Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W267–W270***
doi:10.1093/nar/gki465

MINER: software for phylogenetic motif identification

David La and Dennis R. Livesay[1,2,*]

Department of Biological Sciences, [1]Department of Chemistry and [2]Center for Macromolecular Modeling and Materials Design, California State Polytechnic University, Pomona, CA 91767, USA

Received February 11, 2005; Revised March 25, 2005; Accepted April 13, 2005

ABSTRACT

MINER is web-based software for phylogenetic motif (PM) identification. PMs are sequence regions (fragments) that conserve the overall familial phylogeny. PMs have been shown to correspond to a wide variety of catalytic regions, substrate-binding sites and protein interfaces, making them ideal functional site predictions. The MINER output provides an intuitive interface for interactive PM sequence analysis and structural visualization. The web implementation of MINER is freely available at http://www.pmap.csupomona.edu/MINER/. Source code is available to the academic community on request.

INTRODUCTION

Because of the exponential growth of available sequence data, the development of accurate computational strategies for functional site identification has become one of the most important post-genomic challenges (1). Many methods attempt to predict functional sites from sequence alone. Highly conserved positions within sequence alignments are strong candidates for functional sites. Although attractive owing to their relative simplicity, conservation-based approaches frequently result in too many false positives to be satisfactory. In addition, sequence regions with significant variability can also be critical to function, especially when their composition may define subfamily specificity. Frequently, these regions correspond to residues that are critical in molecular recognition and binding specificity.

In this report, we present MINER, software for phylogenetic motif (PM) identification. PMs are sequence alignment regions that conserve the overall phylogeny of the complete family. Through comparison with structural and biochemical data, we have shown that PMs represent good functional site predictions in a wide variety of protein systems (2). Our results indicate that, despite little overall proximity in sequence, PMs are structurally clustered around key functionality across a wide variety of structural examples. PMs correspond to a variety of structural features, including solvent exposed loops, active site clefts and buried regions surrounding prosthetic groups (Figure 1). Our results also indicate that PMs are generally conserved in sequence, indicating that PMs tend to be motifs in the traditional sense. Consequently, PM results bridge evolutionary (3–5) and traditional motif (6,7) approaches. In spite of the small alignment window size, PM tree significance has been demonstrated using bootstrapping.

IMPLEMENTATION

MINER takes as an input any multiple sequence alignment (MSA). If sequences are unaligned, MINER will align them for the user by using ClustalW (8). A sliding sequence window algorithm is used to quantitatively evaluate the phylogenetic similarity between each sequence region and the whole sequence. Distance-based trees are calculated both for the whole alignment and each window. Phylogenetic similarity is based on tree topology, which is calculated using the partition metric algorithm (9). The partition metric counts the number of topological differences between the two trees. Partition metric values are recast as *Z*-scores. Overlapping sequence windows scoring past some preset phylogenetic similarity *Z*-score (PSZ) threshold are identified as PMs. Empirically, we have determined that a window width between 5 and 10, and a PSZ threshold between −1.5 and −2.2 (lower scores indicate greater similarity) represent ideal default parameters for functional site prediction. MINER allows the user to easily change these parameters as desired. By default, alignment positions with >50% gaps are eliminated (masked). However, the user retains the option to handle gaps as described previously (2).

MINER can now automatically determine the PSZ threshold without human subjectivity. The automated algorithm relies on significant raw data preprocessing to improve signal detection. Subsequently, Partition Around Medoids Clustering of the similarity scores assesses those sequence fragments whose annotation remains in doubt. The accuracy of the approach has been confirmed through comparisons with our manual results (2,10). A preprint more thoroughly describing the automated algorithm is available at the MINER

*To whom correspondence should be addressed. Tel: +1 909 869 4409; Fax: +1 909 869 4344; Email: drlivesay@csupomona.edu

website. With the automated algorithm in hand, we have precomputed all PMs for the most recent version of the COG database (11). These results are also available at the MINER website.

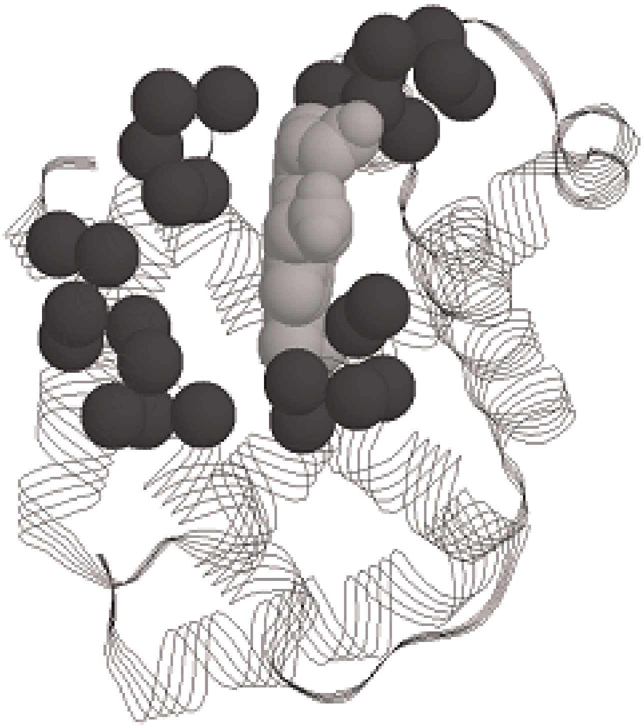


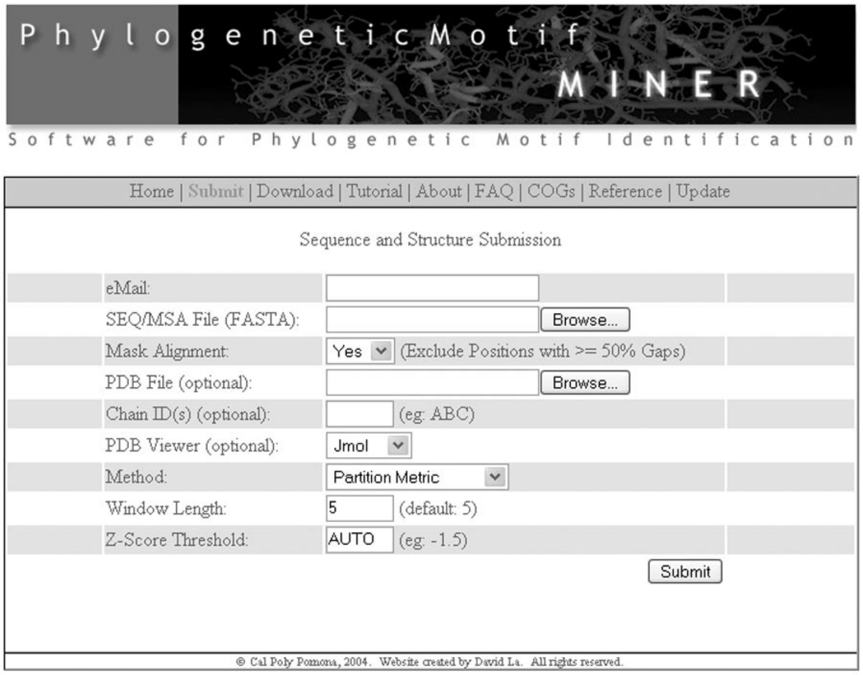
Figure 1. The four best-scoring (PSZ threshold = −1.5) PMs identified from the myoglobin protein family are mapped onto structure (pdbid: 1MBA). The α-carbons of the PMs are shown as black spheres; the heme is shown in gray.

MINER is available as standalone (command-line based) software and through the Web via a user-friendly interface. The standalone version is written in PERL and can be easily modified. A CGI facade is implemented over the standalone version for ease of use. After the web-based calculation is complete, MINER sends an email with a hyperlink directing the user to their results. The user has 1 week to access and download their outputs. MINER is part of the larger Protein Motif Analysis Portal at California State Polytechnic University, Pomona (12).

INPUT

MINER requires a minimum of five sequences in the FASTA format. However, we recommend using 25 or more sequences to ensure sufficient evolutionary diversity. With the exception of gaps, all non-alphabetic characters found in the input will be purged. Optionally, a Protein Data Bank (PDB) structure may be submitted to better highlight PM regions. MINER will automatically add the PDB sequence to a dataset of unaligned sequences if it does not exist. However, user-provided alignments must already include the PDB sequence as part of the alignment.

There are several default MINER options that can be customized before submission (Figure 2). Enabling the masking feature will purge alignment positions with >50% gaps. Although masking is optional, we find that eliminating these positions significantly increases the quality of functional site predictions, especially in more divergent families. MINER also provides three methods for identifying motifs. By default, MINER identifies functional sites as described above. Alternatively, MINER also provides the option to identify traditional motifs using the False Positive Expectation (FPE) of a regular

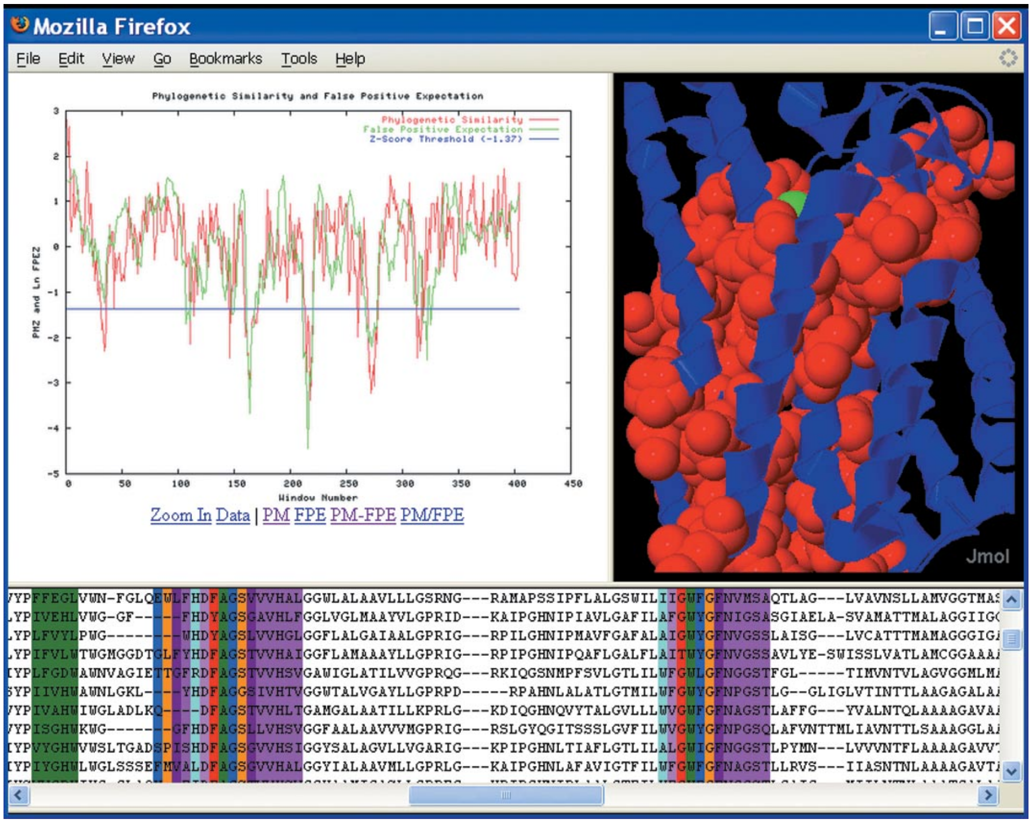

PhylogeneticMotif
MINER
Software for Phylogenetic Motif Identification
Home | Submit | Download | Tutorial | About | FAQ | COGs | Reference | Update
Browse...
Browse...
Submit

Sequence and Structure Submission

eMail:	
SEQ/MSA File (FASTA):	
Mask Alignment:	Yes (Exclude Positions with >= 50% Gaps)
PDB File (optional):	
Chain ID(s) (optional):	(eg: ABC)
PDB Viewer (optional):	Jmol
Method:	Partition Metric
Window Length:	5 (default: 5)
Z-Score Threshold:	AUTO (eg: -1.5)

© Cal Poly Pomona, 2004. Website created by David La. All rights reserved.

Figure 2. Screenshot of the MINER input page.


Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Phylogenetic Similarity and False Positive Expectation
Phylogenetic Similarity
False Positive Expectation
Z-Score Threshold (-1.37)
PMZ and Ln FPEZ
3
2
1
0
-1
-2
-3
-4
-5
0
50
100
150
200
250
300
350
400
450
Window Number
Zoom In Data | PM FPE PM-FPE PM/FPE
Jmol

```
VYPFFEGLVWN-FGLQEWLFHDFAGSVVVHALGGWLALAAVLLLGSRNG---RAMAPSSIPFLALGSWILIIGWFGFNVMSAQTLAG---LVAVNSLLAMVGGTMAS
LYPIVEHLVWG-GF-----FHDYAGSGAVHLFGGLVGLMAAYVLGPRID---KAIPGHNIPIAVLGAFILAFGWYGFNIGSASGIAELA-SVAMATTMALAGGIIGC
LYPLFVYLPWG--------WHDYAGSLVVHGLGGFLALGAIAALGPRIG---RPILGHNIPMAVFGAFALAIGWYGFNVGSSLAISG---LVCATTTMAMAGGGIGA
LYPIFVLWTWGMGGDTGLFYHDFAGSTVVHAIGGFLAMAAAYLLGPRIG---RPIPGHNIPQAFLGALFLAITWYGFNVGSSAVLYE-SWISSLVATLAMCGGAAAA
IYPLFGDWAWNVAGIETTGFRDFAGSTVVHSVGAWIGLATILVVGPRQG---RKIQGSNMPFSVLGTLILWFGWLGFNGGSTFGL-----TIMVNTVLAGVGGMLMA
5YPIIVHWAWNLGKL----YHDFAGGSIVHTVGGWTALVGAYLLGPRPD-----RPAHNLALATLGTMILWFGWYGFNPGSTLG--GLIGLVTINTTLAAGAGALAA
VYPIVAHWIWGLADLKQ----DFAGSTVVHLTGAMGALAATILLKPRLG---KDIQGHNQVYTALGVLLLWVGWFGFNAGSTLAFFG---YVALNTQLAAAAGAVAA
VYPISGHWKWG--------GFHDFAGSLLVHSVGGFAALAAVVVMGPRIG---RSLGYQGITSSSLGVFILWVGWYGFNPGSQLAFVNTTMLIAVNTTLSAAAGGLAA
IYPVYGHWVWSLTGADSPISHDFAGSGVVHSIGGYSALAGVLLVGARIG---KPIPGHNLTIAFLGTLILALGWIGFNGGSTLPYMN---LVVVNTFLAAAAGAVVT
IYPIYGHWLWGLSSSEFMVALDFAGSGVVHALGGYIALAAVMLLGPRLG---KAIPGHNLAFAVIGTFILWFGWFGFNAGSTLLRVS---IIASNTNLAAAAGAVTA
```

Figure 3. Screenshot of the MINER output applied to the ammonia channel. The sequence window (which can toggle between aligned and ungapped) is hyperlinked to the structure viewer and WebLogos. The upper-left window can toggle between both PM (red) and FPE (green) results. In all cases, the raw data are easily accessible for export.

expression or profile. Both approaches are described in detail within the tutorial at the MINER website. When used in conjunction with the PM results, these alternative approaches often provide synergistic information. In addition, the width of the sliding window can also be modified. By default, the width is set to five alignment positions, which we find it to be ideal for identifying functional sites (2). However, large windows are more appropriate when exploiting 'motif-ness' (e.g. using PMs to de-ORFan uncharacterized sequences). The *Z*-score threshold is automatically determined by default, but can be manually set any value ≤ -1. Finally, either Jmol (default) or Chime viewers can be used for interactive structure visualization.

OUTPUT

The MINER output is a framed HTML file (Figure 3) that provides (i) phylogenetic similarity versus window number plots, (ii) an annotated structure and (iii) an annotated MSA. PM regions in the PDB structure are annotated by writing the PSZ to the temperature factor column. Furthermore, interactive structural visualization of the identified PMs is achieved with the option of using either Jmol or Chime. Each PM within the alignment is hyperlinked such that clicking it will highlight the corresponding structural region. PM sequence logos, generated by WebLogo (13), are also hyperlinked from the MSA. In all cases, the raw data are available for easy export to auxiliary programs. With the masking feature enabled, regions of the MSA colored light gray represent alignment positions that have been purged before PM identification. At the MINER website, a full tutorial and frequently asked questions page is provided. The tutorial guides one through the output of triosephosphate isomerase results, which is the center of discussion in our previous reports (2,10).

CONCLUSIONS

MINER is a convenient web-based program for PM discovery. MINER utilizes a sliding sequence window algorithm to systematically evaluate all regions of an MSA input. Phylogenetic similarity is determined by comparing tree topology, which is calculated using the partition metric algorithm consequently resulting in a PSZ value. The sensitivity in the PM identification is constrained using a PSZ threshold, which is automatically determined by default. The resulting MINER output uses Jmol or Chime PDB viewers allowing protein structure and corresponding PM regions to be interactively visualized. The standalone version of MINER is freely available for academic download on request.

ACKNOWLEDGEMENTS

This work is supported by a grant from the American Chemical Society-Petroleum Research Fund (36848-GB4). Funding to pay the Open Access publication charges for this article was provided by the College of Science at California State Polytechnic University.

Conflict of interest statement. None declared.

REFERENCES

1. Jones,S. and Thornton,J.M. (2004) Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.*, **8**, 3–7.
2. La,D., Sutch,B. and Livesay,D.R. (2005) Predicting protein functional sites with phylogenetic motifs. *Proteins*, **58**, 309–320.
3. Armon,A., Graur,D. and Ben Tal,N. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.*, **307**, 447–463.
4. del Sol,M.A., Pazos,F. and Valencia,A. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.
5. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
6. Liu,A.H., Zhang,X., Stolovitzky,G.A., Califano,A. and Firestein,S.J. (2003) Motif-based construction of a functional map for mammalian olfactory receptors. *Genomics*, **81**, 443–456.
7. Puntervoll,P., Linding,R., Gemund,C., Chabanis-Davidson,S., Mattingsdal,M., Cameron,S., Martin,D.M., Ausiello,G., Brannetti,B., Costantini,A. *et al.* (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625–3630.
8. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
9. Penny,D. and Hendy,M. (1985) The use of tree comparison metrics. *Syst. Zool.*, **34**, 75–82.
10. Livesay,D.R. and La,D. (2005) The evolutionary origins and catalytic importance of conserved electrostatic networks within TIM-barrel proteins. *Protein Sci.*, **14**, 1158–1170.
11. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
12. La,D., Silver,M., Edgar,R.C. and Livesay,D.R. (2003) Using motif-based methods in multiple genome analyses: a case study comparing orthologous mesophilic and thermophilic proteins. *Biochemistry*, **42**, 8988–8998.
13. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.