Accepted for publication in Bionformatics.

BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–2

The LabelHash Server and Tools for Substructure-Based Functional Annotation

Mark Moll [1]*, Drew H. Bryant [1] and Lydia E. Kavraki [1,2,3]

[1]Department of Computer Science, Rice University, Houston, TX 77005, [2]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX 77030, [3]Department of Bioengineering, Rice University, Houston, TX 77005 USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: The LabelHash server and tools are designed for large-scale substructure comparison. The main use is to predict the function of unknown proteins. Given a set of (putative) functional residues, LabelHash finds all occurrences of matching substructures in the entire Protein Data Bank, along with a statistical significance estimate and known functional annotations for each match. The results can be downloaded for further analysis in any molecular viewer. For Chimera there is a plugin to facilitate this process.

Availability: The website is free and open to all users with no login requirements at http://labelhash.kavrakilab.org.

Contact: mmoll@rice.edu

1 INTRODUCTION

The function of many proteins is still poorly understood. For proteins with low sequence identity to any other proteins, structural comparison can be used to identify distant homologs. Both the number and diversity of structures is rapidly increasing. This helps make the results of a structural approach to functional annotation more informative, but at the same time also computationally more challenging. The function of a protein can often be characterized by a structural *motif:* a representative substructure formed by functional residues. The function of an unknown protein can therefore be inferred from high similarity of a substructure in this protein to a motif with known function. A variety of methods have been proposed that compute matches to a motif in a collection of structures, some of which are also accessible through a web server or as downloadable program (Laskowski *et al.*, 2005; Kleywegt, 1999; Stark and Russell, 2003; Kinjo and Nakamura, 2007; Konc and Janezic, 2010; Ren *et al.*, 2010). They differ significantly in one or more of the following aspects: representation of a motif, matching algorithm, statistical model, and the set of targets a motif can be matched against. The underlying algorithm for the LabelHash web server makes it possible to match motifs against the entire Protein Databank with minimal restrictions and still obtain results within a matter of minutes.

2 METHODS

The LabelHash algorithm (Moll *et al.*, 2010) consists of two stages: a preprocessing stage and a matching stage. During the preprocessing stage the algorithm builds up lookup tables for n-tuples of residues that occur in a set of target structures. These n-tuples are indexed by their residue labels. For a given n-tuple of residues we can instantly find all occurrences in all 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 target structures.

During the matching stage the algorithm first looks for partial matches of size n and incrementally expands these partial matches to complete matches in a depth-first fashion. The motif information required by LabelHash is very simple. Motifs consist of the C_α positions of the functionally important (possibly non-sequential) residues, labeled with the admissible residue types. There is no need for importance ranking of motif residues. To determine the statistical significance of a match we use a non-parametric model based on the LRMSD's of all matches that were found (Fofanov *et al.*, 2008). This model corrects for systematic algorithmic bias. We have shown that LabelHash can achieve extremely high specificity with high sensitivity on a benchmark set of motifs for twenty different Enzyme Commission classes (Moll and Kavraki, 2008). On average, we obtained a sensitivity of 86% and a specificity of 99.94% at a p-value of 0.001 for these motifs.

By default, the LabelHash algorithm returns only the best complete match for each target, but it can optionally also compute partial matches (of a given minimum size) and multiple matches per protein. With these features the user can filter and postprocess the matches in variety of ways, e.g., by employing a different selection mechanism for the "best" match per target.

3 THE LABELHASH WEB SERVER AND TOOLS

The LabelHash server provides an easy-to-use front-end to our implementation of the LabelHash algorithm. The server uses a LabelHash table that contains all the indexing information for the full Protein Data Bank (PDB). Each PDB entry often has more than one chain. Each chain is inserted as a separate target in the LabelHash table. This results in more than 180,000 target structures. The user can specify a motif by filling out a web form that asks for a PDB ID, a chain ID, and a number of residue sequence numbers. For each residue the user can specify allowed substitutions. It is also possible to upload an arbitrary PDB file and use that to define a motif. The motif can be matched against either the PDB, the non-redundant PDB (at different sequence identity levels), or a single target. Upon submission of the motif, a job will be submitted to a scheduler, and the user is presented with a URL of the page where the match results will appear, typically, after a few minutes. If a user filled out an email address on the main page, an email will be sent when the results are ready.

*To whom correspondence should be addressed.

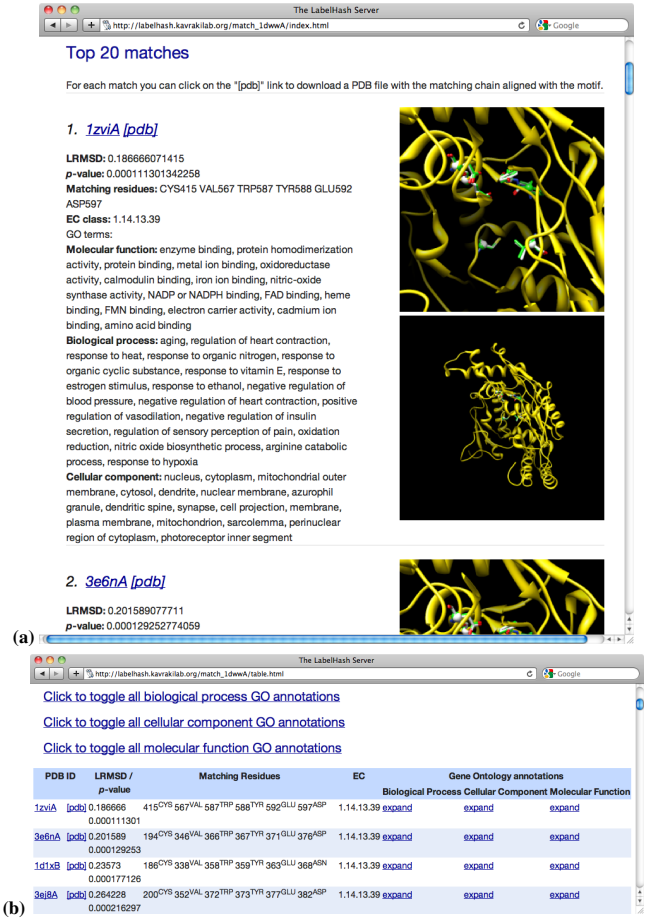

Fig. 1. **(a)** The main results page showing each matching substructure (in green) superimposed with a motif (in white). The rest of the matching structure is shown in ribbon representation. For each match we also include known annotations. **(b)** The table view of results. The first two columns contain links to the PDBsum entry and the aligned structure in PDB format. The Gene Ontology terms in the last 3 columns can be expanded individually or all simultaneously per category.

The match results are made available in different formats so as to accommodate several different workflows. The main results page shows the top twenty matches rendered in alignment with the motif in individual images. For each match the corresponding Enzyme Commission classification and Gene Ontology terms are shown, when they are available (see figure 1a). On a second results page, a more compact table view of the top 100 matches with known annotations is shown (see figure 1b). From both pages the user can download individual PDB files of the aligned structures or an archive with PDB files of the top 100 aligned structures. This is intended for users who want to further analyze the results in any molecular viewer. Finally, the user can also download an XML file with all matches. This file can be viewed as plain text, but one can also open this file in Chimera, a molecular modeling program, using a plugin we developed called ViewMatch. This plugin allows the user to quickly scroll through the matches, see their annotations, and filter them by different attributes.

Although the LabelHash web server is very easy to use, it offers very little flexibility in how the matching is performed. For more advanced users we have created a set of command line tools that allow for more extensive experimentation. First, the user can create LabelHash tables for arbitrary sets of PDB files. Second, all the parameters to the algorithm can be set by the user. In particular, one can choose to enable partial matches or multiple matches per protein. The tools include a Python module. This, combined with a simple XML input/output format, facilitates pre- and postprocessing.

The LabelHash algorithm enables applications that benefit from structural analysis and go beyond what was described in the introduction. For example, we were able to identify subtle patterns of substructural variation in large protein (super)families that correlate with, e.g., phylogenetic distance, conformational rearrangement upon binding, or homology (Bryant *et al.*, 2010).

ACKNOWLEDGEMENTS

The authors thank Dr. Paul Finn for many suggestions to improve the usability of the site. Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081) (Pettersen *et al.*, 2004).

Funding: National Science Foundation (DBI-0960612, GRFP DGE-0237081 to DHB), the John & Ann Doerr Fund for Computational Biomedicine at Rice University, the Texas Higher Education Coordinating Board NHARP 01907, and the Sloan Foundation.

Conflict of Interest: none declared.

REFERENCES

Bryant, D. H., Moll, M., Chen, B. Y., Fofanov, V. Y., and Kavraki, L. E. (2010). Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction. *BMC Bioinformatics*, **11**(242).

Fofanov, V. Y., Chen, B. Y., Bryant, D. H., Moll, M., Lichtarge, O., Kavraki, L. E., and Kimmel, M. (2008). A statistical model to correct systematic bias introduced by algorithmic thresholds in protein structural comparison algorithms. In *IEEE Intl. Conf. on Bioinformatics and Biomedicine Workshops (BIBMW)*, pages 1–8.

Kinjo, A. R. and Nakamura, H. (2007). Similarity search for local protein structures at atomic resolution by exploiting a database management system. *Biophysics*, **3**, 75–84.

Kleywegt, G. J. (1999). Recognition of spatial motifs in protein structures. *J Mol Biol*, **285**(4), 1887–1897.

Konc, J. and Janezic, D. (2010). ProBiS: a web server for detection of structurally similar protein binding sites. *Nucleic Acids Res*, **38**(Web Server issue), W436–40.

Laskowski, R. A., Watson, J. D., and Thornton, J. M. (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Research*, **33**, W89–W93.

Moll, M. and Kavraki, L. E. (2008). Matching of structural motifs using hashing on residue labels and geometric filtering for protein function prediction. In *The Seventh Annual International Conference on Computational Systems Bioinformatics (CSB2008)*, pages 157–168.

Moll, M., Bryant, D. H., and Kavraki, L. E. (2010). The LabelHash algorithm for substructure matching. *BMC Bioinformatics*, **11**(555).

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, **25**(13), 1605–1612.

Ren, J., Xie, L., Li, W. W., and Bourne, P. E. (2010). SMAP-WS: a parallel web service for structural proteome-wide ligand-binding site comparison. *Nucleic Acids Res*, **38**(Web Server issue), W441–4.

Stark, A. and Russell, R. B. (2003). Annotation in three dimensions. PINTS: Patterns in non-homologous tertiary structures. *Nucleic Acids Research*, **31**(13), 3341–3344.