LabelHash: A Flexible and Extensible Method for Matching Structural Motifs

Mark Moll and Lydia E. Kavraki
Computer Science Department, Rice University, Houston, TX 77005, USA

Motivation

We want to represent protein function with a structural motif. Given a motif, we would like to quickly identify all proteins that are functionally related in a statistically significant way. It is difficult to design motifs that have high sensitivity as well as high specificity.

Problem statement:

What functionally relevant information about a large set of proteins (such as the whole PDB) can be stored in tables *in a scalable way* such that we quickly find matches to a point-based motif designed with any method?

LabelHash Algorithm

The LabelHash algorithm consists of two phases:

- **Preprocessing phase.** Hash tables are built for a large collection of proteins. The tables contain n-tuples of residues that are close together and close to the surface, indexed by their residue labels.
- **Matching phase.** For a given motif we can instantly look up partial matches of size n. Using a constrained depth-first search partial matches are augmented to complete matches.

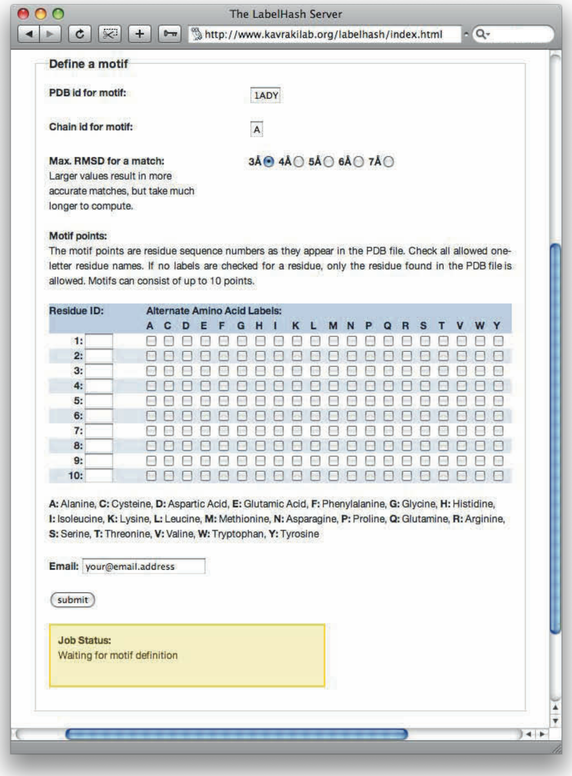
Statistical significance of matches is determined with a nonparametric model.

Initial Results

The algorithm has been tested with motifs for 20 enzyme (EC) classes. We created LabelHash tables for the non-redundant PDB at a 95% ID threshold. On average, we obtained a false positive rate of about 0.04% and a true positive rate of 84%. The runtime varies from minutes to hours, depending on the size of the motif and the number of alternate labels for each motif point.

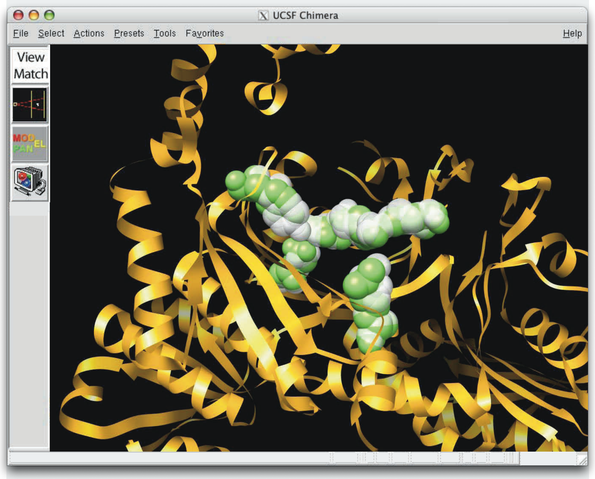
The LabelHash program is accessible through a web interface and available for download at:

http://kavrakilab.org/labelhash/

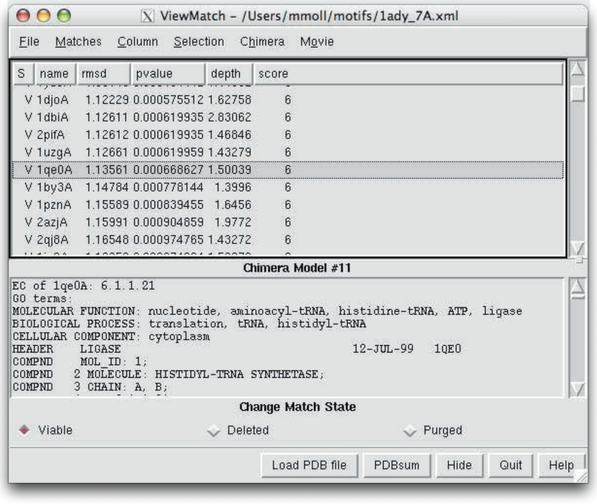


Match Visualization

We have developed a plugin for Chimera that can read in a file of matches and visualize the results. Below, a motif (in white) is shown superimposed on a match (in green). The rest of the matching protein is shown in ribbon representation.



In a controller window we can scroll through a list of matches. The bottom half of the window shows additional information about the selected match.



Conclusion

We have developed a practical new algorithm for partial structure comparison. It is a highly sensitive and specific method for matching structural point-based motifs (designed with any method). Typically, the number of false positives is much smaller than the number of true positives. The results are easily visualized and analyzed in Chimera. In addition, the LabelHash output XML files with matches are very amenable to post-processing. For instance, matches can easily be clustered or filtered out based on additional constraints.

References

M. Moll and L.E. Kavraki. Matching Of Structural Motifs Using Hashing On Residue Labels And Geometric Filtering For Protein Function Prediction, *Conf. Computational Systems Bioinformatics (CSB)*, 2008.

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

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.