Published in final edited form as:

J Struct Funct Genomics. 2015 December ; 16(0): 101–111. doi:10.1007/s10969-015-9199-0.

Automated Protein Motif Generation in the Structure-Based Protein Function Prediction Tool ProMOL

Mikhail Osipovitch¹, Mitchell Lambrecht², Cameron Baker¹, Shariq Madha¹, Jeffrey L. Mills², Paul A. Craig², and Herbert J. Bernstein²

¹Gosnell School of Life Sciences, Rochester Institute of Technology

²School of Chemistry and Materials Science, Rochester Institute of Technology

Abstract

ProMOL, a plugin for the PyMOL molecular graphics system, is a structure-based protein function prediction tool. ProMOL includes a set of routines for building motif templates that are used for screening query structures for enzyme active sites. Previously, each motif template was generated manually and required supervision in the optimization of parameters for sensitivity and selectivity. We developed an algorithm and workflow for the automation of motif building and testing routines in ProMOL. The algorithm uses a set of empirically derived parameters for optimization and requires little user intervention. The automated motif generation algorithm was first tested in a performance comparison with a set of manually generated motifs based on identical active sites from the same 112 PDB entries. The two sets of motifs were equally effective in identifying alignments with homologs and in rejecting alignments with unrelated structures. A second set of 296 active site motifs were generated automatically, based on Catalytic Site Atlas entries with literature citations, as an expansion of the library of existing manually generated motif templates. The new motif templates exhibited comparable performance to the existing ones in terms of hit rates against native structures, homologs with the same EC and Pfam designations, and randomly selected unrelated structures with a different EC designation at the first EC digit, as well as in terms of RMSD values obtained from local structural alignments of motifs and query structures. This research is supported by NIH grant GM078077.

Keywords

template-based alignment; ProMOL; PyMOL; molecular visualization; structural bioinformatics; enzyme; catalytic site motif

Introduction

As part of the Structural Biology Extensible Visualization Scripting Language (SBEVSL) initiative, ProMOL [1] is a plugin for the molecular graphics system PyMOL [2]. ProMOL is written in Python 2.7 and utilizes a template-based function prediction method in an attempt to classify proteins of known structure but unknown function [1]. The method involves compilation of an extensive library of structural motif templates based on well

characterized enzyme structures with active sites that are defined in the Catalytic Site Atlas (CSA) [3]. Initial screening of approximately 3,500 PDB structures of unknown function with ProMOL led to tentative function assignments for 65 of these structures [4]. In ProMOL, the motifs are generated with the Motif Maker interface and are defined by the PDB ID's of the native structures, EC designations, types, numbers, functional components, and chain affiliations of catalytic site residues (Fig. 1). The motifs are stored as Python script files and include all functional atoms in the side chains of the amino acids that comprise the active site, stringency parameters such as Tolerance Factor (2.00 Å by default) and optional inclusion of backbone atoms. The motif templates in ProMOL are defined by the distance between the atoms of the amino acids involved. The Tolerance Factor represents the acceptable variation in distance between atoms that ProMOL uses when deciding if a query structure is a match for the motif template. Presently, the ProMOL library contains 181 motifs that were generated and tested manually. The Motif Finder interface is used to screen the motifs against query structures. The method of motif search is based upon relative distances between catalytic site residues (Fig. 2).

As of mid-2015, the Protein Data Bank (PDB [5]) contains over 4,000 structures without an assigned function. Expanding the motif library will increase the predictive capacity of ProMOL. The larger the number of motifs in the library, the greater the chance of identifying enzymatic functions in these proteins. The automated generation algorithm allows users to create motifs at a higher rate and in a more systematic fashion.

Materials and Methods

Automated Motif Generation Algorithm

ProMOL contains a motif library of enzyme active sites that is divided into “sets”. The previous release of ProMOL included a library of 181 motifs that were manually generated using the Motif Maker in ProMOL; these motifs are designated the “P set”. In this manuscript, we describe the creation of two additional sets: the “AP set” and the “A set”. Motif templates in the AP set are generated automatically based on the exact structures and residues that were used to manually create the P set. Motif templates in the A set were subsequently generated automatically based on literature-based entries in the Catalytic Site Atlas.

To test the algorithm, the AP set of 112 motifs was generated automatically based on the original P set of 181 motif templates. Here we explain why the number 181 was pruned first to 157, then to 116, and finally to 112. Motif definitions for the new AP set were created directly from P set motifs. Residue names, loci, chain, and EC code were obtained from Python files for P set motifs. Identification of functional components of residues as S (sidechain), as N (backbone amide nitrogen) or as O (backbone carbonyl oxygen) was obtained from the CSA. The only difference between P and AP set motifs is the inclusion of functional component information that is required in the auto generation algorithm that may include backbone atoms for selected residues, when those backbone atoms contribute to catalysis. When the auto generation algorithm was developed, we included only active site entries found in the CSA (1) that contained between 2 and 10 residues and (2) that had the active site located on chain A. If an enzyme had active sites on multiple chains, a motif was

automatically generated only in cases where the active site on chain A was identical to active sites on all other chains, i.e., no bifunctional enzymes. Given those rules, motif definitions were not created for 24 PDB entries found in the P set, leaving 157 candidates for automated motif definition. After the auto generation algorithm was run for the remaining 157 definitions, 116 motifs were saved. In the auto generation algorithm a motif is saved only if the matching subset is identical to the motif definition with tolerance factor

1.00. Of the 157 definitions, 41 motifs were not created because they did not select the matching set of residues in the proper range for tolerance factors. Our concern is that tolerance factors below 1.00 would allow for very little variation in position for active site residues, potentially reducing the number of true positives dramatically. Four additional motif definitions were removed from the test pool due to ambiguities or disagreements on EC class assignments between CSA and the PDB, leaving a total of 112 motif definitions available for comparisons between the P set and the AP set.

Testing against “itself”

To initiate the automated motif generation algorithm, a user must enter a comma separated list of PDB entries in the file `pdb_list.txt` which is located in the `/modules/pmg_tk/startup/ProMol` folder (location depends on the operating system). Once the user clicks on the Auto button (Figure 1A), the algorithm builds the motif and begins by screening the motif against its native structure. When the screening is completed, ProMOL returns a collection of residues that satisfy the distance constraints. The matching subset may include residues that satisfy the distance constraints but that are not found in the motif definition, because of inclusion of atoms from other residues of the same name at similar distances and inclusion of atoms from residues in multiple chains. These anomalies are highly unlikely for motifs of four or more residues. The algorithm stores the motif only if the matching subset returns the identical collection of residues as are found in the motif definition. If the initial matching subset contains extraneous residues, the algorithm first increases the motif stringency by including the backbone atoms of all residues in the definition and rescreens the motif against the native structure. If the extraneous residues are still present in the matching subset after the inclusion of the backbone atoms, the algorithm proceeds to decrease the Tolerance Factor (TF, 2.00 Å by default). At each iteration, the algorithm decreases the TF by 0.10 Å and rescreens the motif against its native structure. If the extraneous residues are not eliminated with the TF decreased to 1.00 Å, the algorithm abandons the current motif definition. This process is summarized in Figure 3. Motifs that are successfully created by the automated process are then saved in the ProMOL motifs folder. The motifs folder is a subfolder to the ProMOL folder that is created at the same level as the `promol.py` plugin code as part of the ProMOL installation process.

Testing against homologs

These motifs were then tested against homologs. Initial testing of the automatically generated motifs was performed against homologs that were defined simply as PDB entries that shared the same EC class. For subsequent studies, the definition of the homolog was refined to be PDB entries that shared the same EC class and Pfam [11] family. The reason for this change is explained in more detail for the testing of the A set of 296 motifs below.

Testing against unrelated structures

Alignments were conducted in ProMOL using PDB entries chosen at random, using a copy of one of the PDB entry indices in Microsoft Excel, introducing a column of randomly generated numbers and sorting on this random number. Randomly chosen structures were limited to EC classes that differed at the first digit when compared to the motif definition.

Following these initial tests, the A set of motifs was created from selected entries in the CSA, which contains more than 18,000 motifs. The library of motif definitions was compiled from the literature-based entries in the Catalytic Site Atlas (CSA) database [3]; each entry contains a list of components from enzyme active sites:

- Residue type (amino acid)
- Residue number
- Chain
- Functional component (backbone, sidechain, or both)
- EC number of the native structure

Of the 18,000 motifs in the CSA, there are only several hundred distinct structural active site motifs with good specificity and good support in the literature. A library of 388 motifs was initially created with the automated tool based only on CSA motifs, which were chosen for the following characteristics:

- They are based on literature citations.
- They contain more than one amino acid residue.
- If an active site was found to occur in more than one chain in a PDB entry, the motif was only created if the catalytic site residues were the same in all identical chains (see Conclusions and Future Plans)

The automated motif collection was reduced from 388 to 296 due to a low number of homologs in some cases and to motif instability during testing for others. The Motif Maker tab in ProMOL was modified to include the algorithm to automatically create motifs based on catalytic site definitions found in the CSA.

Resources

The code is open source and available under a GPL license on SourceForge at https://sourceforge.net/p/sbevsl/code/HEAD/tree/trunk/promol_auto/. Automated motif generation is available in ProMOL version 5.4 r452 and greater.

Results and Discussion

We initially set out to validate the performance of the automated motif generator by comparing the predictive capability of automated motifs to manually defined motifs. The automated motif generator was used to render a duplicate of 112 of the existing 181¹

¹The motif generation and selection process is explained in Materials and Methods.

manually generated motif templates (P); the new motifs were designated AP motifs². For a given motif, we estimated the true positive rate as the percentage of known homologs against which the motif produces a positive hit. The false positive rate for a motif is estimated as the percentage of unrelated structures against which the motif produces a positive hit. Then the performance of the motif sets was compared in terms of hit rates in native structures, known homologs with the same EC and Pfam designation, and 100 randomly selected unrelated structures (Table 1). The two motif sets based on the same set of native structures but generated with different methods exhibited similar hit rates against homologous structures (70.25% and 71.59% for the P Set and the AP Set, respectively) and against unrelated structures (3.79% and 4.32% for the P Set and the AP Set, respectively). The true positive rate for both sets of motifs is different from the 62.63% value reported previously for the full P set of 181 motifs [1]. The differences result from the following changes in the matching algorithm: (1) the use of a more restricted definition of a true positive to be one with a Levenshtein distances = 1, rather than accepting Levenshtein distances of 2 and 3 as in the prior results, (2) the use in the present instance of a smaller set of the most reliable 112 motifs for which positive results with Levenshtein distances = 1 could be achieved, and (3) restricting the definition of a homolog to be a molecule agreeing in both EC number and Pfam family, rather than the EC number alone. Perhaps the greatest impact of adopting this new approach to homolog definition is a reduction of the false positive rate from the value of 18.51% reported previously [1] to ~4% reported here.

Common PDB Entries

Each motif was screened against itself, against homologs of the same EC and Pfam designation, and against 100 random unrelated structures. The hit rates and average RMSD values for structural alignments are comparable for the two sets generated with different methods.

In addition to comparing hit rates in homologous and unrelated structures, a second approach to testing the performance of the automated motifs vs. the manually defined motifs was by analysis of the variance in the average root-mean-square deviation (RMSD) measurements obtained from the respective local structural alignments. ProMOL has the built-in capability to calculate RMSD values between all atoms, α - and β -carbon atoms, and all α -carbon atoms in local structural alignments of motif template to the tested structure. The AP set motifs have slightly higher hit rates against homologous and unrelated structures and slightly higher average RMSD measurements (Table 1 and Figure 4). Combined, these results suggest higher sensitivity for AP Set motifs. Nevertheless, for both motif sets, the all atom RMSD measurements in alignments to homologs were found to concentrate predominantly within 0 to 1 Angstroms (Figure 4A and 4B). We observed no significant differences in average RMSD measurements between P and AP Set motifs (Figure 4C, and 4D). In addition, for both motif sets, the RMSD distributions indicate that the local alignments against homologs are of much higher quality than those against unrelated structures. Taken together, a comparison of the hit percentage on homologous or unrelated

²It should be noted that there is minimal overlap between the AP set and the A set described later in the manuscript.

structures and associated RMSD measurements indicates that automatically generated motifs perform just as well as manually generated motifs.

Density plot of all atom RMSD distribution in alignments of AP Set motifs to homologous and unrelated structures. (C, bottom left) Bar plot of all atom RMSD from alignments of P and AP motifs to their homologous structures, 95% CIs are shown. (D, bottom right) Bar plot of all atom RMSD from alignments of P and AP motifs to their unrelated structures, 95% CIs are shown. For both motif sets, the majority of RMSDs from unrelated structures concentrate between 1 and 5 Å and around 5% fall outside the range of 0 to 25 Å. Analysis of variance did not reveal significant differences in distribution of RMSD measurements obtained with P and AP Set motifs. Similar patterns were observed for α -carbon and α - and β -carbon RMSDs (results not shown).

After validating our approach to automated motif generation, we subsequently created 296 additional automatically generated motifs and tested them against all of their homologs, based on EC class - a total of 23,065 functionally annotated PDB entries. Using this approach, only 42% of homologs were found based on alignment with the appropriate motif templates by ProMOL. This number was significantly lower than the hit rate of 62.63% reported in [1], so we pursued the following series of structural and sequence alignment studies.

Even with these adjustments, the hit rate was still significantly lower than expected for some sets of homologs, so we explored two cases more deeply: EC class 4.2.3.3, the methylglyoxal synthases; and EC class 2.4.2.31, the NAD⁺-protein-arginine ADP-ribosyltransferases. EC class 4.2.3.3, the methylglyoxal synthases, has nine entries in the PDB. A multiple sequence alignment with MUSCLE [12] revealed that the six catalytic site residues listed in the Catalytic Site Atlas were highly conserved for all members of EC class 4.2.3.3, with the exception that two structures (PDB entries 1S8A and 1S89) are site-directed mutants at position 98 (Fig 5A) [13]. Motifs were generated manually for the seven structures containing all six active site residues: HIS-19, GLY-66, ASP-71, ASP-91, HIS-98, ASP-101 (residue numbers taken from PDB entry 1B93 are used for this discussion). For all generated motifs, the Tolerance Factor was kept at the default of 2.00 Å. In each case, the backbone atoms for GLY were added to the motif. Six of the seven motifs returned only the six correct residues when tested against themselves. In the case of 1VMD, the backbone atoms for the ASP-91 were included in the motif to eliminate additional residues when tested against itself. To see if the motif's native structure has an effect on its true positive hit rate, we screened the seven motifs against the nine PDB entries from EC class 4.2.3.3 (Figure 5). We found that motifs based on different homologous native structures exhibit variable true positive rates. For example, the motifs based on 1B93 [14] and 1IK4 [15], were each found in four of their eight homologs while the motif based on 1VMD [16] was found in six of its eight homologs; in two cases one residue from the motif template (ASP-101) was not detected in the query. The motif based on 2XW6 [17] was found in six of its homologs as well, but in four cases, residue ASP-101 was again not detected in the query. ProMOL did not align the two query structures containing site-directed mutations at position 98 (1S8A and 1S89) with any of the motif templates and these were omitted from Fig. 5. [13].

These results suggest subtle structural differences in active sites of enzymes of the same EC designation. In cases, such as the query of 1B93 against the motif template for 2XW6, ProMOL did not find ASP-101 in the query structure (Fig. 6). Four of the residues aligned quite nicely: HIS-19, GLY-66, ASP-71, and ASP-91. The two HIS-98 residues were somewhat distant from each other in the alignment, but had a similar orientation and were found by ProMOL. The ASP-101 residues, on the other hand, are oriented in different directions and were not found by the alignment algorithm in ProMOL. The results in Table 1 are pertinent to this argument as well. The RMSD values for the AP motifs (automatically generated versions of motifs from the P Set, which were manually generated) are about one tenth of an Å higher than with the P motifs. These results indicate that catalytic site structural variations within the same EC class can be significant enough to necessitate additional motif definitions to allow for successful prediction. For instance, inclusion of three motif definitions in ProMOL for the methylglyoxal synthases, based on 1B93, 1VMD, and 2XW6, would be sufficient to allow successful prediction of all the EC class members.

We performed a similar analysis of the NAD⁺-protein-arginine ADP-ribosyltransferases (EC class 2.4.2.31), which contain four catalytic residues: SER-147, GLU-159, ARG-184 and GLU-189. Only four of the fifteen EC 2.4.2.31 structures in the PDB contained all four catalytic residues. A motif template based on one of these four structures, 1OG1 [20], aligned well with 1GXY [21] in ProMOL/PyMOL. The other two structures were determined with different crystal lattices (1GXZ [21], 1GY0 [21]), which may explain the lack of alignment by ProMOL. Two of the structures contained site-directed mutants at this position (1OG3 E189I [20]; 1OG4 E189A [20]). In this case, then, the best possible alignment results in ProMOL would have been 4/15 (27%) but ProMOL actually found 2/15 alignments (13%).

Sequence comparison studies (either directly in PyMOL or with Clustal Omega [22]) revealed that a significant percentage of the overlooked homologs shared function with the templates but did not share the same active site residues because of residue substitution. For example, in the case of EC class 2.1.2.2, only 10 of 32 structures have all 4 active site residues based on a multiple sequence alignment in Clustal Omega. The other 22 PDB entries have threonine in place of serine in the active sites. ProMOL still finds 15 of 31³ homologs of PDB entry 1CDE [23], a serine-containing member of EC class 2.1.2.2. A motif based on the active site of PDB entry 3DA8 [24], which contains a threonine in place of the serine in the active site, aligned with 12 of 31³ homologs. This includes 7 that aligned with 1CDE (with a Levenshtein distance of 1, reflecting the threonine for serine substitution) and 5 that were uniquely aligned to 3DA8. All of the homologs that aligned with the 3DA8 template contained threonine. In the case of EC class 2.1.2.2, then, the addition of a separate motif containing threonine in place of serine, would have enabled ProMOL to identify 5 additional members of this EC class.

A number of other factors also contribute to a reduction in the hit rate.

³The discrepancy between 31 homologs and 32 structures is because the template structures (1CDE and 3DA8) are not counted as homologs in the comparison.

- *Truncated PDB entries.* PDB entry 1PL3 [25] (EC class 1.1.99.18) is a truncated structure that does not contain the active site residues.
- *Catalytic residue substitution.* A motif template based on PDB entry 1GA8 [26] failed to align with a homolog, 1SS9 [27], that is a member of the same EC class and is a member of the same Pfam family. Close examination of 1SS9 revealed an E189Q mutation in the active site. Any such substitution to a catalytic residue would interfere with current versions of ProMOL recognizing a homolog.
- *Residue differences in CSA active site definitions.* Careful study of the entries in the CSA revealed active sites from the same EC class that contained different residues. In the case of EC class 1.11.1.10, the following residues are listed for the active sites on four different structures.
 - 1A8Q [28] TRP-28 SER-94 MET-95 ASP-223 HIS-252
 - 1A7U [28] PHE-32 SER-98 MET-99 ASP-228 HIS-257
 - 1A8S [28] TRP-28 SER-94 THR-95 SER-119 ASP-224 HIS-253
 - 2CPO [29] HIS-105 GLU-183
- *Prosthetic groups and metal ions.* Many enzymes contain prosthetic groups that are not included in the existing motifs. For instance, PDB entry 1WNW [30] (Heme Oxygenase, EC class 1.14.99.3) contains a heme group that is not present in the CSA motif. This is addressed under future work below.

Even taking all of these examples into account, we still had cases for which the ProMOL motifs as used in Motif Finder identified active site residues in less than 20% of homologs, as defined by EC class. Sequence comparison of all members of these low-scoring EC classes revealed that these EC classes frequently contained members with poor sequence alignments that were missing some or all of the active site residues defined by the CSA for that EC class. For example, PDB entry 1ZRZ [31] is a member of EC class 2.7.11.13, which has 59 member structures in the PDB. ProMOL found active site residue alignments in 18 of the 59 structures. Of the 18 alignments in ProMOL, 15 were found to come from the same Pfam family, PF00069.

We saw this pattern repeatedly and, therefore, decided to limit the definition of homologs as those structures that shared both EC class and Pfam family. Using the 296 automatically generated motifs, we found that the hit rate was 66.2%.

Conclusions and Future Plans

According to our tests, motifs created by the automated motif maker perform at the same level as motifs that have been created and tested manually. In the process of testing, we have conducted alignments of over 23,000 PDB entries against 296 automatically generated motifs. The overall hit rate for homolog alignment is still lower than desirable (45% if we define homologs as being members of the same EC class), but we have identified a number of factors that contribute to the low hit rate. Enzymes are assigned to EC classes “according to the reaction they catalyse” (<http://www.chem.qmul.ac.uk/iubmb/enzyme/rules.html>),

rather than the sequence or structure of a protein. When homologs were defined as structures that shared the same EC class and Pfam family, the hit rate increased to between 66% and 72% for the A set, the P set and the AP set. Other factors that affect the hit rate are related to changes in the active site residues, since alignments in ProMOL are based on active site templates. These include active site residue substitutions, either by nature or by design, truncated protein structures that lack some or all active site residues, and the presence of metal ions and prosthetic groups in enzyme active sites.

Future versions of ProMOL will address these concerns. The next version of ProMOL will include common residue substitutions, such as SER/THR, ASP/GLU, and ASN/GLN, as well as metal ions and common prosthetic groups. As was mentioned in the case of methylglyoxal synthases, we often find that one template in a particular EC class or Pfam family is not effective in identifying all homologs, sometimes even in cases where we see more than 50% sequence identity, which suggests that including more than one motif definition from the same EC class increases the chance of correct function assignment. This is analogous to Pfam families, which are not defined by single sequences, but are built from “seeds” the minimal set of sequences needed to build the family [11]. In order to completely cover an EC class, we may need to create motifs from multiple enzymes in that family. Originally our focus was on structures that were characterized by a single active site replicated identically in all chains. While useful in some cases, we have found this to be unnecessary, and will adopt the Pfam approach of treating each chain as an independent molecule, extending the motifs library and search algorithm to identify motifs that are different in different chains, or active sites that exist at interfaces between two chains. We also plan to develop methods to identify structures that lack the active site residues, either because of truncation or site-directed mutagenesis.

Up to this point, we have limited the number of catalytic site motifs to achieve reasonable ProMOL/PyMOL search times. With the implementation of our database (<http://promoldb.blogspot.com/>, manuscript in preparation), this is less of a concern. Automated motif generation will facilitate this process and should enable us to create and test over 10,000 new motifs in the next year, based on the pool of active site templates in the CSA. We then plan to use our expanded motif library to analyze all protein structures in the PDB and enter these results in our database. A database-ready version of ProMOL is under development that first will scan the database for existing alignments and then return these to the user (a process that will take seconds over a normal Internet connection), rather than the current 5–10 minute run time for each individual protein query.

Acknowledgments

The authors thank the invaluable assistance of Frances C. Bernstein and current and former students who have worked at Dowling College and at RIT on the SBEVSL project. Funding for the project has been provided by NIH GM078077, NSF-DUE 0402408, Rochester Institute of Technology and Dowling College.

References

1. Hanson B, Westin C, Rosa M, Grier A, Osipovitch M, MacDonald ML, Dodge G, Boli PM, Corwin CW, Kessler H, McKay T, Bernstein HJ, Craig PA. Estimation of protein function using template-

- based alignment of enzyme active sites. BMC Bioinformatics. 2014; 15:87.10.1186/1471-2105-15-87 [PubMed: 24669788]
2. Delano W. Pymol: An open-source molecular graphics tool. CCP4 Newsl Protein Crystallogr. 2002; 40:44–53.
 3. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res. 2004; 32:D129–D133.10.1093/nar/gkh028 [PubMed: 14681376]
 4. McKay T, Hart K, Horn A, Kessler H, Dodge G, Bardhi K, Bardhi K, Mills JL, Bernstein HJ, Craig PA. Annotation of proteins of unknown function: initial enzyme results. J Struct Funct Genomics. 2015; 16:43–54.10.1007/s10969-015-9194-5 [PubMed: 25630330]
 5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242.10.1093/nar/28.1.235 [PubMed: 10592235]
 6. West AH, Martinez-Hackert E, Stock AM. Crystal Structure of the Catalytic Domain of the Chemotaxis Receptor Methyltransferase, CheB. J Mol Biol. 1995; 250:276–290.10.1006/jmbi.1995.0376 [PubMed: 7608974]
 7. Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Sov Phys Dokl. 1966; 10:707–710.
 8. Djordjevic S, Goudreau PN, Xu Q, Stock AM, West AH. Structural basis for methyltransferase CheB regulation by a phosphorylation-activated domain. Proc Natl Acad Sci USA. 1998; 95:1381–1386. [PubMed: 9465023]
 9. Altschul S, Gish W, Miller W, Myers EW, Lipman D. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410. [PubMed: 2231712]
 10. Holm L, Rosenström P. Dali server: conservation mapping in 3D. Nucleic Acids Res. 2010; 38:W545–549. [PubMed: 20457744]
 11. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. Pfam: the protein families database. Nucleic Acids Res. 2014; 42:D222–D230.10.1093/nar/gkt1223 [PubMed: 24288371]
 12. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797.10.1093/nar/gkh340 [PubMed: 15034147]
 13. Marks GT, Susler M, Harrison DHT. Mutagenic Studies on Histidine 98 of Methylglyoxal Synthase: Effects on Mechanism and Conformational Change. Biochemistry. 2004; 43:3802–3813.10.1021/bi035838o [PubMed: 15049687]
 14. Saadat D, Harrison DH. The crystal structure of methylglyoxal synthase from *Escherichia coli*. Structure. 1999; 7:309–317.10.1016/S0969-2126(99)80041-0 [PubMed: 10368300]
 15. Marks GT, Harris TK, Massiah MA, Mildvan AS, Harrison DHT. Mechanistic Implications of Methylglyoxal Synthase Complexed with Phosphoglycolohydroxamic Acid As Observed by X-ray Crystallography and NMR Spectroscopy. Biochemistry. 2001; 40:6805–6818.10.1021/bi0028237 [PubMed: 11389594]
 16. Joint Center for Structural Genomics (JCSG). Crystal structure of Methylglyoxal synthase (TM1185) from *Thermotoga maritima* at 2.06 Å resolution. 200410.2210/pdb1vmd/pdb
 17. Shahsavari A, Erfani Moghaddam M, Antonyuk SV, Khajeh K, Naderi-Manesh H. Atomic Resolution Structure of Methylglyoxal Synthase from *Thermus* Sp. Gh5 Bound to Phosphate: Insights Into the Distinctive Effects of Phosphate on the Enzyme Structure. 201010.2210/pdb2xw6/pdb
 18. Sugahara M, Kunishima N. Crystal structure of methylglyoxal synthase from *Thermus thermophilus* HB8. 200410.2210/pdb1wo8/pdb
 19. Shahsavari A, Erfani Moghaddam M, Antonyuk SV, Khajeh K, Naderi-Manesh H. Crystal Structures of Methylglyoxal Synthase from *Thermus* Sp. Gh5 in the Open and Closed Conformational States Provide Insight Into the Mechanism of Allosteric Regulation. 201010.2210/pdb2x8w/pdb
 20. Ritter H, Koch-Nolte F, Marquez VE, Schulz GE. Substrate Binding and Catalysis of Ecto-ADP-ribosyltransferase 2.2 from Rat. Biochemistry. 2003; 42:10155–10162.10.1021/bi034625w [PubMed: 12939142]

21. Mueller-Dieckmann C, Ritter H, Haag F, Koch-Nolte F, Schulz GE. Structure of the Ecto-ADP-ribosyl Transferase ART2.2 from Rat. *J Mol Biol.* 2002; 322:687–696.10.1016/S0022-2836(02)00818-5 [PubMed: 12270706]
22. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol.* 2011; 7:539.10.1038/msb.2011.75 [PubMed: 21988835]
23. Almasy RJ, Janson CA, Kan CC, Hostomska Z. Structures of apo and complexed *Escherichia coli* glycinamide ribonucleotide transformylase. *Proc Natl Acad Sci USA.* 1992; 89:6114–6118. [PubMed: 1631098]
24. Zhang Z, Caradoc-Davies TT, Dickson JM, Baker EN, Squire CJ. Structures of Glycinamide Ribonucleotide Transformylase (PurN) from *Mycobacterium tuberculosis* Reveal a Novel Dimer with Relevance to Drug Discovery. *J Mol Biol.* 2009; 389:722–733.10.1016/j.jmb.2009.04.044 [PubMed: 19394344]
25. Rotsaert FAJ, Hallberg BM, de Vries S, Moenne-Loccoz P, Divne C, Renganathan V, Gold MH. Biophysical and Structural Analysis of a Novel Heme b Iron Ligation in the Flavocytochrome Cellobiose Dehydrogenase. *J Biol Chem.* 2003; 278:33224–33231.10.1074/jbc.M302653200 [PubMed: 12796496]
26. Persson K, Ly H, Dieckelmann M, Wakarchuk W, Withers S, Strynadka N. Crystal structure of the retaining galactosyltransferase LgtC from *Neisseria meningitidis* in complex with donor and acceptor sugar analogs. *Nat Struct Biol.* 2001; 8:166–175. [PubMed: 11175908]
27. Lairson LL, Chiu CPC, Ly HD, He S, Wakarchuk WW, Strynadka NCJ, Withers SG. Intermediate Trapping on a Mutant Retaining α -Galactosyltransferase Identifies an Unexpected Aspartate Residue. *J Biol Chem.* 2004; 279:28339–28344.10.1074/jbc.M400451200 [PubMed: 15075344]
28. Hofmann B, Tölzer S, Pelletier I, Altenbuchner J, van Pée KH, Hecht HJ. Structural investigation of the cofactor-free chloroperoxidases. *J Mol Biol.* 1998; 279:889–900.10.1006/jmbi.1998.1802 [PubMed: 9642069]
29. Sundaramoorthy M, Turner J, Poulos TL. The crystal structure of chloroperoxidase: a heme peroxidase cytochrome P450 functional hybrid. *Structure.* 1995; 3:1367–1378.10.1016/S0969-2126(01)00274-X [PubMed: 8747463]
30. Matsui T, Furukawa M, Unno M, Tomita T, Ikeda-Saito M. Roles of Distal Asp in Heme Oxygenase from *Corynebacterium diphtheriae*, HmuO: A Water-Driven Oxygen Activation Mechanism. *J Biol Chem.* 2005; 280:2981–2989.10.1074/jbc.M410263200 [PubMed: 15528205]
31. Messerschmidt A, Macieira S, Velarde M, Baedeker M, Benda C, Jestel A, Brandstetter H, Neufeind T, Blaesse M. Crystal Structure of the Catalytic Domain of Human Atypical Protein Kinase C- ι Reveals Interaction Mode of Phosphorylation Site in Turn Motif. *J Mol Biol.* 2005; 352:918–931.10.1016/j.jmb.2005.07.060 [PubMed: 16125198]

1A

1B

Figure 1.

Motif Definition and Matching Subset. (A) ProMOL's Motif Maker interface holding the motif definition based on the active site of 1CHD [6], a chemotaxis receptor methyltransferase. The motif is defined by the PDB ID and EC designation of the native structure, amino acid types, amino acid positions, chain affiliations, and functional components of active residues. (B) Graphical representation of the corresponding atoms included in the active site motif definition of 1CHD. Selected distances between β -carbons are shown. Query structures can

be screened for motifs composed of the same residue types and within similar distances of each other.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

2A

2B

Figure 2.

Screening Query Structures. (A) ProMOL's Motif Finder interface showing the result of screening the active site motif of 1CHD, a chemotaxis receptor methylesterase, against one of its homologs, PDB entry 1A2O. The results include a Levenshtein distance of 0 [7] and RMSD values for local structural alignments. (B) A local structural alignment of 1CHD (white) with its homolog, 1A2O (red) [8]. All catalytic residue types and numbers are identical in the two methylesterases. Moreover, this functional similarity is not detected with

BLAST [9] and while Dali [10] is able to recognize 1A2O as structurally similar, the similarity is only moderate (Z-score 5.1, RMSD 3.5 Å, identity 9%).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3.

Automated Motif Generation Algorithm. When the motif is tested against itself (*i.e.* against the structure on which it is based), ProMOL returns a matching subset, a collection of residues that satisfy the distance constraints. The matching subset may include residues that are not found in the motif definition. The algorithm stores the motif only if the matching subset returns an identical collection of residues as in the motif definition.

4A

4B

4C

4D

Figure 4. Distribution of All Atom RMSD Measurements. The RMSD measurements were calculated by ProMOL based on all atoms from local structural alignments of P and AP motifs against homologous and unrelated structures. (A, top left) Density plot of all atom RMSD distribution in alignments of P Set motifs to homologous and unrelated structures. (B, top right)

```

1WO8 -----MKALALIAHDAKKDEMVAFCRLHKDVLARYPLLATGTTGARIQEAT
2X8W -----MRALALIAHDAKKEEMVAFQQRHREVLARFPLVATGTTGRRIEEAT
2XW6 -----SHMRALALIAHDAKKEEMVAFQQRHREVLARFPLVATGTTGRRIEEAT
1S89 MELTTRT----LPARKHIALVAHDHCKQMLMSWVERHQPLLEQHVLYATGTTGNLISRA
1S8A MELTTRT----LPARKHIALVAHDHCKQMLMSWVERHQPLLEQHVLYATGTTGNLISRA
1B93 MELTTRT----LPARKHIALVAHDHCKQMLMSWVERHQPLLEQHVLYATGTTGNLISRA
1EGH MELTTRT----LPARKHIALVAHDHCKQMLMSWVERHQPLLEQHVLYATGTTGNLISRA
1IK4 MELTTRT----LPARKHIALVAHDHCKQMLMSWVERHQPLLEQHVLYATGTTGNLISRA
1VMD MSDRPRRYKIFMDKKKRIALIAHDDRKRDLLEWVSFNLGTLKHELYATGTTGALLQEKL
      . :*:*** * :: : : * .. * ***** :.

1WO8 GLAVERVLSGPLGGDLQIGARVAEGKVLAVVFLQDPLTAKPHEPDVQALMRVCNVHGVPL
2X8W GLTVEKLLSGPLGGDQMGARVAEGRILAVIFFRDPLTAQPEPDVQALLRVCDVHGVPL
2XW6 GLTVEKLLSGPLGGDQMGARVAEGRILAVIFFRDPLTAQPEPDVQALLRVCDVHGVPL
1S89 GMNVNAMLSPMGGDQQVGALISEGKIDVLIFFWDPLNAVPHDPDKALLRLATVWNIPV
1S8A GMNVNAMLSPMGGDQQVGALISEGKIDVLIFFWDPLNAVPHDPDKALLRLATVWNIPV
1B93 GMNVNAMLSPMGGDQQVGALISEGKIDVLIFFWDPLNAVPHDPDKALLRLATVWNIPV
1EGH GMNVNAMLSPMGGDQQVGALISEGKIDVLIFFWDPLNAVPHDPDKALLRLATVWNIPV
1IK4 GMNVNAMLSPMGGDQQVGALISEGKIDVLIFFWDPLNAVPHDPDKALLRLATVWNIPV
1VMD GLKVHRLKSGPLGGDQIGAMIAEGKIDVLIFFWDPLEPQAHDPDKALIRIATVYNIPV
      *: * : ***:*** *:*** :*: :. :*: :*** . :*: ***:***: * :*:

```

Motif	Query Structures						
	1B93	1VMD	1EGH	1IK4	2X8W	1WO8	2XW6
1B93							
1VMD						ASP-101	ASP-101
1EGH							
1IK4							
1WO8							
2X8W							
2XW6	ASP-101	ASP-101	ASP-101	ASP-101			

Figure 5.

Methylglyoxal Synthase Motif Testing. (A). A sequence alignment of the nine members of EC 4.2.3.3 for which PDB entries exist. Conserved catalytic residues are shown in red and active site mutants are shown in green. (B). To investigate the hit rates among homologs, we created seven motifs from EC Class 4.2.3.3. Blue blocks indicate positive hits that contained all residues. Individual residues are identified for cases where five residues aligned, but one of the residues in the motif template was not found in the query structure. Additional citations include 1EGH [14], 1WO8 [18], 2X8W [19].

Figure 6.

Catalytic site residue alignment for 1B93 and 2XW6. The query structure (1B93) is shown in red while the motif template (2XW6) is shown in white. The closely aligned residues can be found at the top of the image, while the more distant HIS-98 and the undetected ASP-101 are at the bottom of the image. The dotted lines represent the differences between the β carbons. The longer β -carbon distance for ASP-101 may explain why it was not found by the alignment algorithm.

Table 1
Quality Assessment of 112 Manually and Automatically Generated Motifs Based on

Motif Set	Generation Method	Query Structures	Hit Rate (%)	All Atom RMSD (Å)	α-Carbon RMSD (Å)	α- & β-Carbon RMSD (Å)
P Subset	Manual	Native	97.32	~0	~0	~0
		Homolog	70.25	0.58	0.48	0.51
		Unrelated	3.79	8.26	7.50	7.58
AP Set	Automatic	Native	100.00	~0	~0	~0
		Homolog	71.59	0.66	0.55	0.58
		Unrelated	4.32	9.35	8.57	8.71