Automated Protein Structural Motif Generation in ProMOL

Mikhail Osipovitch[1], Paul A. Craig[2], Herbert J. Bernstein[3]
[1]RIT School of Life Sciences; [2]RIT School of Chemistry and Materials Science; [3]Dowling College, Department of Mathematics and Computer Science

ABSTRACT

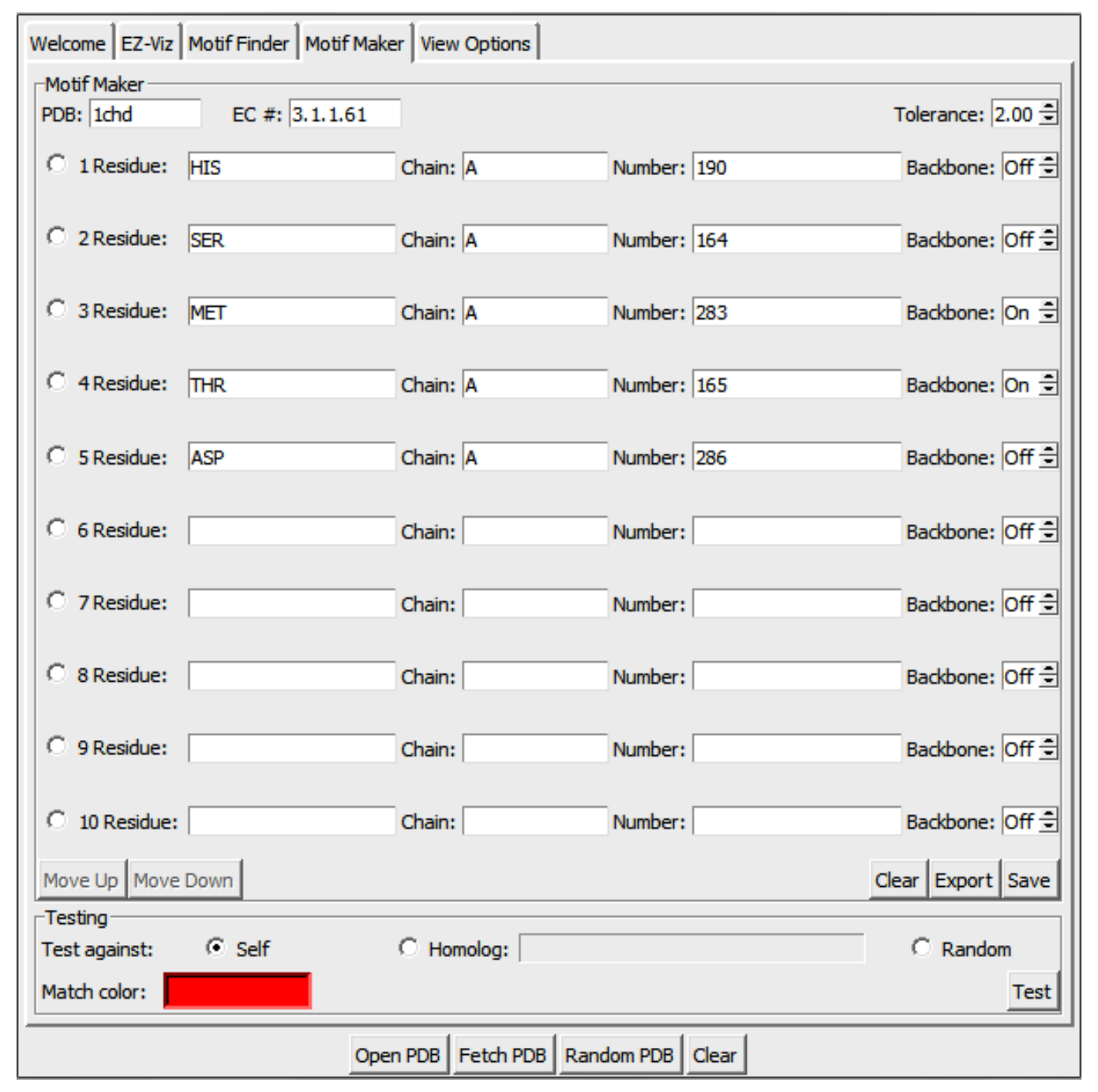
ProMOL, an extension for molecular graphics system PyMOL, is a structure-based protein function prediction software. ProMOL implements a set of routines for building motif templates screening them against query structures. Presently, each motif is generated individually and requires user intervention in optimization of parameters for sensitivity and selectivity. An algorithm was developed to automate motif building and testing routines. The algorithm uses a set of empirically derived parameters for optimization, requires little user intervention, and provides a possibility for extending the ProMOL library of motif templates at a higher pace. As a result, 388 motifs were generated automatically as an expansion of the library of 181 individually generated ones. The new motif templates exhibited comparable performance to the existing ones in terms of hit rates against native structures, homologous of the same EC designation, and randomly selected unrelated structures of a different EC designation at the first EC digit, as well as in terms of RMSD values obtained from structural alignments of motif templates and matching subsets of query structures. The research is supported by the NSF and NIH.

INTRODUCTION

ProMOL is written in Python and is an extension for the molecular visualization system PyMOL. As part of the Structural Biology Extensible Visualization Scripting Language (SBEVSL) initiative, ProMOL is developed in an attempt to classify proteins of known structure but an unknown function. The structure-based function prediction method involves compilation of an extensive library of structural motifs based on active sites that are defined in Catalytic Site Atlas (CSA) [1]. In ProMOL, the motifs are defined by the PDB ID's of the native structures, EC designations, types, numbers, functional components, and chain affiliations of catalytic site residues ***(Figure 1)***. The motifs are saved as Python script files and include stringency parameters such as Tolerance Factor (2.00 Å by default) and backbone atoms inclusion. Presently, the motifs are generated individually using the ProMOL's Motif Maker interface. The Motif Finder interface is used to asses the performance by testing the motif against its native, homologous, and unrelated structures ***(Figure 2)***. Additionally, the Motif Finder can screen the generated motifs against proteins of unknown function. The method of motif search used by ProMOL is based upon relative distances between catalytic site residues. In a query structure, the algorithm first seeks by residue type in the CSA-defined motif (e.g., all aspartates), then looks at the distances for each of the aspartates to all other residues found in the catalytic site (e.g., serines and histidines in serine proteases).

Presently, ProMOL library contains 181 motifs that were generated individually with the Motif Maker interface, and 261 JESS motifs that were converted from XML into Python files. As of today, Protein Data Bank (PDB) contains over 3,500 protein structures without an assigned function [2]. Expanding the motif library increases the predictive capacity of ProMOL. The larger the number of distinct functional motifs in the library, the greater the chance of inferring the function of an unclassified protein. To this end, an algorithm was developed to automate the motif generation and testing routines of ProMOL to create motifs at a higher pace and in a more systematic fashion.

1A



PDB: 1chd EC #: 3.1.1.61 Tolerance: 2.00

	Residue	Chain	Number	Backbone
○ 1	HIS	A	190	Off
○ 2	SER	A	164	Off
○ 3	MET	A	283	On
○ 4	THR	A	165	On
○ 5	ASP	A	286	Off
○ 6				Off
○ 7				Off
○ 8				Off
○ 9				Off
○ 10				Off

Test against: ● Self ○ Homolog: ○ Random
Match color:

1B

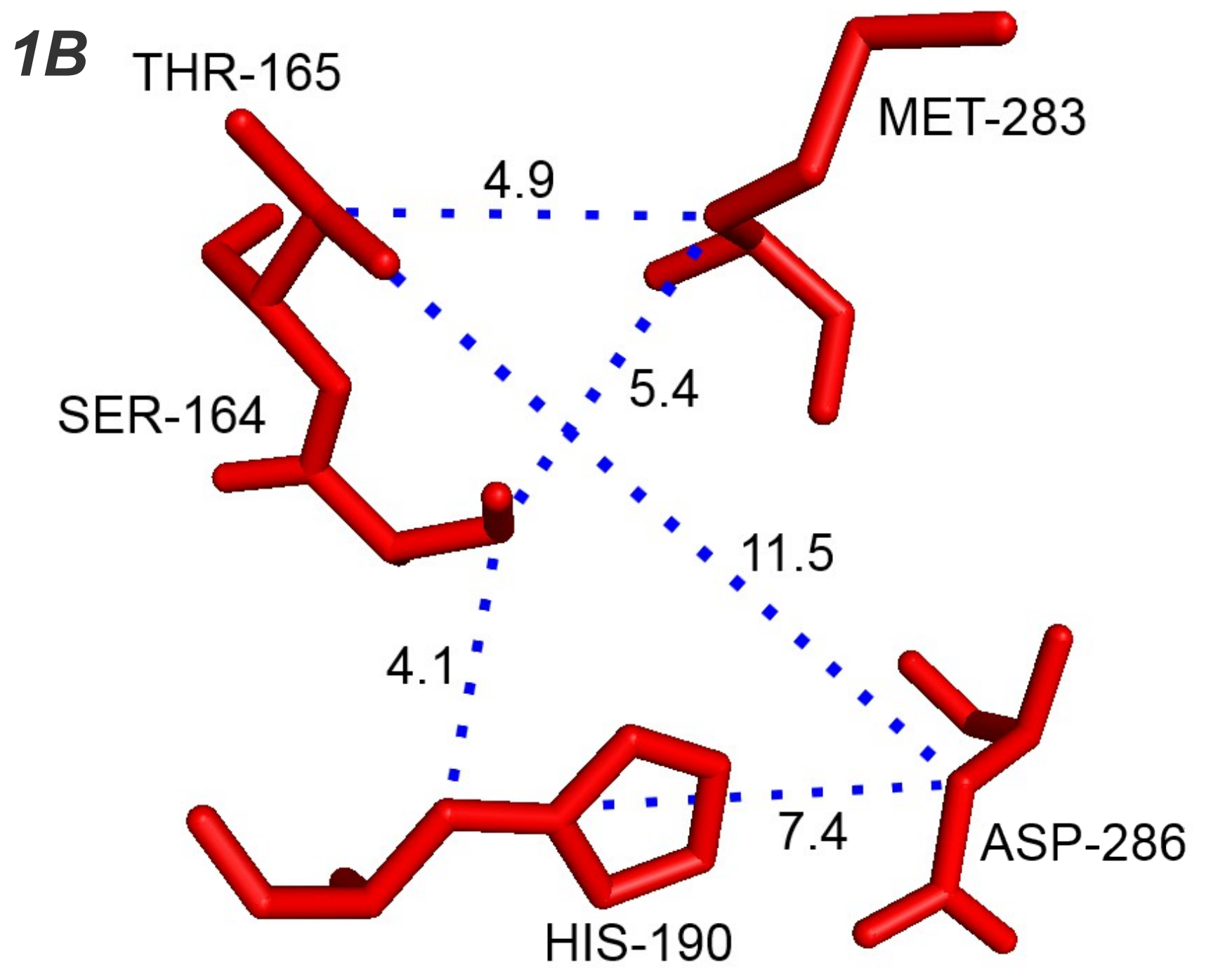
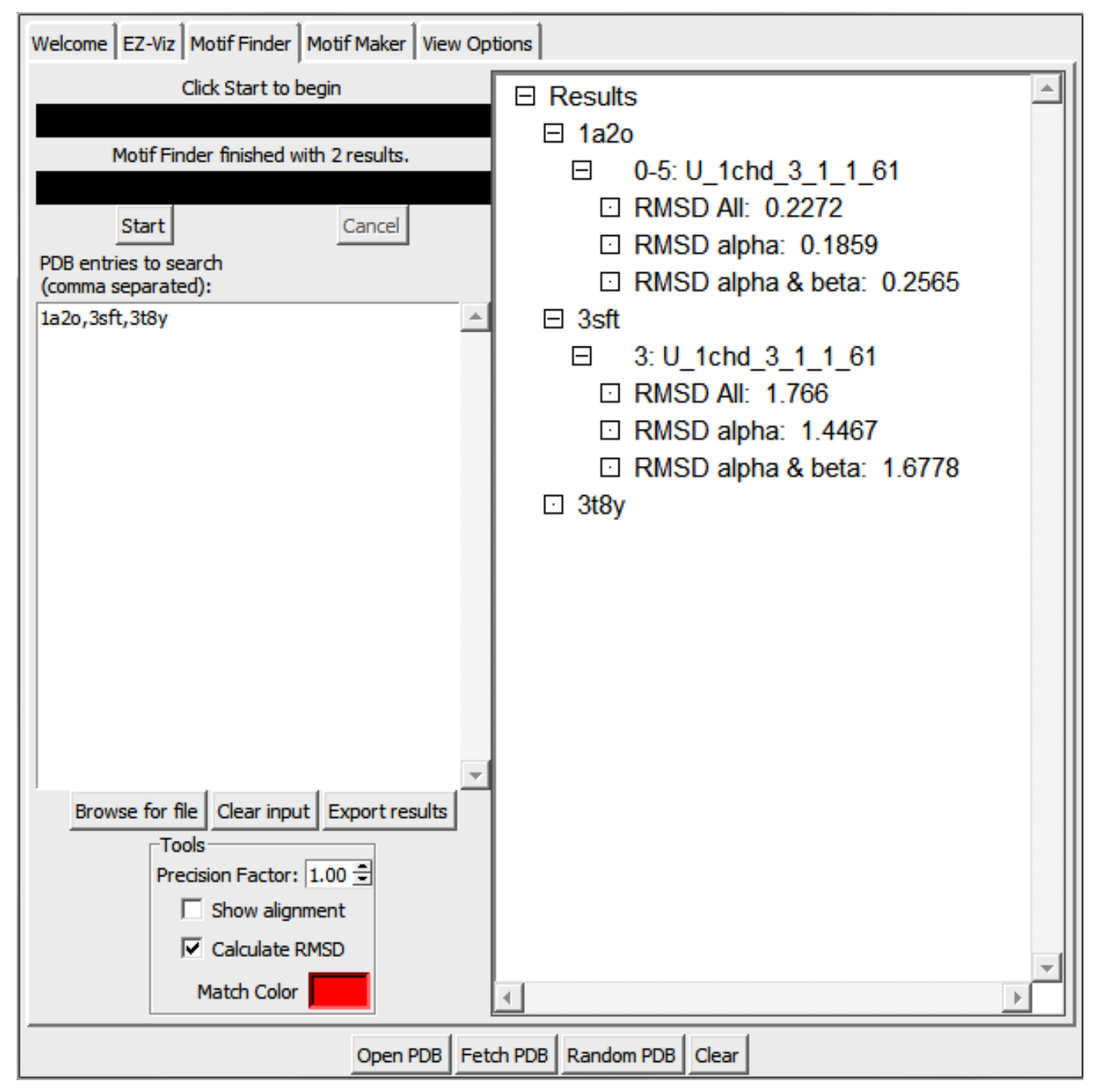


Figure 1**: Motif Definition and Matching Subset.* ***(A) ProMOL's Motif Maker interface holding the motif definition based on the structure of 1CHD, a chemotaxis receptor methylesterase. The motif definition includes the PDB ID and the EC designation of the native structure and corresponding residues types, chain affiliation, and numbers of active residues. ***(B)*** Graphical representation of the matching subset corresponding to the active site motif definition of 1CHD. Selected distances between β carbons are shown. Unknown structures can be screened for motifs composed of the same residue types and within similar distances of each other.

2A



Click Start to begin
Motif Finder finished with 2 results.
PDB entries to search (comma separated):
1a2o,3sft,3t8y

Precision Factor: 1.00
☐ Show alignment
☑ Calculate RMSD
Match Color

Results
- 1a2o
  - 0-5: U_1chd_3_1_1_61
    - RMSD All: 0.2272
    - RMSD alpha: 0.1859
    - RMSD alpha & beta: 0.2565
- 3sft
  - 3: U_1chd_3_1_1_61
    - RMSD All: 1.766
    - RMSD alpha: 1.4467
    - RMSD alpha & beta: 1.6778
- 3t8y

2B

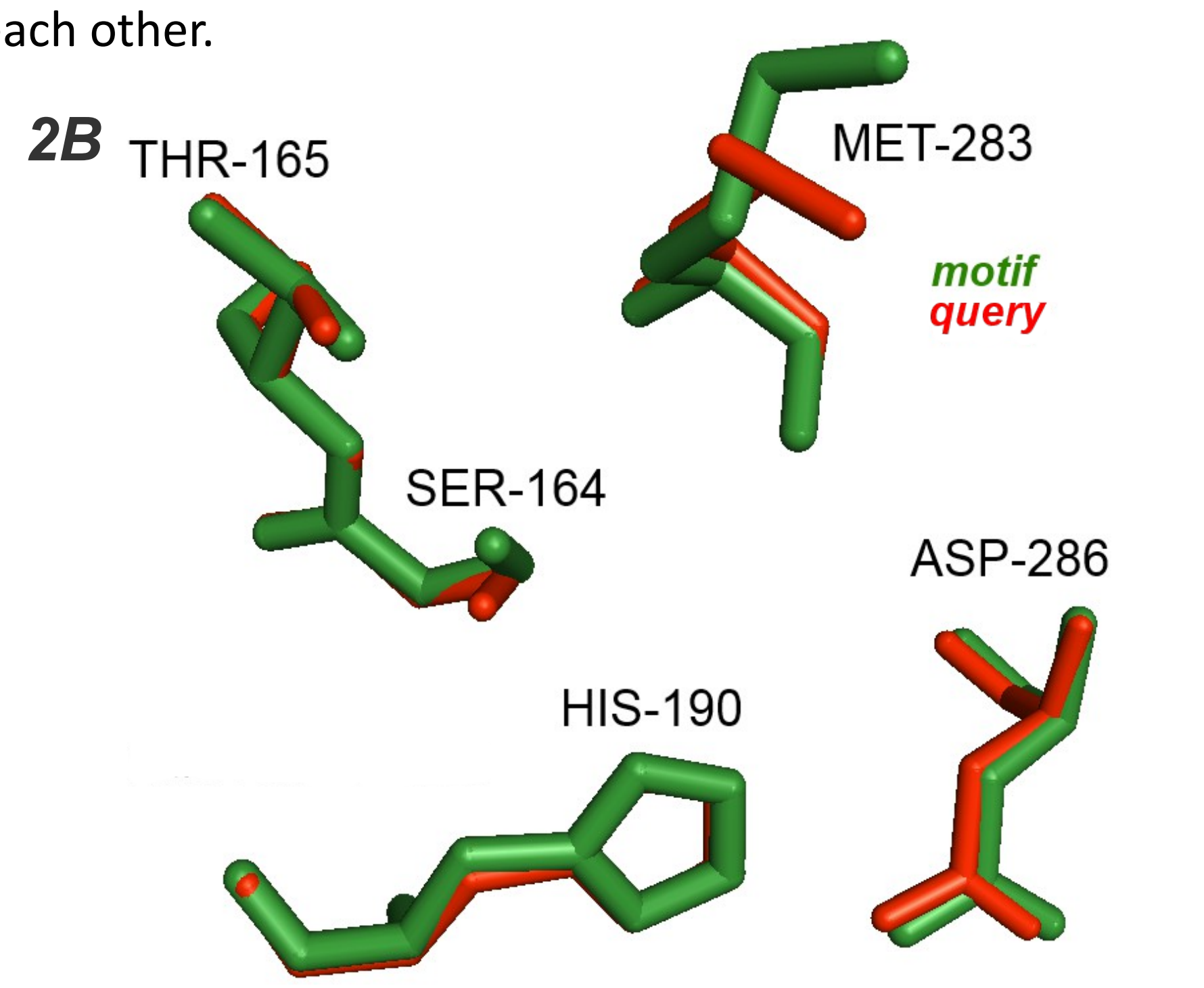


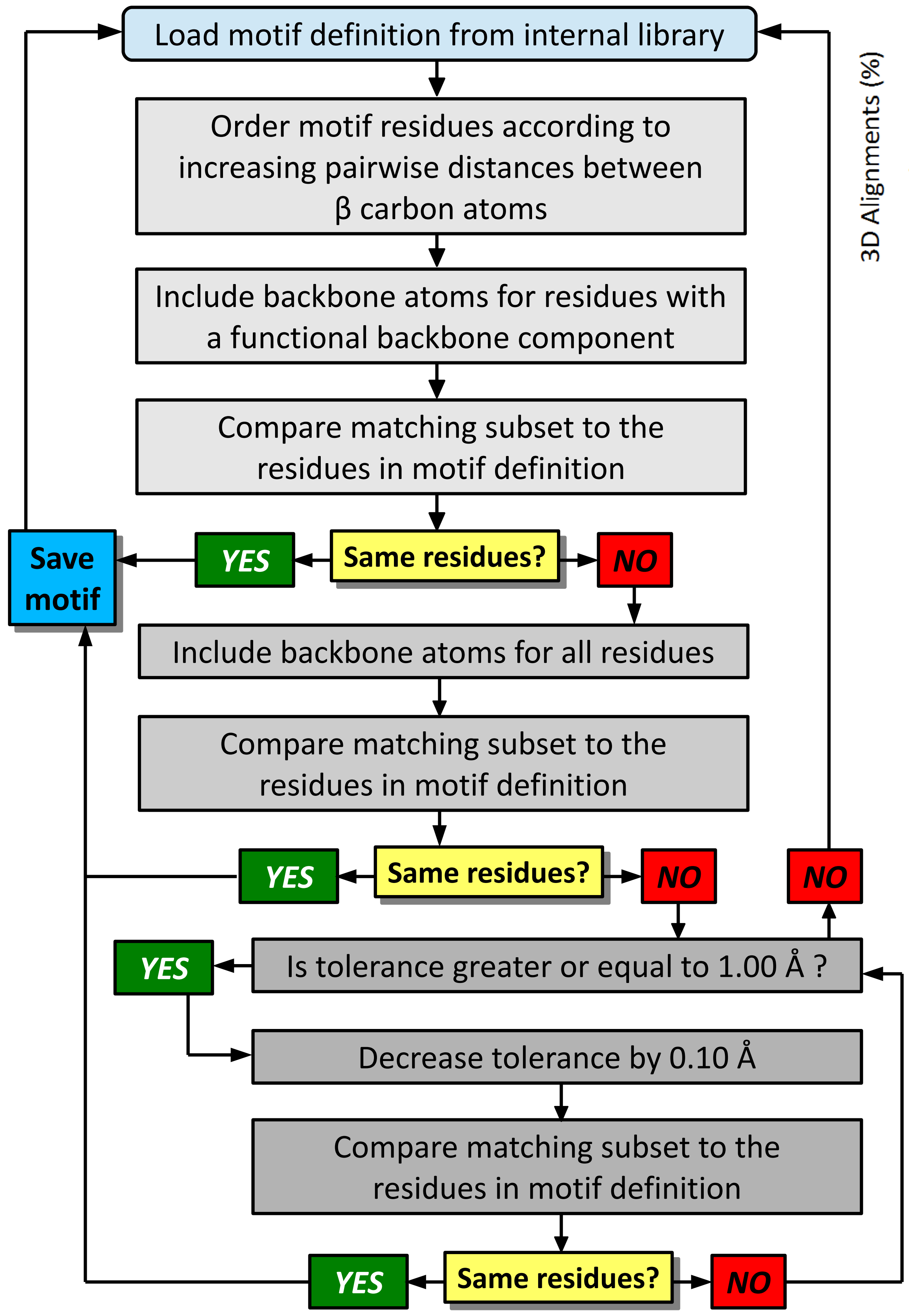
Figure 2**: Screening Homologous Structures.* ***(A) ProMOL's Motif Finder interface showing the result of screening the active site motif of 1CHD, a chemotaxis receptor methylesterase, against its homologues. The results include the Levenshtein distances and RMSD values for structural alignments. ***(B)*** A structural alignment of 1CHD (green) with its homologous structure 1A2O (red). All catalytic residue types and numbers are identical in the two methylesterases.

MATERIALS and METHODS

Motif Generation. A library of motif definitions was compiled with the CSA database version 2.2.12 [1]. Motif definitions were selected from CSA literature-based entries. Each definition corresponded to a catalytic site composed of the same set of two or more residues on one or more chains, and contained no catalytic co-factors. The electronic library was used to supply the information into ProMOL's Motif Maker routines. The Python module handling the motif generation was modified to include the algorithm to automatically create motifs based on the motif definitions library ***(Figure 3)***.

Quality Assessment. To asses the validity of the method, we used the automated algorithm to render a duplicate of the individually generated P Set motifs. The performance of the two identical sets was compared in terms hit rates in native structures, known homologues with the same EC designation, and 100 randomly selected unrelated structures with a different EC designation at the first EC digit. Similarly, we assessed the quality of the unique motif templates rendered with the automatic generation algorithm. In addition to the hit rates, the performance of individually and automatically generated motifs was assessed in terms of the average root-mean-squared deviation (RMSD) values obtained from local structural alignments.

- For the number of motif definitions:


Load motif definition from internal library
Order motif residues according to increasing pairwise distances between β carbon atoms
Include backbone atoms for residues with a functional backbone component
Compare matching subset to the residues in motif definition
Save motif
YES
Same residues?
NO
Include backbone atoms for all residues
Compare matching subset to the residues in motif definition
YES
Same residues?
NO
NO
YES
Is tolerance greater or equal to 1.00 Å ?
Decrease tolerance by 0.10 Å
Compare matching subset to the residues in motif definition
YES
Same residues?
NO

- Stringency control of motif parameters:

▢ : least stringent ▨ : moderately stringent ■ : most stringent

***Figure 3**:* Automated motif generation algorithm. When motif is screened in the native structure, ProMOL returns a matching subset, a collection of residues that satisfy the distance constraint. The matching subset may include residues that are not found in the motif definition. The algorithm only saves the motifs if matching subset returns the identical collection of residues as in the motif definition. The motif stringency is increased by including backbone atoms and decreasing the tolerance factor. Higher stringency promotes selectivity when motifs are screened against unassigned structure; however, overly stringent motifs may result in diminished sensitivity.

RESULTS and DISCUSSION

Motif Set	*Generation Method*	*Screening Structures*	*Hit Rate (%)*	*All Atom RMSD (Å)*	*α Carbon RMSD (Å)*	*α & β Carbon RMSD (Å)*
P Set (181 Motifs)	Individual	Native	96.69	~0	~0	~0
		Homologues	62.63	1.67	1.50	1.54
		Unrelated	18.64	6.96	6.27	6.35
AP Set (181 Motifs)	Automatic	Native	100.00	~0	~0	~0
		Homologues	63.73	2.08	1.98	2.02
		Unrelated	18.20	7.32	6.62	6.72
A Set (388 Motifs)	Automatic	Native	100.00	~0	~0	~0
		Homologues	51.51	1.67	1.46	1.52
		Unrelated	16.52	6.54	5.84	5.92

***Figure 4**: Quality Assessment of Individually and Automatically Generated Motifs*. The quality of individually and automatically generated motifs was assessed by screening each motif against its native structure, homologues of the same EC designation, and 100 randomly selected unrelated structures of a different EC designation at the first EC digit. The AP Set was generated automatically and represents a duplicate of the individually generated P Set. The hit rates and RMSD values for structural alignments are nearly identical for the two sets generated with different methods.

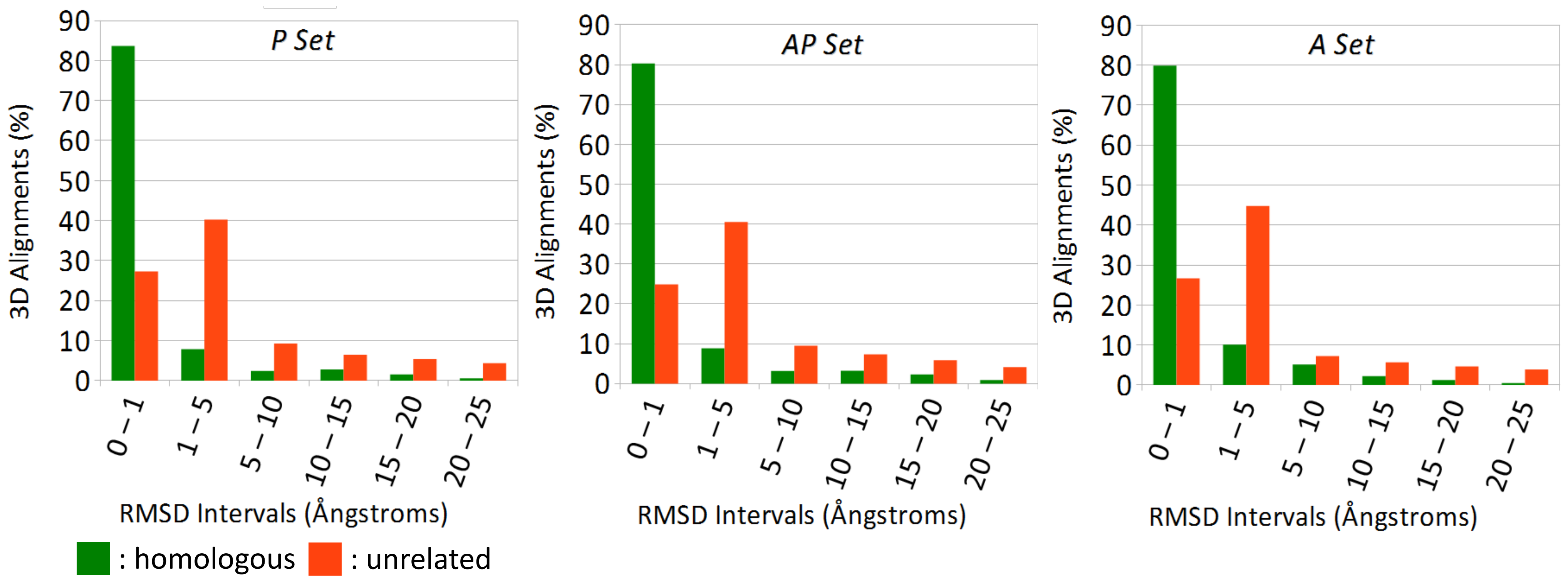


***Figure 5**: All Atom RMSD Distribution*. The RMSD values were obtained from structural alignments of motifs of P, AP, and A Sets against their homologous and unrelated structures. ProMOL calculates the RMSD values only when a motif is found in the query structure. For all three motif sets, around 80% of RMSD's from homologous structures fall between 0 and 1 Ångstroms. Of the RMSD's from unrelated structures, only around 25% are found in that range. For all three motif sets, the majority of RMSD's from unrelated structures concentrate between 1 and 5 Ångstroms and around 5% fall outside of the range of 0 to 25 Ångstroms. Similar patterns were observed for α Carbon and α & β Carbon RMSD's (results not shown). The RMSD distributions indicate that alignments against homologous structures are of higher quality than the alignments against unrelated structures.

Motif	*Screened Structures*								
	1B93	1VMD	1EGH	1IK4	1WO8	1S8A	1S89	2X8W	2XW6
1B93	🟩		🟩	🟩				🟩	
1VMD	🟩	🟩	🟩	🟩	🟩			🟩	🟩
1EGH	🟩		🟩	🟩					
1IK4	🟩		🟩	🟩				🟩	
1WO8	🟩				🟩			🟩	
1S8A						🟩			
1S89							🟩		

***Figure 6**: Methylglyoxal Synthase Motifs*. To investigate the low hit rates in homologues, we created 7 motifs based on structures of the same EC designation keeping all parameters constant. Green blocks indicate positive hits. The motif based on 1B93 is a part of the A Set and apart from the native structure, it was found in 3 out of 8 homologues while the motif based 1VMD was found in 6 out 8 homologues. The results suggest structural differences in active sites of enzymes of the same reaction type.

CONCLUSIONS and FUTURE PLANS

It is intended to work on further optimization of motif parameters influencing the sensitivity and selectivity of existing and newly generated motifs. Additionally, the investigations suggest that including more than one motif of the same enzyme superfamily increases the chance of correct function assignment. Meanwhile, present work has tripled the size of ProMOL's motif library. As compared to individually created motifs, the automatic generation produced motifs of comparable quality in terms of hit rates in homologous and unrelated structures. Moreover, the automated algorithm allows for a highly systematic and time-saving motif generation approach with a potential of creating new motifs at a higher pace to expand the predictive capacity of ProMOL.

REFERENCES

[1] Porter Craig T., Bartlett Gail J., Thornton Janet M. The Catalytic Site Atlas: a Resource of Catalytic Sites and Residues Identified in Enzymes Using Structural Data. (2004) *Nucleic Acids Research*, 32: D129-D133
[2] Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. The Protein Data Bank. (2000) *Nucleic Acids Research*, 28: 235-242
[3] Torrance James W., Bartlett Gail J., Porter Craig T., Thornton Janet M. Using a Library of Structural Templates to Recognize Catalytic Sites and Explore Their Evolution in Homologous Families. (2005) *J Mol Biol*. 347:565-81

ACKNOWLEDGEMENTS

Rochester Institute of Technology, Dowling College, NSF-DUE 0402408, NIGMS 1R15GM078077, 3R15GM078077-01S, 2R15GM078077-02, 3R15GM078077-02S1.