Real-time structural motif searching in proteins using an inverted index strategy

Sebastian Bittrich[1*], Stephen K. Burley[1,2,3,4], Alexander S. Rose[1]

1 RCSB Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA, USA
2 RCSB Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
3 Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA
4 Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, USA

* sebastian.bittrich@rcsb.org

Abstract

Biochemical and biological functions of proteins are the product of both the overall fold of the polypeptide chain, and, typically, structural motifs made up of smaller numbers of amino acids constituting a catalytic center or a binding site. Detection of such structural motifs can provide valuable insights into the function(s) of previously uncharacterized proteins. Technically, this remains an extremely challenging problem because of the size of the Protein Data Bank (PDB) archive. Existing methods depend on a clustering by sequence similarity and can be computationally slow. We have developed a new approach that uses an inverted index strategy capable of analyzing >160,000 PDB structures with unmatched speed. The efficiency of the inverted index method depends critically on identifying the small number of structures containing the query motif and ignoring most of the structures that are irrelevant. Our approach (implemented at motif.rcsb.org) enables real-time retrieval and superposition of structural motifs, either extracted from a reference structure or uploaded by the user. Herein, we describe the method and present five case studies that exemplify its efficacy and speed for analyzing 3D structures of both proteins and nucleic acids.

Author summary

The Protein Data Bank (PDB) provides open access to more than 160,000 three-dimensional structures of proteins, nucleic acids, and biological complexes. Similarities between PDB structures give valuable functional and evolutionary insights but such resemblance may not be evident at sequence or global structure level. Throughout the database, there are recurring structural motifs – groups of modest numbers of residues in proximity that, for example, support catalytic activity. Identification of common structural motifs can unveil subtle similarities between proteins and serve as fingerprints for configurations such as the His-Asp-Ser catalytic triad found in serine proteases or the zinc coordination site found in Zinc Finger DNA-binding domains. We present a highly efficient yet flexible strategy that allows users for the first time to search for arbitrary structural motifs across the entire PDB

archive in real-time. Our approach scales favorably with the increasing number and complexity of deposited structures, and, also, has the potential to be adapted for other applications in a macromolecular context.

Introduction

Within proteins, structural motifs are stereotypical arrangements of amino residues, which may or may not be near one another in the linear polypeptide chain. They can be described using three properties, including spatial proximity, relative arrangement in three-dimensions (3D), and physicochemical properties [1]. Many simple structural motifs contribute directly to the biochemical or biological function of a protein, such as the well-known His-Asp-Ser catalytic triad (Figure 1A) found in serine proteases [2]. More complex structural motifs contribute indirectly to function by acting as binding scaffolds (*e.g.*, five residues coordinating two zinc ions within the active center of bovine lens leucine aminopeptidase, PDB ID 1lap [3], Figure 1B). Detection of structural motifs can reveal subtle evolutionary relations between proteins and provide insights into function for as yet uncharacterized biomolecules [1, 4]. In other cases, structural motifs common to two proteins are the result of convergent evolution that allows two polypeptide chains of quite different 3D folds to catalyze very similar, if not identical, chemical reactions [1]. Thus, identification of structural motifs represents a powerful tool for finding functional similarities between proteins, which may not be evident when evaluating solely primary, secondary, tertiary, or even quaternary structures.

Fig 1. Structural motif case studies. (**A**) Active sites of serine proteases can be made up of multiple polypeptide chains [2]. (**B**) Active site sidechains in aminopeptidases [3] coordinate two adjacent ions. (**C**) Zinc Finger DNA-binding domains [5] are stabilized by zinc ions (N.B.: Cys:F-212 was not used to define the search query.) (**D**) Position-specific exchanges (additional *label_comp_id*) can be used to identify enolase superfamily members accurately [6]. (**E**) RNA G-tetrads can be formed between one, two, or four nucleic acid strands [7]. *label_comp_id*, *auth_asym_id*, and *auth_seq_id* as residue identifiers. Rendering by Mol* [8].

Searching for structural motifs can be viewed as a computationally expensive task that requires exhaustive evaluation of many possible combinations in 3D. For a polypeptide chain of length n, there are $\binom{n}{k}$ combinations of motifs of size k. Conversely, deciding whether or not a particular structural motif is present in a given protein can be formulated as subgraph isomorphism problem, which is NP-complete [9]. It is also challenging to automatically classify structural motifs [1, 4, 10, 11], a task commonly realized by identifying overrepresented structural motifs in a protein family [10, 11]. An alternative approach involves manual or semi-automatic biocuration by subject-matter experts. For example, the Catalytic Site Atlas gathers such definitions of structural motifs [12].

Several structural motif search routines have been implemented over the past three decades. They commonly employ geometric hashing techniques or graph theoretical methods (reviews in [1, 4]). Geometric hashing strategies [10, 13–16] describe the relative arrangement between two (or three) residues in a rotation invariant fashion. Typically, geometric descriptors are used to create a reduced hash representation of complex 3D data (*e.g.*, distances between C_α atoms in motif residues). Such simplifications allow geometric hashing approaches to perform computationally expensive tasks only once during a preprocessing step and then reuses the results of this initial step to support rapid query responses [4]. This strategy requires additional storage but permits repeated reuse of the computed reduced low-dimensional representations.

An alternative approach is found in graph theoretic methods [17,18]. Proteins can be represented as graphs with residues being the vertices and edges capturing spatial proximities. A structural motif search then can be posed asking to whether or not a corresponding subgraph occurs in a graph defined by the whole protein. There are also combinatorial approaches such as Fit3D [19] which make exhaustive searching feasible by rejecting candidates early in the process. This step is crucial for an exhaustive search, as it would otherwise require evaluation of every structure file in the search space. Uniquely, the Suns [9] tool adapts strategies from web search engines and enables real-time discovery of similar motifs for protein design. However, its speed is achieved by reporting only modest numbers of results. This approach is inadequate for more general applications such as statistical analyses [20,21].

Given ∼10% year-on-year growth of 3D biostructure data in the PDB [22,23], practicable search implementations must scale linearly with the size of the search space [24]. For global structure comparison, we recently published an efficient approach using BioZernike moments [24]. Users can perform queries on >160,000 PDB structures and retrieve results instantaneously. There is currently an unmet need for a complementary strategy capable of near instantaneous searching for structural motifs across the entire PDB archive.

Herein, we present a real-time structural motif search algorithm that returns results within seconds using readily available computational resources. A novel indexing strategy, inspired by web search indices, describes the relative spatial arrangement of residue pairs ("words" in a text search context) composed of amino acids and/or nucleotides. The structural motif search problem can then be formulated as search for a collection of residue pairs in a set of protein structures ("documents" in a text search context). Search engines tackle similar problems by creating an inverted index [25] that keeps track of all documents in which a certain word occurs. For our motif search problem, such a word-level inverted index keeps track of the polymer sequence positions at which certain residue pairs occur. This strategy enables quick retrieval of residue pairs in a large data corpus similar to the index at the end of a printed volume. It can be used to support real-time structural motif searching across the entire PDB archive and other large 3D structural data sets. Our approach supports several advanced features, including analysis of motifs distributed across multiple polymer chains, support for nucleic acid motifs and post-translationally modified residues, and position-specific exchanges for analyzing evolutionary relationships. Dissimilarity among ensembles of structural motif hits is quantified using the root-mean-square deviation (R.M.S.D.) measure.

Results and discussion

Structural motif searching using inverted indexes. Structural motif searching using an inverted index approach involves creating a lookup table for all arrangements of residues present in the PDB (see Methods). This strategy recasts the search problem to one of loading all occurrences from the inverted index. To achieve this goal, a query motif (Fig 2A) is specified as a collection of residues. The query is fragmented into residue pairs and each pair is represented by a rotation-invariant descriptor (Fig 2B) that is extracted from the known coordinates given by the query. The descriptors are based on residue labels (*e.g.*, serine, aspartic acid, and histidine for the catalytic triad), the backbone distance d_b (C_α for amino acids), the sidechain distance d_s (C_β for amino acids), and the relative angle θ defined by the two vectors connecting backbone and sidechain of each residue (see Methods). Exact values are binned for distances (width=1 Å) and angles (20°). Similar occurrences of residue pairs are thereby sorted to the same bin.

Fig 2. Structural motif search workflow. (**A**) Fragmentation into residue pairs. (**B**) Computation of geometric descriptors. (**C**) Inverted index lookup. All similar occurrences are retrieved for each descriptor. (**D**) Checking for correspondence to ensure that candidate resembles query motif. (**E**) Structures not fulfilling requirements are ignored. Only relevant residues are loaded. (**F**) R.M.S.D. quantifies structural similarity.

The inverted index approach supports lookup of all occurrences of a certain residue pair descriptor sharing similar geometric properties (Fig 2C). The result is an exhaustive map of all PDB IDs wherein this residue pair occurs at least once. Individual occurrences are identified by an expression such as 1–87, wherein 1 corresponds to the first assembly generation operation and 87 refers to the residue with index 87 within a given PDB structure. The inverted index enumerates all residues within a structure, independent of polymer chain locations. Thus, our approach generates a word-level inverted index (*i.e.*, it reports both the PDB ID and the unique position of its occurrence). The binning strategy for distances (or angles) requires that the lookup surveils neighboring bins (Fig 2C), to ensure that occurrences close to bin boundaries are not lost (Fig 3). By default, a tolerance value of 1 is employed, resulting in lookup in three bins for each geometric descriptor. The inverted index contains information on pairs of residues; thus, partial results have to be combined to represent motifs with 3 or more residues (Fig 2D). Valid candidates are collections of residues that fulfill the requirements imposed by the structural motif query. 3D structures in the search space can be ignored if they do not contain all residue pairs present in the query motif. Moreover, all residue pairs of a valid candidate must occur in the correct spatial arrangement. This approach allows us to determine whether or not a certain PDB structure contains a query motif without actually loading any structural data.

For example, no compatible residue pairs occur between histidine and serine (descriptor HS-8-7-5) for PDB IDs 1aab and 1aac during Ser-Asp-His structural motif candidate assembly (Fig 2D). This finding permits rejection of these PDB IDs in all subsequent candidate assembly operations. In practice, qualifying valid candidates allows the structural motif search to be narrowed to hundreds or at most thousands of PDB structures. S1 Table shows how the serine protease query requires loading of only 4,577 PDB structures, ignoring ∼97% of the PDB archive because the inverted index reported no valid candidates for these structures.

Fig 3. Sensitivity of geometric descriptors. Ground truth for catalytic triad query was determined by an exhaustive search routine [26]. Most low R.M.S.D. hits are found by our approach (blue points). Biologically relevant hits exhibit geometric descriptors (area shaded in blue) similar to the query motif (black horizontal line).

Structure data is read for candidates that contain residue pairs similar to the query (Fig 2E). Coordinates are retrieved from a database that allows random access to individual residues of the archive. Each hit is then aligned to the query motif and its dissimilarity is scored by the R.M.S.D. (Fig 2F). The computed R.M.S.D. can be used for downstream operations such as filtering or sorting of hits. Finally, a result list is composed of all returned hits. See Materials and Methods for details.

Runtime analysis. Table 1 provides results of runtime analyses. Our approach provides results within 1 or 2 seconds and scales linearly with the number of structures in the search space (S1 Fig). The majority of the computation time is devoted to reading the inverted index and qualifying valid candidates. The remainder of the runtime is dedicated to retrieving coordinates and computing R.M.S.D. values for each putative hit.

Table 1. Runtime and sensitivity analyses.

Structural Motif Search	Hits	Time	First FN	FNR <1 Å
serine protease [2]	3,498	0.92 s	0.72 Å (1hcg)	1.86%
aminopeptidase [3]	350	0.46 s	N/A	0.00%
zinc coordination [5]	1,056	0.13 s	0.42 Å (2eou)	8.52%
enolase superfamily [6]	288	0.36 s	0.89 Å (3mkc)	2.33%
enolase superfamily (exchanges)	308	0.87 s	0.89 Å (3mkc)	5.62%
RNA G-tetrad [7]	84	1.10 s	1.53 Å (2rsk)	0.00%

Each reported runtime represents the average over 10 benchmark runs. Results are compared to an exhaustive search strategy [26] and the minimum R.M.S.D. of any false negative (FN) hit is reported. The false negative rate (FNR) below 1 Å (*i.e.*, a general cutoff below which we consider hits to be biologically meaningful) is given. S1 Table shows that false-negative rate can be reduced by a moderate runtime increase.

The interplay between structural motif definition and runtime is complex. Motif size and composition show no clear influence on the runtime. The same is true for the number of hits over which R.M.S.D. values are computed. Most significant in determining runtime is the structure of the corresponding bin of the inverted index (see below). Highly populated bins require more time for input/output and also require more logic operations to qualify valid candidates. In comparison to other methods, our approach is indisputably superior in computational efficiency. It does not resort to reporting only a small number of similar motifs [9], nor does it require that the search space be filtered in advance to eliminate redundancy. Searching the entire PDB is necessary to integrate the structural motif search routine with other RCSB PDB (rcsb.org) search capabilities, such as sequence similarity, text search, or shape search [22, 24, 27]. Moreover, our approach will not collapse under the weight of the relentlessly expanding PDB, because its complexity varies linearly with the number of structures in the archive (S1 Fig). Queries with position-specific exchanges (as shown for the enolase superfamily template) require many more read operations in the inverted index but exhibit only a slightly higher response time (S1 Table).

Case studies

We exemplify uses of our approach with five previously characterized structural motifs that are well represented in the PDB (Figure 1). These examples support diverse biological/biochemical functions and differ in size and complexity. The catalytic triad found in serine proteases [2] is a frequently showcased structural motif. Some occurrences are difficult to detect because the motif may be distributed among multiple polypeptide chains. The active site of leucine aminopeptidase [3] is a larger motif encompassing five residues, collectively responsible for coordinating two zinc ions. The His_2/Cys_2 Zinc Finger motif [5] provides an alternative example of a structural motif responsible for metal binding that stabilizes the structures of many DNA-binding domains found in many eukaryotic transcription factors. Complex evolutionary aspects can be represented by structural motifs with position-specific exchanges [19] as seen for the enolase superfamily [6]. G-tetrads are a prominent nucleic acid association motif [7].

Serine proteases: Detection of the catalytic triad Many hydrolases use a serine nucleophile during catalysis. Canonical serine protease catalytic triads are composed of His, Asp, and Ser residues (Figure 1A – PDB ID 4cha – His:B-57, Asp:B-102, Ser:C-195). They typically occur within two polypeptide chains, because

many proteases are initially made as zymogens that require activation by proteolytic processing [2] to prevent uncontrolled digestion of proteins within the cell. A His, Asp, and Ser catalytic triad query based on the configuration of these three residues in PDB ID 4cha returns 3,498 hits in ∼0.9 s. We assessed the false negative rate by comparison to an established, exhaustive search strategy represented by Fit3D [19]. Using an R.M.S.D. cutoff of 1 Å to distinguish positive from negative, our inverted index method gave a false negative rate of <2%. Our approach is not sensitive to the number of polymer chains over which the catalytic triad structural motif is distributed. We successfully detected both known examples involving three polypeptide chains, including trypsin (PDB ID 1ept – His:B-41, Asp:C-52, Ser:A-40 – R.M.S.D.=0.3 Å) and thrombin (PDB ID 2hnt – His:C-26, Asp:D-51, Ser:B-43 – R.M.S.D.=0.3 Å). Both proteins support proteolytic activity. In these rare cases, zymogen activation by peptide cleavage yielded the structural catalytic motif distributed over 3 chains instead of the canonical two chains. This example shows that our approach comes very close to replicating the results of an exhaustive search strategy in less than a second, even for a structural motif that is highly abundant in the PDB archive. For comparison the runtime of the exhaustive search [26] for the catalytic triad benchmark on server hardware required 131 s. We used the least restrictive BLASTe-80 target list (36,213 structures evaluated), Fit3D returned 538 matches below 1 Å R.M.S.D. Our approach returned 2,976 hits <1 Å after evaluating the entire PDB archive. It also successfully identifies inter-chain arrangements of residues matching the query, and the method is agnostic as to polymer type(s) (protein *versus* nucleic acid) and the presence of chemical modifications (*e.g.*, phosphorylation).

Figure 3 depicts how R.M.S.D. values relate to the significance of a given hit detected by our inverted index method. The density distribution shows two distinct sets of hits: One <1.0 Å R.M.S.D. and one >1.5 Å. Hits with high R.M.S.D. values likely encompass arrangements of histidine, aspartic acid, and serine that happen to be in proximity but do not function as serine proteases. The ensemble of hits with higher R.M.S.D. values also show high variance with respect to the discussed geometric descriptors. In contrast, hits with lower R.M.S.D. values all exhibit geometric properties that resemble the query motif (horizontal black line). A further speedup of our method can be achieved by identifying candidates that yield low R.M.S.D. values and ignoring the remainder. The tolerance parameter specifies how much deviation from the query motif is tolerated (area shaded in blue). We used R.M.S.D. <1 Å as a preliminary cutoff for the remaining cases discussed in this paper.

Aminopeptidase: Retrieval of all occurrences of a rigid motif

Aminopeptidases play important roles in protein degradation by removing residues from the N-terminus of a polypeptide chain [3]. Bovine leucine aminopeptidase (BLLAP) is a hexameric enzyme with 3_2 symmetry. The active site of BLLAP contains two adjacent zinc ions separated by ∼2.9 Å and coordinated by the sidechains of five conserved residues Lys, Asp, Asp, Asp, and Glu (Figure 1B – PDB ID 1lap – Lys:A-250, Asp:A-255, Asp:A-273, Asp:A-332, Glu:A-334). This five-residue query could be executed in <0.5 s, approximately half the time required for the catalytic triad search. Query motifs with more than three residues are simplified by a minimum spanning tree approach (see Methods), which removes cycles from the graph defined by the query. This results in $n-1$ constraints with n referring to the number of residues. In this case, all 9 BLLAP structures (UniProt ID P00727) in the PDB were detected with this structural motif search. The remaining 48 hits (R.M.S.D.<1 Å) represent leucine aminopeptidases from other organisms or leucine-aminopeptidase-like enzymes, such as the bottromycin maturation enzyme (PDB ID 5lhj, R.M.S.D.=0.4 Å, sequence identity to BLLAP=35.3%). One PDB structure similar in sequence to BLLAP was not

detected by our query search due to a deviating arrangement of motif residues (motif R.M.S.D.=2.2 Å). PDB ID 3pei a cytosolic aminopeptidase from *F. tularensis* (the causative agent of tularemia) is similar to BLLAP in both sequence (amino acid identity=38%) and structure (R.M.S.D.=1.0 Å for 291 alpha carbon atomic pairs). It also has 3_2 symmetry and a very similar quaternary arrangement to that of BLLAP. This particular structure was reported to be the five residues used for the motif search query present in its active, but in a more flexible arrangement. (Fit3D also failed to detect 3pei.) This bacterial aminopeptidase structure was determined in the absence of divalent metal ions, which may explain the apparent active site structural differences *versus* BLLAP. Alternatively, the prokaryotic enzyme may be dependent on an alternative divalent metal. The foregoing discussion serves to underscore the fact that motif definitions vary in their utility and it may be challenging to find optimal representations for more flexible motifs [28].

Zinc Fingers: Motif definition is key Eukaryotic transcription factors often contain His_2/Cys_2 Zinc Finger domains (Figure 1C – PDB ID 1g2f – Cys:F-207, Cys:F-212, His:F-225, His:F-229). These motifs are composed of two cysteine and two histidine residues which stabilize a small $\beta\beta\alpha$ domain structure enveloping a single zinc ion [5]. In the absence of the zinc ion, these domains do not adopt compact, folded structures and are incapable of binding DNA. Our experience with the His_2/Cys_2 motif reflects the importance of how the query is constructed. We found that the cysteine corresponding to F-212 in the query search occurs commonly in a distinct orientation *versus* the original motif definition. Consequently, the number of false negatives identified in this four-residue motif definition is higher than expected (∼67.0%). To overcome the fact that there is this subtle structural variation within the His_2/Cys_2 Zinc Finger family of DNA-binding domains, we employed a simplified query definition omitting position 212. S2 Fig shows the variability of the angle descriptor between residues 1 and 2 that gave rise to the unexpectedly large number of false negatives. The first false negative (PDB ID 2elv) is depicted in S3 Fig for illustrative purposes. The polypeptide chain backbone geometry in the vicinity of Cys:F-212 varies within this DNA-binding domain family.

A simplified 3-residue search query (PDB ID 1g2f – Cys:F-207, His:F-225, His:F-229) yielded more hits and compares more favorably to the results of an exhaustive search (false negative rate=8.5%). The runtime for the three-residue query increases only slightly *versus* the four-residue query, which can be attributed to more structures included in the R.M.S.D. value computation. These results show that too rigorous a query definition can lead to an unacceptable number of false negatives. In contrast to the histidine residues that are part of the α-helix, the cysteine residues corresponding to Cys:F-212 occur in loops [5], which may exhibit greater structural flexibility than residues in defined secondary structural elements. This problem could be mitigated by defining suitable curated queries for popular motifs within rcsb.org or provide this information as a dedicated public resource akin to the Catalytic Site Atlas [12]. The efficiency of our method allows us to recommend that users vary their motif definitions and seek consensus among multiple runs of similar but different queries.

Enolase superfamily: Efficient searching with position-specific exchanges The enolase superfamily is a group of proteins diverse in sequence, yet largely similar in 3D structure that all catalyze abstraction of a proton from a carboxylic acid [29]. The structural motif supporting this catalytic function [6] is represented in PDB ID 2mnr (Figure 1D – Lys/His:A-164, Asp:A-195, Glu:A-221, Glu/Asp/Asn:A-247, His/Lys:A-297). This particular case is one of the more challenging for motif searching in 3D. Isofunctional exchanges between histidine and lysine have been observed for the

first and last position. Similarly, the glutamic acid at A-247 can be substituted by aspartic acid or asparagine [6]. When no exchanges are considered, the simplest possible query returned 288 hits within <0.4 s. When position-specific exchanges are incorporated within the query, the number of hits with R.M.S.D. <1 Å increases to 308 (computation time <0.9 s). Additional hits with exchanges exhibit R.M.S.D. values ∼1 Å. As before, increased tolerance values result in retrieval of all hits (S1 Table). An increased tolerance value results in no false negatives below 1 Å for the query with position-specific exchanges. The enolase superfamily example demonstrates that our method can process even combinatorially complex queries very efficiently. The number of candidates to evaluate increases dramatically when increased tolerance values and position-specific exchanges are involved at only modest cost in terms of computational time (Table 1).

RNA G-tetrad: Nucleotide motifs in biological assemblies G-tetrads are a common nucleic acid association motif. They are composed of guanine and stabilized by Hoogsteen base pairings (Figure 1E). The four O6 oxygen atoms coordinate monovalent ions, such as K^+, and individual tetrads tend to be stacked one atop the other [7]. A query for the G-tetrad motif takes ∼1 s to complete and returns 84 hits. Interestingly, some G-tetrads arrangements were detected within larger assemblies. For example, PDB ID 3mij (A-4 and A-10) with a R.M.S.D of ∼0.7 Å was detected in a telomeric RNA G-quadraplex. These results document that our approach provides support for facile searching of structural motifs in nucleic acids, and also indexes occurrences in homo- and hetero-meric biological assemblies, including protein-nucleic acid complexes (data not shown).

Structure of the Inverted Index We constructed the inverted index for the PDB as of 2/17/20 encompassing 160,467 distinct structures. The size of the index of amino acid pairs is ∼55 GB, distributed among 239,034 bins (unique combinations of amino acids, distances, and angles). The index contains 6,814,159,549 residue pairs with distance d_b of up to 20 Å, with the largest 21,487 bins representing >50% of all occurrences. Each bin contains an average of 28,514 occurrences (positions referenced wherein this residue pair is observed) distributed over an average of 11,965 PDB structures. The ten largest bins (Table 2) contain combinations of alanine, glutamic acid, glycine, and leucine residues. The most frequent residue pairs tend to capture sequence neighbors (as indicated by alpha carbon separations of ∼4 Å). Other common residue pairs are those with distances d_b near the cutoff value of 20 Å. In general, bins are of particular interest when they contain an elevated number of entries reflecting function (*e.g.*, catalytic triads in serine proteases [2]) or non-covalent interactions responsible for protein structure stabilization.

As described above, our inverted index approach drastically reduces the search space even for individual residue pairs. Even the largest bin AL-4-5-4 (Table 2) eliminates approximately one third of the PDB archive from consideration during a search involving that particular combination of residues. Inverted index-based queries for functional motifs usually consist of multiple residue pairs. cases, the inverted indexing strategy pinpoints a few hundred to thousands of PDB structures that contain the query (S1 Table). The computational frugality of this strategy underpins the speed of our approach, which is particularly relevant given the increasing complexity of more recent PDB deposited depositions [22]. We assessed the impact of the increasing size of the PDB by comparing the current PDB (160,000 structures) to the holdings of the archive at the end of 2012 (roughly half the size, 78,237 structures). The 2012 inverted index contains 2,483,893,161 residue pairs (∼37% of the 6,814,159,549 residues pairs for the current PDB). Between 2012 and 2020, the number of PDB structures increased by

Table 2. Ten most abundant residue pairs in the inverted index

Residue Pair Descriptor	# Structures	# Occurrences
AL-4-5-4	110,080	903,128
GL-19-19-2	103,954	555,431
AL-10-11-5	103,570	643,736
EL-4-5-4	103,470	623,808
EL-5-7-6	103,351	582,756
AE-4-5-4	103,191	639,248
EL-10-11-5	102,304	544,341
GL-18-18-2	101,624	510,194
GL-19-19-3	101,494	520,548
AL-19-19-2	101,308	612,042

Sorted by number of structures containing this residue pairs. Only amino acid pairs were considered.

∼105%, while the size of the inverted index increased by ∼174%. The fact that the inverted index did not grow in strict 1:1 proportion with the growth in the number of new structures reflects the fact that the average size of a structure deposited to the PDB has been increasing year-on-year [30]. The average PDB structure size (number of residues) rose from 620 to 885, comparing the 2012 archive snapshot to the entire contents of the current PDB.

Conclusion

The search for structural motifs is a computationally challenging task because it involves evaluating a large number of possible combinations. We have developed a robust, efficient method that utilizes a reduced representation of pairs of residues with two distance descriptors and one angle descriptor. Our approach enables composition of an inverted index that groups similar occurrences together. Our database approach for coordinate retrieval is entirely independent of the number of residues comprising each structure, which is essential with increasing complexity of deposited structures [22, 27]. The inverted index is used to retrieve the position of all occurrences similar to a query within seconds. R.M.S.D. values are computed for all occurrences. Our implementation requires loading of large amounts of structure data and is realized by a custom coordinate database which minimizes I/O operations. We provide a feature-rich implementation (supporting multiple chains, bioassemblies, and position-specific exchanges), which is available as an open-source project (github.com/rcsb/strucmotif-search) and will be used to augment the search capabilities provided by RCSB PDB [22, 27] on rcsb.com.

We set out to augment the RCSB PDB search capabilities by developing a scalable structural motif search strategy that goes well beyond existing technologies. Our implementation yielded an efficient and accurate processes that supports even quite sophisticated structural motif searches across the >160,000 structures represented in the PDB archive. Run times will scale linearly with the growth of the archive. The new tool will enable rapid testing of hypotheses (*e.g.*, by defining varying motif definitions [28]) for the millions of users who frequent the RCSB PDB website (rcsb.com) on an annual basis. This work builds upon previous experience and ongoing developments at RCSB PDB, all of which all aim at ensuring the efficient management of the ongoing 3D data deluge [8, 24, 31, 32].

Materials and methods

Amino acid and nucleotide representation During a structural motif query, the labels of the residues are known. In case of the catalytic triad, the query encompasses one histidine, one aspartic acid, and one serine. Thus, the problem can be simplified by ignoring 17 of 20 amino acids. Furthermore, residues occur in a specific distance from one another and their arrangement is constrained by the requirement to adapt a functional conformation. We determine three additional properties of each residue pair in the query. We propose to represent amino acids and nucleotides in a generic manner (Fig 4) whereby all amino acids are typified in a comparable way. We assume that the most valuable information is provided by the position of functional groups in the side chain. Therefore, we try to balance the influence of backbone and side chain atoms to accommodate cases with position-specific exchanges. For example, it is difficult to represent the side chains of a tryptophan and an alanine in a comparable way. In consequence, we choose to represent residues by backbone and sidechain coordinates. For amino acids, we consider C_α as backbone and C_β as sidechain representative, respectively. In case of glycine, the C_β atom is approximated by superimposing coordinates of a prototypic alanine. For nucleotides, we utilize $C4'$ as backbone and $C1'$ as sidechain representative. We represent query motifs and database structures by forming all residue pairs that are less than 20 Å apart with respect to their backbone distance d_b. Motifs can have an arbitrarily large extent as long as they are composed of residue pairs with less than 20 Å distance. We chose a threshold of 20 Å as this can comfortably represent all common motifs.

Fig 4. Representation of residue pairs. Residue pairs are represented by 3 descriptors that are transformation invariant: backbone distance d_b, sidechain distance d_s, and angle θ. This constitutes a compact representation of residue pairs and enables quick retrieval of similar pairs.

Geometric description Backbone distance d_b, sidechain distance d_s, and the angle θ can be used to describe the geometric arrangement of residue pairs in the search space (Fig 4). Target structures contain a hit when they contain residue pairs with the same properties as the query motif. We employ a binning approach to assess similarity at residue pair level. Distances are binned into 1 Å groups and angles are grouped by 20° intervals. This binning approach allows to represent the characteristics of each reside pair by a single integer value that captures both residue types, the d_p bin, the d_s bin, as well as the θ bin. Similar residue pairs are sorted into the same bin during index creation. Searching requires reading all occurrences registered in a bin to lookup all residue pairs that are relevant for a query. Every lookup operation should encompass reading of neighboring bins, otherwise highly similar residue pairs may be result in false negatives because they are separated by a bin border. We chose bin sizes to capture the majority of an exhaustive search strategy as represented by Fit3D [26]. We provide the option to increase the tolerance parameter which will report more hits but also is computationally more expensive.

Storage of pair occurrences The inverted index is implemented by a file system-based approach. Each bin is represented by a dedicated file which maps between PDB identifiers as keys and an array of residue pair positions as values. Occurrences are identified by the assembly generation operator and the index of the residue in the protein of origin. In other terms: This constitutes a word-level inverted index whereby words are pairs of residues in a certain arrangement and documents are PDB entries.

The inverted index readily provides the sequence position of all residues which makes the retrieval of individual residues convenient and fast. All information is stored in a custom binary data format using the MessagePack codec.

Assembly of residue pairs into candidates Speed and sensitivity are prime reasons to minimize the number of operations on the inverted index. Every lookup operation requires I/O time. Also, more constraints are introduced when all residue pairs of the query are required to be present in a protein structure. This may lead to an elevated number of false negatives. The constraints implied by the residue pairs of the query are transitive (given a constraint between residues A and B as well as B and C, then there is also some constraint on A and C). Therefore, it is advantageous to compose a list of query residue pairs that is as sparse as possible while still avoiding specious structures. Motifs with 4 or more residues are pruned by determining the minimum spanning tree of residue pairs using Kruskal's algorithm. Only the selected residue pairs are used to lookup occurrences and perform the search. Candidates that fulfill the query are determined by reading all occurrences for each residue pairs. Within the given tolerance range, all residue pair occurrences are pooled together into a map (PDB identifiers as key, collections of occurrences in each structure as value). Valid candidates are determined by enforcing that a certain PDB entry contains all residue pairs specified by the query. Furthermore, the individual residue pairs have to be connected, meaning that they are located close together and not scattered throughout the structure. This is done by testing whether the graph defined by the query motif is isomorphic to any graph present in the PDB entry of interest.

Management of structure data The inverted indexing strategy and subsequent screening for candidates reduces the number of coordinate files to assess to the order of hundreds to thousands. During development, we found that reading structure data even on this scale accounts for the majority of computation time of each run. We omitted hydrogen atoms and models with numbers unequal to 1, if present. The first location was kept when alternate positions existed for an atom. Furthermore, non-polymer groups such as water, ions, and ligands were removed. The precision of coordinates was decreased to 1 decimal place (in contrast to the normal 3) which has a minimal effect on the computed R.M.S.D. values but further eases storage requirements (and therefore time spent on I/O operations). We found that even efficient serialization strategy for macromolecular data such as MMTF [31, 32] constitute a performance bottleneck. We this issue by a database that allows the direct retrieval of individual residues based on their assembly operator and their sequence index (without reading information on other residues). This is especially beneficial for large ribosome structures or viral complexes. The corresponding database has a size of 39.4 GB and allows optimal throughput with respect to the number of candidates evaluated per second.

Scoring and significance of hits The inverted index simplifies R.M.S.D. calculation by providing the correspondence between residues of query and hit. Otherwise, all combinations of ambiguous labels would need to be evaluated (as is the case for the three aspartic acids in the aminopeptidase example). When position-specific exchanges occur at a position, we form the intersection of atom names and compute the R.M.S.D. for that subset of compatible atoms. No efforts are made to find correspondence between ambiguously labelled atoms [33] (*e.g.*, C_δ and C_ϵ atoms of tyrosine). The computational load to compute the R.M.S.D. is negligible when a quaternion-based solution [34] is applied. The R.M.S.D. values enable sorting of hits by dissimilarity.

It would be desirable to quantify the significance of hits: *i.e.*, assess the probability that the corresponding hit is the result of a random arrangement of amino acids and not biologically meaningful. Statistical models try to address this problem [20, 21], however no widely accepted solution exists and, maybe, it is even impossible to compute the significance of structural motifs in a truly objective way [1]. We assume that the best way to assess significance is still expert knowledge. The newly added visualization capabilities of Mol* [8] and the cross-references provided by the RCSB PDB web page are an excellent starting point to put a notable hit into further context.

Visualization of hits The challenges in significance assessment emphasize the importance of visualization. Expert knowledge is the most reliable way of identifying meaningful hits. This allows to visually inspect whether the fold of a hit corresponds to that of the query protein. Other characteristics to investigate are the solvent exposure of a hit that may be required for active sites as well as the presence of ligands or ions. The prototype implementation (motif.rcsb.org) is based on the NGL viewer [35]. The prototype allows the definition and submission of custom queries. A list of all results is presented, and the user can align individual motifs to the query as well as align the complete structure that contains query or motif.

Benchmarking setup Runtime measurements were performed on a 3.2 GHz Intel Core i7 CPU with 12 cores, 16 GB memory, and macOS. The inverted index and the coordinate database are stored on an SSD. Benchmarks of Java implementations were executed using JMH Java Benchmark Harness using Oracle JDK (HotSpot) 11.0.4. There are some difficulties pertaining benchmarking the search routine. Some time is required to warm-up and optimize the code just-in-time. Therefore, 5 warm-up and 10 measurement iterations performed.

Exhaustive search for structural motifs We assessed whether our approach finds the same set of relevant hits (*i.e.*, with a R.M.S.D. value below 1 Å) as an established, exhaustive search strategy. The Fit3D web server [26] was used to retrieve a result list for each query motif (Figure 1) using the *BLASTe-80* target list. The R.M.S.D. of the first false negative hit was determined as well as the overall false negative rate. Fit3D allows hits in regions composed solely of alpha carbon atoms. Such hits are likely false positives and cannot be found with our approach because no beta carbons exist to compute descriptors. Therefore, we removed them from the exhaustive list. This methodology was also used to prepare Figure 3. For the G-tetrad motif, we submitted a custom target list of all RNA sequences in the PDB archive.

Acknowledgments

We gratefully acknowledge discussions and feedback by the RCSB team and Chris Randle specifically for setting up the service at motif.rcsb.org. We are grateful to Florian Kaiser for sharing his expertise on structural motif searching and for providing the abstract depiction of amino acids and nucleotides. We also thank Michael Schroeder for noting the connection between motif definition and motif search. RCSB PDB is funded by grants to SKB from the National Science Foundation (DBI-1832184), the US Department of Energy (DE-SC0019749), and the National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences of the National Institutes of Health under grant R01GM133198.

References

1. Via A, Tramontano A. Protein Structural Motifs: Identification, Annotation and Use in Function Prediction. In: Sequence and Genome Analysis II – Methods and Applications; 2011. p. 1–21.

2. Hedstrom L. Serine protease mechanism and specificity. Chemical reviews. 2002;102(12):4501–4524.

3. Burley SK, David PR, Taylor A, Lipscomb WN. Molecular structure of leucine aminopeptidase at 2.7-A resolution. Proceedings of the National Academy of Sciences. 1990;87(17):6878–6882.

4. Nilmeier JP, Meng EC, Polacco BJ, Babbitt PC. 3D Motifs. In: From Protein Structure to Function with Bioinformatics. Springer; 2017. p. 361–392.

5. Pabo CO, Peisach E, Grant RA. Design and selection of novel Cys2His2 zinc finger proteins. Annual review of biochemistry. 2001;70(1):313–340.

6. Meng EC, Polacco BJ, Babbitt PC. Superfamily active site templates. PROTEINS: Structure, Function, and Bioinformatics. 2004;55(4):962–976.

7. Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S. Quadruplex DNA: sequence, topology and structure. Nucleic acids research. 2006;34(19):5402–5415.

8. Sehnal D, Rose A, Koča J, Burley S, Velankar S. Mol*: towards a common library and tools for web molecular graphics. In: Proceedings of the Workshop on Molecular Graphics and Visual Analysis of Molecular Data. Eurographics Association; 2018. p. 29–33.

9. Gonzalez G, Hannigan B, DeGrado WF. A real-time all-atom structural search engine for proteins. PLoS computational biology. 2014;10(7):e1003750.

10. Nussinov R, Wolfson HJ. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. Proceedings of the National Academy of Sciences. 1991;88(23):10495–10499.

11. Kaiser F, Labudde D. Unsupervised Discovery of Geometrically Common Structural Motifs and Long-Range Contacts in Protein 3D Structures. IEEE/ACM transactions on computational biology and bioinformatics. 2017;16(2):671–680.

12. Ribeiro AJM, Holliday GL, Furnham N, Tyzack JD, Ferris K, Thornton JM. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucleic acids research. 2017;46(D1):D618–D623.

13. Pennec X, Ayache N. A geometric algorithm to find small but highly similar 3D substructures in proteins. Bioinformatics (Oxford, England). 1998;14(6):516–522.

14. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein science. 1997;6(11):2308–2323.

15. Moll M, Bryant DH, Kavraki LE. The LabelHash algorithm for substructure matching. BMC bioinformatics. 2010;11(1):555.

16. Wolfson HJ, Rigoutsos I. Geometric hashing: An overview. IEEE computational science and engineering. 1997;4(4):10–21.

17. Konc J, Janežič D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. Bioinformatics. 2010;26(9):1160–1168.

18. Nadzirin N, Gardiner EJ, Willett P, Artymiuk PJ, Firdaus-Raih M. SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures. Nucleic acids research. 2012;40(W1):W380–W386.

19. Kaiser F, Eisold A, Labudde D. A novel algorithm for enhanced structural motif matching in proteins. Journal of Computational Biology. 2015;22(7):698–713.

20. Stark A, Sunyaev S, Russell RB. A model for statistical significance of local similarities in structure. Journal of molecular biology. 2003;326(5):1307–1316.

21. Fofanov VY, Chen BY, Bryant DH, Moll M, Lichtarge O, Kavraki L, et al. A statistical model to correct systematic bias introduced by algorithmic thresholds in protein structural comparison algorithms. In: 2008 IEEE International Conference on Bioinformatics and Biomedicine Workshops. IEEE; 2008. p. 1–8.

22. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic acids research. 2019;47(D1):D464–D474.

23. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic acids research. 2019;47(D1):D520–D528.

24. Guzenko D, Burley SK, Duarte JM. Preparing for the 3D data deluge: real time structural search at the scale of the PDB and beyond. bioRxiv. 2019;doi:10.1101/845123.

25. Knuth DE. The art of computer programming. vol. 3. Pearson Education; 1997.

26. Kaiser F, Eisold A, Bittrich S, Labudde D. Fit3D: a web application for highly accurate screening of spatial residue patterns in protein structure data. Bioinformatics. 2016;32(5):792–794.

27. Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic acids research. 2016; p. gkw1000.

28. Chen BY, Fofanov VY, Bryant DH, Dodson BD, Kristensen DM, Lisewski AM, et al. The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs. Journal of Computational Biology. 2007;14(6):791–816.

29. Babbitt PC, Hasson MS, Wedekind JE, Palmer DR, Barrett WC, Reed GH, et al. The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the α-protons of carboxylic acids. Biochemistry. 1996;35(51):16489–16501.

30. Young JY, Westbrook JD, Feng Z, Sala R, Peisach E, Oldfield TJ, et al. OneDep: unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. Structure. 2017;25(3):536–545.

31. Bradley AR, Rose AS, Pavelka A, Valasatava Y, Duarte JM, Prlić A, et al. MMTF—An efficient file format for the transmission, visualization, and analysis of macromolecular structures. PLoS computational biology. 2017;13(6):e1005575.

32. Valasatava Y, Bradley AR, Rose AS, Duarte JM, Prlić A, Rose PW. Towards an efficient compression of 3D coordinates of macromolecular structures. PloS one. 2017;12(3):e0174846.

33. Coutsias EA, Wester MJ. RMSD and Symmetry. Journal of computational chemistry. 2019;40(15):1496–1508.

34. Liu P, Agrafiotis DK, Theobald DL. Fast determination of the optimal rotational matrix for macromolecular superpositions. Journal of computational chemistry. 2010;31(7):1561–1563.

35. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW. NGL viewer: web-based molecular graphics for large complexes. Bioinformatics. 2018;34(21):3755–3758.

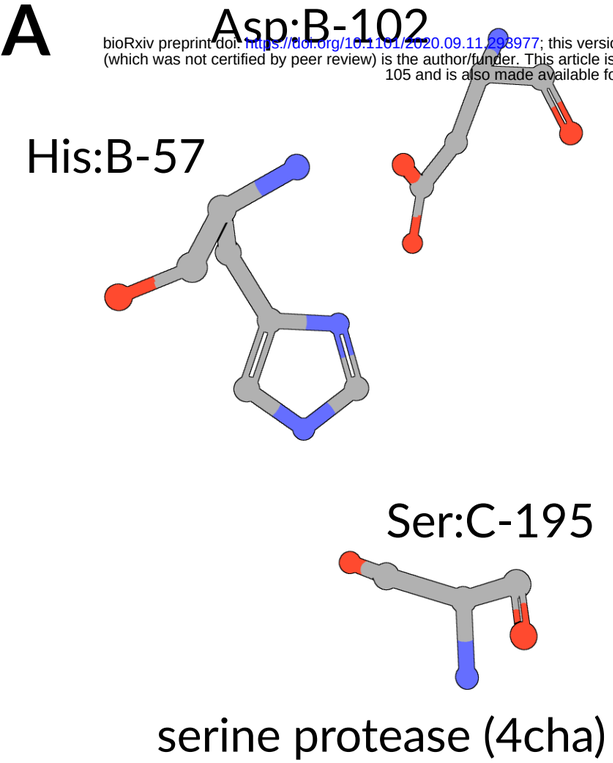
Supporting information

S1 Fig. Impact of archive size on runtime. Archive was sampled. Runtime for the catalytic triad query increases linearly with archive size.

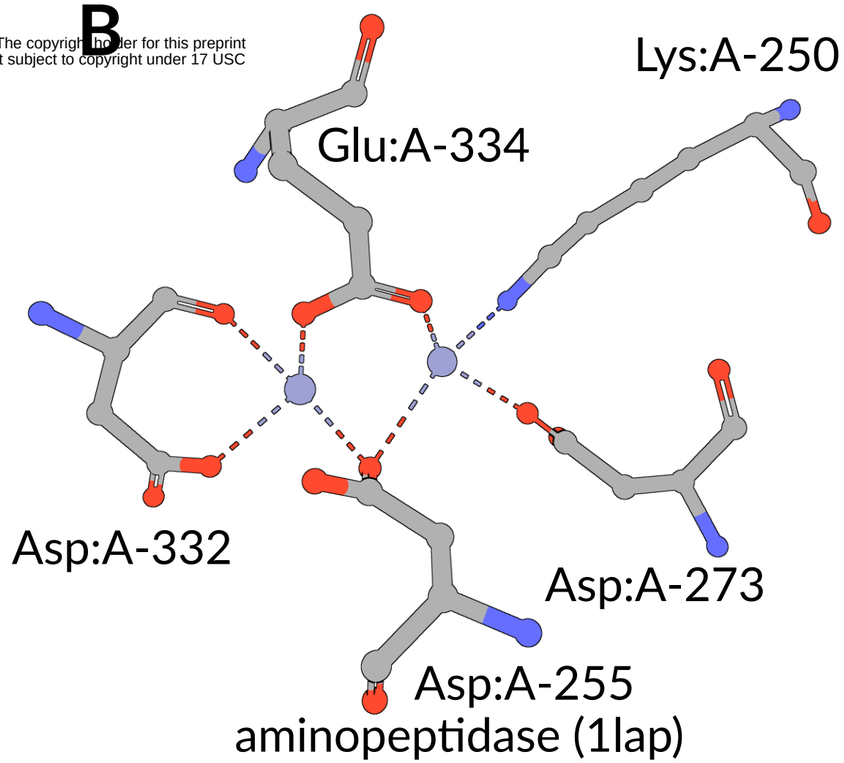
S2 Fig. Sensitivity of geometric descriptors for Zinc Finger motif. A common motif definition, that encompasses four residues, led to an unacceptable number of false negatives. Cysteine F-212 is structurally variable and most hits were not detected when the angle θ between residues 1 and 2 was below 70°.

S3 Fig. Rendering of the first false negative for Zinc Finger motif. The structurally flexible cysteine at A-15 in PDB ID 2elv (colored in green, motif in light grey) causes this hit to be a false negative. The vector between alpha and beta carbon is orthogonal to that of the query motif.

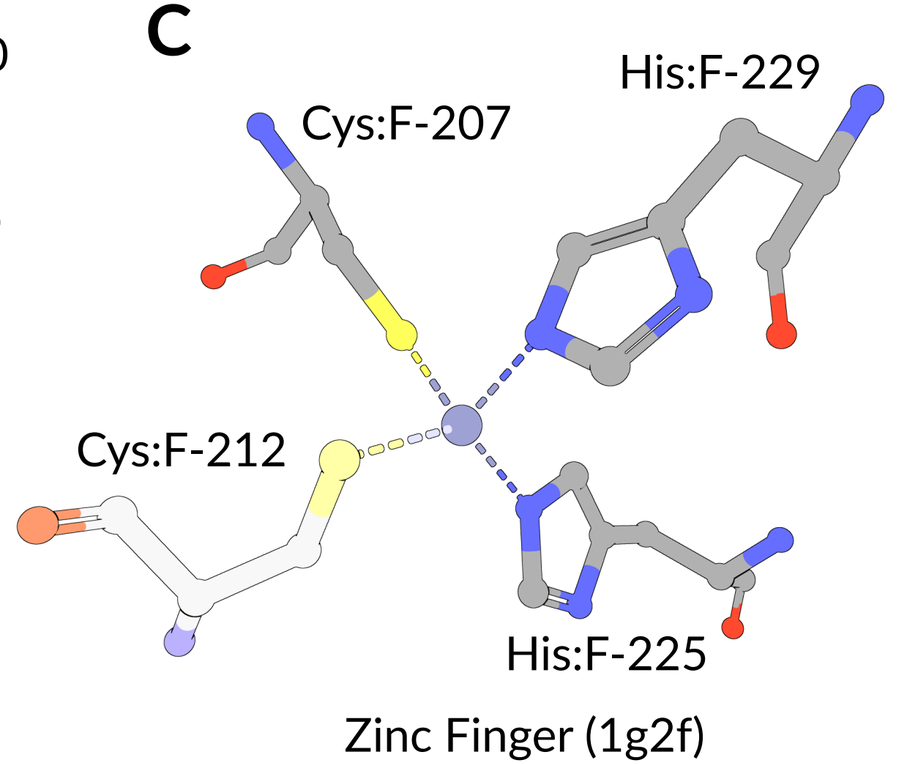
S1 Table. Tolerance analysis. Higher tolerance values lead to higher runtimes. Furthermore, increased tolerance values reduce false negative rate (in comparison to an exhaustive search strategy [19]). We consider hits below 1 Å as biologically meaningful. A tolerance value of 3 does not miss any hits below the threshold.



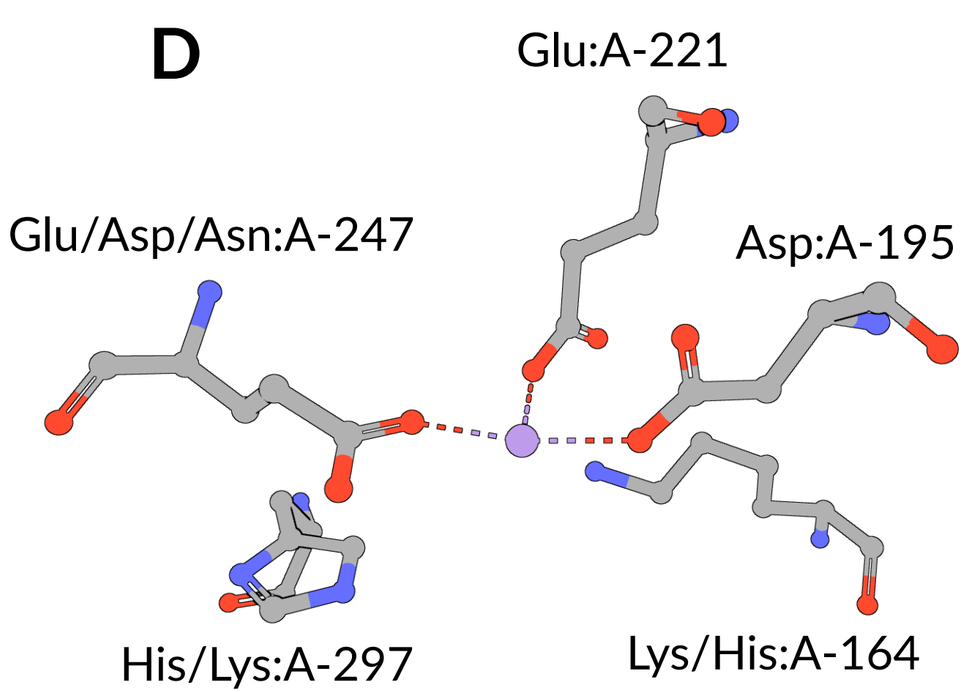
serine protease (4cha)



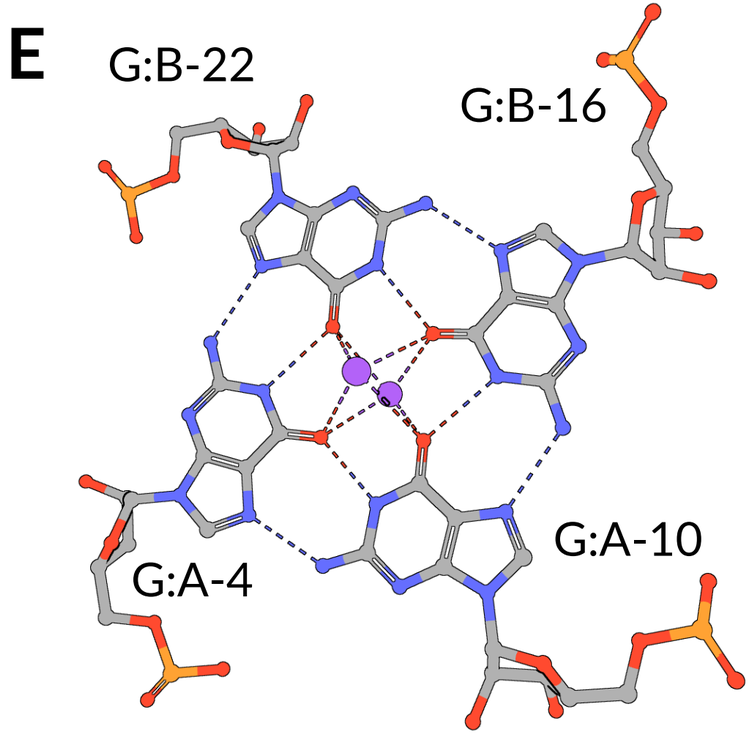
aminopeptidase (1lap)



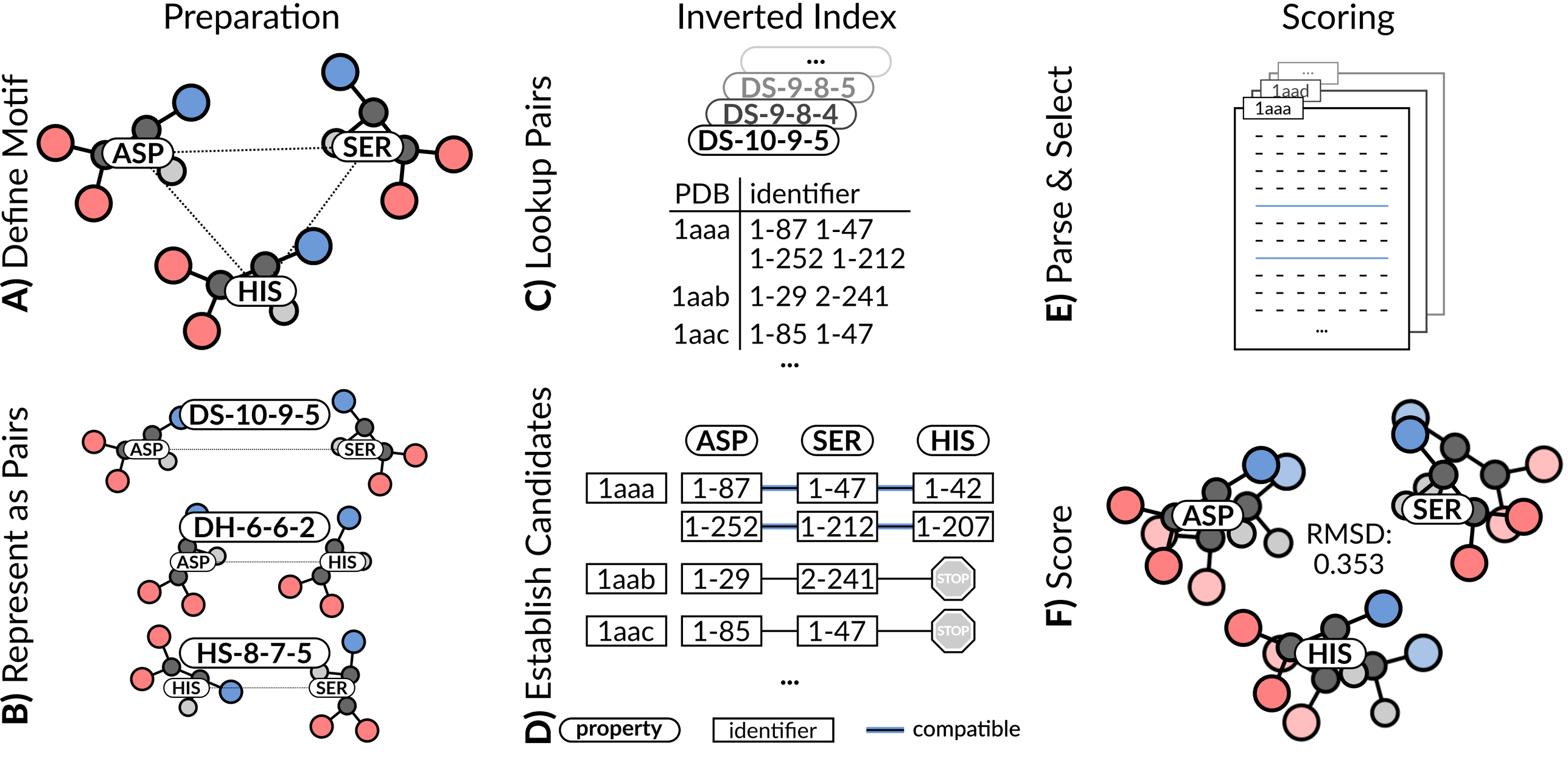
Zinc Finger (1g2f)

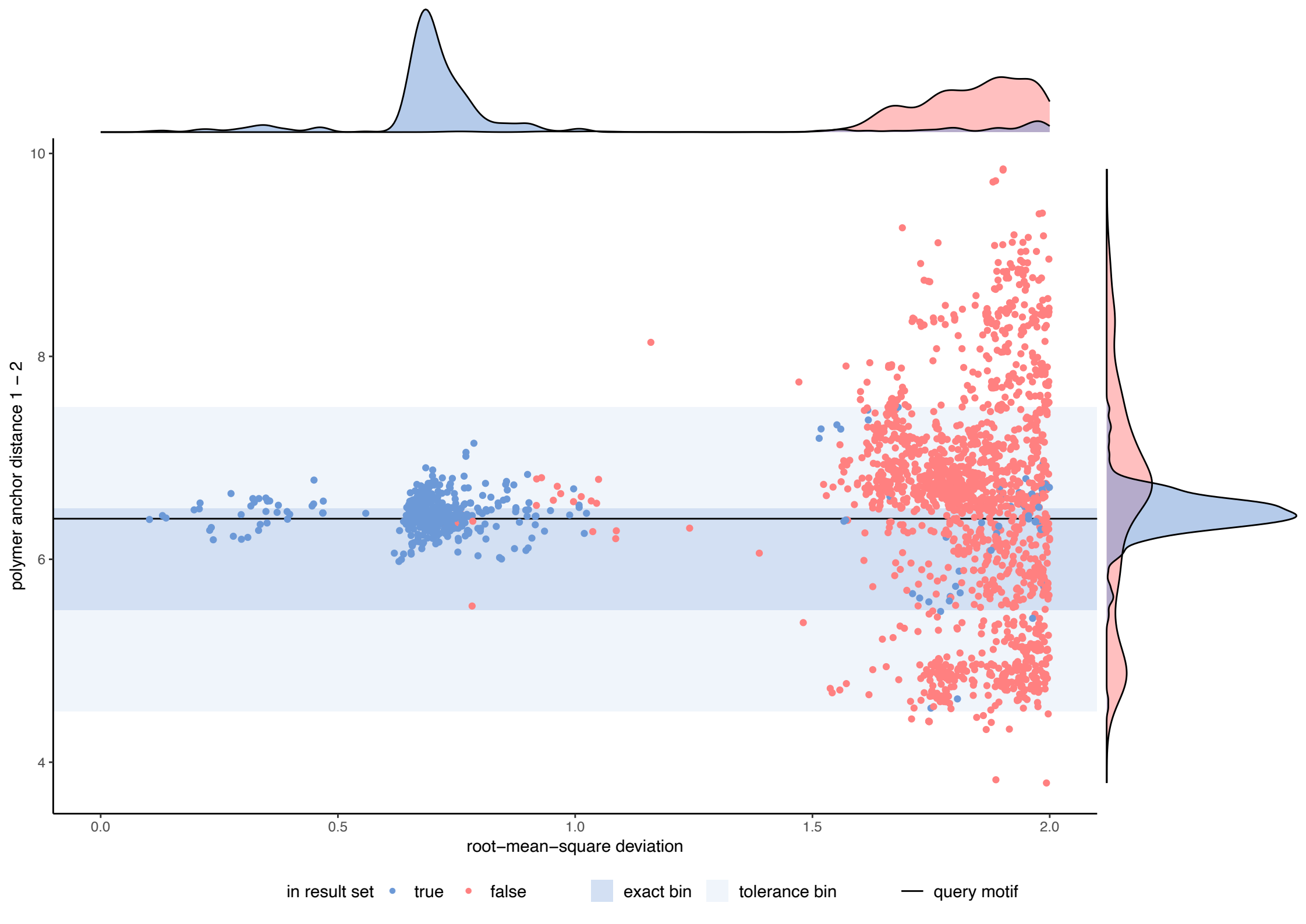


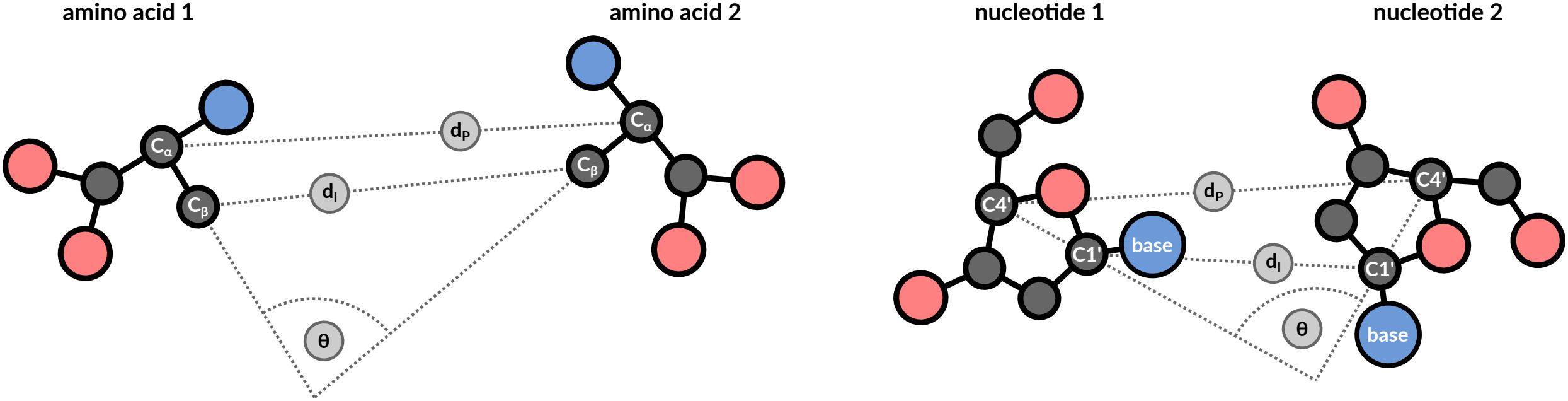
enolase superfamily (2mnr)



G-tetrad (3ibk)


Preparation
Inverted Index
Scoring
A) Define Motif
ASP
SER
HIS
B) Represent as Pairs
DS-10-9-5
ASP
SER
DH-6-6-2
ASP
HIS
HS-8-7-5
HIS
SER
C) Lookup Pairs
...
DS-9-8-5
DS-9-8-4
DS-10-9-5
PDB
identifier
1aaa
1-87 1-47
1-252 1-212
1aab
1-29 2-241
1aac
1-85 1-47
...
D) Establish Candidates
ASP
SER
HIS
1aaa
1-87
1-47
1-42
1-252
1-212
1-207
1aab
1-29
2-241
STOP
1aac
1-85
1-47
STOP
...
property
identifier
compatible
E) Parse & Select
...
1aad
1aaa
...
F) Score
ASP
SER
HIS
RMSD:
0.353


10
8
6
4
polymer anchor distance 1 – 2
0.0
0.5
1.0
1.5
2.0
root–mean–square deviation
in result set
true
false
exact bin
tolerance bin
query motif


amino acid 1
amino acid 2
Cα
Cβ
dP
dl
θ
nucleotide 1
nucleotide 2
C4'
C1'
base
dP
dl
θ