LabelHash: A Flexible and Extensible Method for Matching Structural Motifs

Mark Moll[1,]* and Lydia E. Kavraki[1,2,3,]*
[1] Department of Computer Science, Rice University, Houston, TX 77005, USA
[2] Department of Bioengineering, Rice University, Houston, TX 77005, USA
[3] Struct. & Comp. Biology and Molec. Biophysics, Baylor College of Medicine, Houston, TX 77005, USA
*To whom correspondence should be addressed: {mmoll,kavraki}@cs.rice.edu

Nature Precedings : doi:10.1038/npre.2008.2199.1 : Posted 15 Aug 2008

1. INTRODUCTION

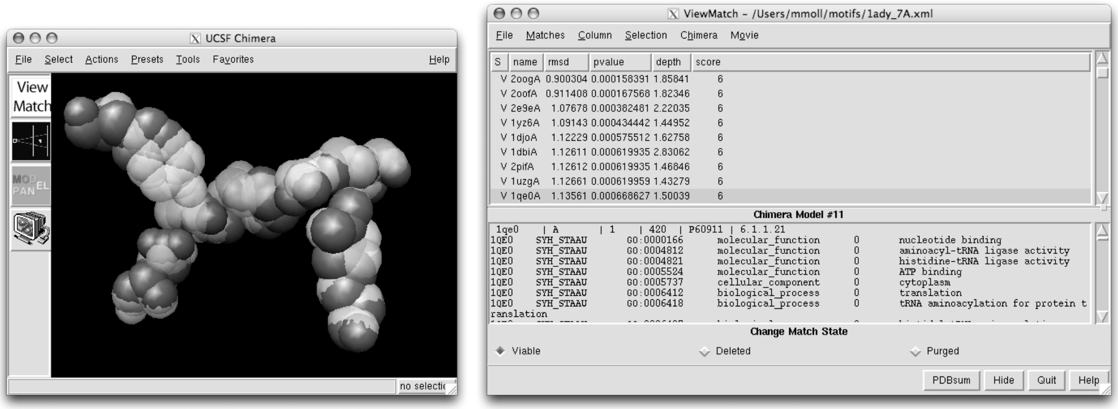
There is an increasing number of proteins with known structure but unknown function. Determining their function would have a significant impact on understanding diseases and designing new therapeutics. Computational methods can facilitate function determination by identifying proteins that have high structural and chemical similarity. Below, we will briefly describe LabelHash, a new method for partial structure comparison. In partial structure comparison, the goal is to find the best geometric and chemical similarity between a set of 3D points called a *motif* and a subset of a set of 3D points called the *target*. Both the motif and targets are represented as sets of labeled 3D points. A motif is ideally composed of the functionally most-relevant residues in a binding site. The labels denote the type of residue. Motif points can have multiple labels to denote that substitutions are allowed. Any subset of the target that has labels that are compatible with the motif's labels is called a *match*. The aim is to find statistically significant matches to a structural motif. Our method preprocesses a background database of targets such as a non-redundant subset of the Protein Data Bank in such a way that we can look up in constant time partial matches to a motif. Using a variant of the previously described match augmentation algorithm (1), we obtain complete matches to our motif. The nonparametric statistical model developed by (2,3) corrects for any bias introduced by our algorithm. This bias is introduced by excluding matches that do not satisfy certain geometric constraints for efficiency reasons.

In the implementation design of our method we focused on flexibility, extensibility, and ease of use. We wanted motifs to be as general as possible to allow for future extensions and to facilitate motif design through a variety of methods. The input should be easy to generate from "raw data" such as PDB files, and the output should be easy to post-process and visualize. Although the ideal of functional annotation is full automation, an exploratory process of iterative and near-interactive motif design and refinement will be extremely valuable. Our simple-to-use and extensible LabelHash algorithm can be a critical component of this process. It is possible to optionally include partial matches or multiple matches per target (although in these cases it is not yet possible to assign a statistical significance to matches). Currently, our motifs are based on C-alpha positions, labeled with residue labels, but an option to use pseudocenters instead is in development. Users can run the LabelHash algorithm on our server and visualize the results in Chimera, a molecular modeling program.

2. THE LABELHASH ALGORITHM AND IMPLEMENTATION

The LabelHash algorithm consists of two stages: a preprocessing stage and a matching stage. During the preprocessing stage we build up hash tables for *n*-tuples of residues that occur in a set of targets. These *n*-tuples are hashed based on residue labels. For each a given *n*-tuple of residues we can instantly find all occurrences in the targets. The *n*-tuples are subject to mild geometric constraints that guarantee spatial coherence and proximity to the molecular surface. The preprocessing stage needs to be executed only once for a given set of targets. The matching stage is a variant of the match augmentation algorithm described in (1). One important difference is that in LabelHash we no longer need an importance ranking of motif residues.

We have developed a plugin for Chimera (a molecular modeling program) called ViewMatch. Figure 1 shows the user interface. In the main window a selected match is shown superimposed with the motif. In the controller window, we see all matches in the top half with their LRMSD to the motif, *p*-value and other attributes. By specifying constraints on the match attributes, the user can restrict the matches that are


UCSF Chimera
File Select Actions Presets Tools Favorites Help
View Match
MOD PANEL
no selectio
ViewMatch – /Users/mmoll/motifs/1ady_7A.xml
File Matches Column Selection Chimera Movie
Chimera Model #11
Change Match State
PDBsum
Hide
Quit
Help

S	name	rmsd	pvalue	depth	score
V	2oogA	0.900304	0.000158391	1.85841	6
V	2oofA	0.911408	0.000167568	1.82346	6
V	2e9eA	1.07678	0.000382481	2.22035	6
V	1yz6A	1.09143	0.000434442	1.44952	6
V	1djoA	1.12229	0.000575512	1.62758	6
V	1dbiA	1.12611	0.000619935	2.83062	6
V	2pifA	1.12612	0.000619935	1.46846	6
V	1uzgA	1.12661	0.000619959	1.43279	6
V	1qe0A	1.13561	0.000668627	1.50039	6

```
 1qe0    | A       | 1    | 420  | P60911 | 6.1.1.21
1QE0    SYH_STAAU       GO:0000166      molecular_function      0       nucleotide binding
1QE0    SYH_STAAU       GO:0004812      molecular_function      0       aminoacyl-tRNA ligase activity
1QE0    SYH_STAAU       GO:0004821      molecular_function      0       histidine-tRNA ligase activity
1QE0    SYH_STAAU       GO:0005524      molecular_function      0       ATP binding
1QE0    SYH_STAAU       GO:0005737      cellular_component      0       cytoplasm
1QE0    SYH_STAAU       GO:0006412      biological_process      0       translation
1QE0    SYH_STAAU       GO:0006418      biological_process      0       tRNA aminoacylation for protein t
ranslation
```

◆ Viable ◇ Deleted ◇ Purged

Figure 1. **The ViewMatch plugin for Chimera allows for easy visualization and analysis of matches.**

shown. The bottom half of the window shows additional information for the selected match, such as EC classification and GO terms. By clicking the PDBsum button, the PDBsum web pages (4) are shown for the selected matches. This gives the user an enormous amount of information about a match. We have also created a command line tool that performs cluster analysis on a set of matches. This often reveals much additional information about why certain motifs may be difficult to match.

3. RESULTS

The LabelHash has been tested on 20 motifs. Our set of targets consisted of a non-redundant version of the PDB, separated into individual chains. This resulted in about 18,000 targets. At a false positive rate of about 0.04% we achieved on average a true positive rate of 84%. Since the number of targets is so large, a small false positive rate could still mean the absolute number of false positives is much larger than the number of true positives, but in our experiments this was not the case. In fact, the number of false positives was usually smaller than the number of true positives.

4. ACKNOWLEDGEMENTS

The project upon which this publication is based was performed pursuant to Baylor College of Medicine Grant No. DBI-054795 from the National Science Foundation. Equipment was funded by NSF CNS 0454333 and NSF CNS-0421109 in partnership with Rice University, AMD and Cray. The authors are indebted to V. Fofanov for many useful discussions on the use of statistical analysis and for his comments on LabelHash. They are also deeply grateful for the help of B. Chen and D. Bryant with match augmentation. O. Lichtarge and D. Kristensen participated in precursors of this work.

5. REFERENCES

1. Chen, B.Y., Fofanov, V.Y., Bryant, D.H., Dodson, B.D., Kristensen, D.M., Lisewski, A.M., Kimmel, M., Lichtarge, O. and Kavraki, L.E. The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs. *J. Comp. Bio.*, 14(6):791–816, 2007.

2. Fofanov, V.Y., Chen, B.Y., Bryant, D.H., Moll, M., Lichtarge, O., Kavraki, L.E. and Kimmel, M. Correcting systematic bias caused by algorithmic thresholds in statistical models of protein sub-structural similarity. *BMC Biology Direct*, 2008. Submitted.

3. Fofanov, V.Y. Statistical Models in Protein Structural Alignments. PhD thesis, Department of Statistics, Rice University, Houston, TX, 2008.

4. Laskowski, R.A. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Research*, 29(1): 221–222, 2001.