaskovic, R.; Simon, M.; Stefan, N.; Schwill, M.; Plückthun, A. Designed ankyrin repeat proteins (DARPins) from research to therapy. *Meth. Enzymol.* **2012**, *503*, 101−134.
(57) Michaely, P.; Tomchick, D. R.; Machius, M.; Anderson, R. G. W. Crystal structure of a 12 ANK repeat stack from human ankyrinR. *EMBO J.* **2002**, *21*, 6387−6396.
(58) Slep, K. C. The role of TOG domains in microtubule plus end dynamics. *Biochem. Soc. Trans.* **2009**, *37*, 1002−1006.
(59) Groves, M. R.; Hanlon, N.; Turowski, P.; Hemmings, B. A.; Barford, D. The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs. *Cell* **1999**, *96*, 99−110.
(60) Russo, A. A.; Tong, L.; Lee, J. O.; Jeffrey, P. D.; Pavletich, N. P. Structural basis for inhibition of the cyclin-dependent kinase Cdk6 by the tumour suppressor p16INK4a. *Nature* **1998**, *395*, 237−243.
(61) Jeffrey, P. D.; Tong, L.; Pavletich, N. P. Structural basis of inhibition of CDK-cyclin complexes by INK4 inhibitors. *Genes Dev.* **2000**, *14*, 3115−3125.
(62) Venkataramani, R. N.; MacLachlan, T. K.; Chai, X.; El-Deiry, W. S.; Marmorstein, R. Structure-based design of p18INK4c proteins with increased thermodynamic stability and cell cycle inhibitory activity. *J. Biol. Chem.* **2002**, *277*, 48827−48833.
(63) Batchelor, A. H.; Piper, D. E.; de la Brousse, F. C.; McKnight, S. L.; Wolberger, C. The structure of GABPalpha/beta: an ETS domain-ankyrin repeat heterodimer bound to DNA. *Science* **1998**, *279*, 1037−1041.
(64) Garvie, C. W.; Hagman, J.; Wolberger, C. Structural studies of Ets-1/Pax5 complex formation on DNA. *Mol. Cell* **2001**, *8*, 1267−1276.
(65) Lamber, E. P.; Vanhille, L.; Textor, L. C.; Kachalova, G. S.; Sieweke, M. H.; Wilmanns, M. Regulation of the transcription factor Ets-1 by DNA-mediated homo-dimerization. *EMBO J.* **2008**, *27*, 2006−2017.
(66) Babayeva, N. D.; Baranovskaya, O. I.; Tahirov, T. H. Structural basis of Ets1 cooperative binding to widely separated sites on promoter DNA. *PLoS ONE* **2012**, *7*, e33698.
(67) Buchwalter, G.; Gross, C.; Wasylyk, B. Ets ternary complex transcription factors. *Gene* **2004**, *324*, 1−14.
(68) Zhao, H.; Msadek, T.; Zapf, J.; Madhusudan; Hoch, J. A.; Varughese, K. I. DNA complexed structure of the key transcription factor initiating development in sporulating bacteria. *Structure* **2002**, *10*, 1041−1050.
(69) Fujii, Y.; Shimizu, T.; Toda, T.; Yanagida, M.; Hakoshima, T. Structural basis for the diversity of DNA recognition by bZIP transcription factors. *Nat. Struct. Biol.* **2000**, *7*, 889−893.
(70) Wintjens, R.; Rooman, M. Structural classification of HTH DNA-binding domains and protein-DNA interaction modes. *J. Mol. Biol.* **1996**, *262*, 294−313.
(71) Krishna, S. S.; Majumdar, I.; Grishin, N. V. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.* **2003**, *31*, 532−550.
(72) Moreland, R. T.; Ryan, J. F.; Pan, C.; Baxevanis, A. D. The Homeodomain Resource: a comprehensive collection of sequence, structure, interaction, genomic and functional information on the homeodomain protein family. *Database (Oxford)* **2009**, *2009*, bap004.
(73) Li, T.; Stark, M. R.; Johnson, A. D.; Wolberger, C. Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. *Science* **1995**, *270*, 262−269.
(74) Beamer, L. J.; Pabo, C. O. Refined 1.8 A crystal structure of the lambda repressor-operator complex. *J. Mol. Biol.* **1992**, *227*, 177−196.
(75) Stoll, R.; Lee, B. M.; Debler, E. W.; Laity, J. H.; Wilson, I. A.; Dyson, H. J.; Wright, P. E. Structure of the Wilms tumor suppressor protein zinc finger domain bound to DNA. *J. Mol. Biol.* **2007**, *372*, 1227−1245.
(76) Schweizer, A.; Roschitzki-Voser, H.; Amstutz, P.; Briand, C.; Gulotti-Georgieva, M.; Prenosil, E.; Binz, H. K.; Capitani, G.; Baici, A.; Plückthun, A.; Grütter, M. G. Inhibition of caspase-2 by a designed ankyrin repeat protein: specificity, structure, and inhibition mechanism. *Structure* **2007**, *15*, 625−636.
(77) Finzel, B. LORE: exploiting database of known structures. *Methods Enzymol.* **1997**, *277*, 230−242.