Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

Leitmotif: protein motif scanning 2.0

Siniša Biđin[1§], Ivan Vujaklija[1§*], Tina Paradžik[2], Ana Bielen[3] and Dušica Vujaklija[2*]

[1]Faculty of Electrical Engineering and Computing[1], University of Zagreb, Unska 3, Zagreb, [2]Department of Physical Chemistry, Institute Ruđer Bošković, Bijenička 54, Zagreb, [3]Department of Biochemical Engineering, Faculty of Food Technology and Biotechnology, Pierottijeva 6, University of Zagreb, Zagreb, Croatia

*To whom correspondence should be addressed; §Authors contributed equally to this work

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Motif-HMM (mHMM) scanning has been shown to possess unique advantages over standardly used sequence-profile search methods (e.g. HMMER, PSI-BLAST) since it is particularly well suited to discriminate proteins with variations inside conserved motifs (e.g. family subtypes) or motifs lacking essential residues (false positives, e.g. pseudoenzymes).
Results: In order to make mHMM widely accessible to a broader scientific community we developed Leitmotif, a mHMM web application with many parametrization options easily accessible through intuitive interface. Substantial improvement of performance (ROC scores) was obtained by using two novel parameters. To the best of our knowledge Leitmotif is the only available mHMM application.
Availability: Leitmotif is freely available at https://leitmotif.irb.hr
Contact: sinisa@heuristika.hr or ivan.vujaklija@fer.hr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Rapid advances in sequencing technologies have resulted in remarkable accumulation of proteomics data. Lack of high-throughput experimental assays for protein annotations necessitates the application of computational methods. However, a major drawback of the computational approach is an ever increasing number of erroneous annotations (Furnham *et al*., 2009). These errors, once introduced, have a negative impact on subsequent analyses (Schäffer *et al*., 2001).

By using a motif scanning approach we found (Vujaklija *et al*., 2016) a surprisingly high number of false positives (20%) within the GDSL protein family in the manually curated Pfam database which uses the state of the art profile-HMM search tool HMMER (hmmer.org; Seo *et al.,* 2018). Our results showed that current profile sequence similarity search methods like HMMER and PSI-BLAST are inadequate to distinguish proteins that share profile similarity but lack catalytically active sites (Vujaklija *et al.*, 2016). Catalytically deficient variants, i.e. pseudoenzymes, have been found in all major enzyme families and it is becoming clear that they have important roles in various cellular processes in all organisms (Eyers *et al*., 2016). In addition, functional variations between family subtypes often depend on Specificity Determining Residues (SDR) (Goldstein *et al.,* 2009). Therefore, the ability to distinguish different residues at specific positions is very important (Kress *et al*., 2018). Considering this, motif scanning can be successfully used for computational detection of enzyme families, subfamilies and pseudoenzymes. By far the most widely used approaches for motif scanning are regular expressions search and Position Specific Scoring Matrix (PSMM). Although very rarely used motif-HMM (mHMM) was recently shown to be very effective for motif scanning (Vujaklija *et al*., 2016). Unlike the well-known profile-HMM (e.g. HMMER), mHMM models only protein motif regions (for details see Supplementary 1.1). The advantage of mHMM over the two aforementioned standard methods is its ability to (probabilistically) model the ordering of multiple motifs as well as inter motif distances. This additional information can provide important clues to substantially improve protein annotations as shown in this study.

2 Methods

Leitmotif enables: (**i**) to define up to 5 motifs; (**ii**) to set expected inter-motif distances and distance to N/C-terminus. This feature was added since protein motifs are often located within a certain distance from each other (Galperin and Frishman, 1999) which can be useful for correct annotation. Users can also (**iii**) penalize deviations from the expected

distances by choosing different Distance Penalty (**DP**) strengths ("None" to "Strong"). This feature is implemented by using a simple heuristic formula (Supplementary 1.2). Finally, users can (**iv**) set any residue at a particular position as Immutable Residues (**IR**) (Fig.1). This forces the

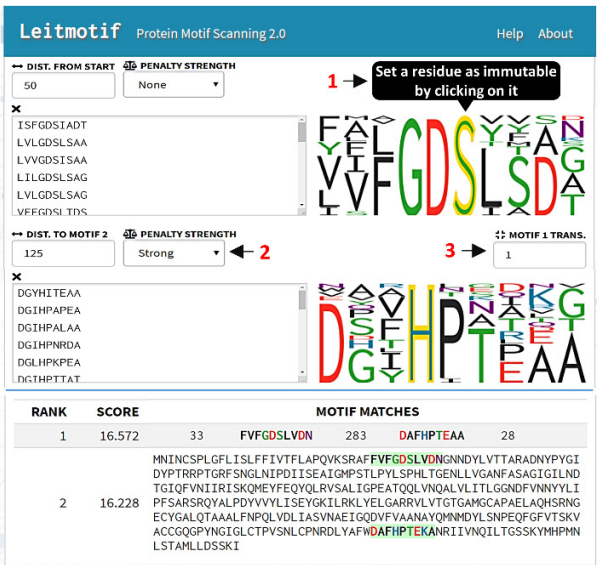

Fig. 1. The **Leitmotif user interface.** The upper panel shows setting: (**1**) immutable residues; (**2**) inter-motif distances and distance penalties; and (**3**) transition probabilities. The lower panel shows an output example.

algorithm to return a motif with the specified residue at selected position(s), thus enabling identification of protein sequences with particular motif pattern(s). This feature can be applied to distinguish various protein families, subfamilies or pseudoenzymes. Additionally, Leitmotif offers 4 different ways of computing emission probabilities and sequence weighting options (Supplementary 1.3).

3 Results

To illustrate benefits of two novel parameters **IR** and **DP** we used two test datasets. Firstly we analyzed a GDSL test dataset used previously (Supplementary 2.2; Supplement GDSL Test sequences). Two essential residues, Ser (1st motif) and His (3rd motif) recently proposed to form an essential catalytic dyad, were set as immutable, (Leščić Ašler *et al*., 2017 and Supplementary 2.2.1.) (Fig 1). Motif distances were set in line with reported data (Vujaklija *et al*., 2016; Upton and Buckley, 1995). GDSL lipases possess 3 most conserved motifs (Bielen *et al*., 2009; Chepyshko *et al*., 2012). We tested **IR/DP** parameters using 3, 2 and 1 motif(s). Transition probabilities were set to 0.99 to allow indels in motif(s). All emission probability algorithms were used. Due to limited space, only ROC scores with Modified Ancestral (**MA**) algorithm are shown in Table 1. We include different ROC scores for comparison (see Supplementary 2.1). As shown, increasing the number of motifs expectedly improves ROC scores (Supplementary 2.2.2 & 2.2.3). Moreover, setting novel parameters (**IR**, **DP**) to selected values leads to substantial improvement in all ROC scores irrespective of the number of motifs (Table 1A). The reasons for selecting these values are explained in Supplementary 2.2.1. The Desaturases First subfamily was chosen as a second test dataset (Supplementary 2.3; Supplement Desaturases First Test sequences) since it has a previously described protein motif signature (Hashimoto *et al*., 2008), thus providing unbiased annotation criteria (Supplementary 2.3.2). Importantly, in this case **IR/DP** values were set exclusively based on the *seed dataset* (Supplementary 2.3.3). Altogether, results presented in Table 1 (A & B) and in the Supplementary (2.2.2, 2.2.3, 2.3.4 and 2.3.5) show that the addition of novel parameters substantially improves protein motif scanning in selected datasets.

Table 1 ROC scores for different IR and DP parameter values on datasets A) the GDSL family and B) the Desaturases First subfamily.

A-GDSL	**IR**	**DP**	**ROC**	**nROC 50**	**nROC 5**	**nROC 1**
3 motifs	[6S, ,4H]	W	0.9902	0.9651	0.7968	0.7324
	[, ,]	/	0.9784	0.9239	0.7202	0.5994
2 motifs	[6S,4H]	M	0.9879	0.9588	0.8458	0.7901
	[,]	/	0.9665	0.9009	0.4965	0.2564
1 motif	[6S]	S	0.8916	0.6661	0.1083	0.0080
	[ ]	/	0.7517	0.2483	0.0144	0.0048
B-Desaturase First subfamily						
3 motifs	[1H 2R 6H, 1H 2R 4H 5H, 1H 2N 4H 5H]	S*	0.9885	0.9088	0.6419	0.5713
	[, ,]	/	0.8741	0.4959	0.0536	0.0019

IR-Immutable Residues: Numbers stand for IR positions in motifs and letters stand for amino acids, different motifs are separated by commas; **DP**-Distance Penalty: "/"-None; "W"-Weak, "M"-Medium and "S"-Strong. While GDSL (Table 1A) uses the same DP for all distances, Desaturases First (Table 1B) uses Strong between the 1st & 2nd motif only (S*).

Funding

This research was funded in part by the Croatian Government and the EU through the European Regional Development Fund - the Competitiveness and Cohesion Operational Programme (KK.01.1.1.01), The Scientific Centre of Excellence for Marine Bioprospecting–BioProCro.

Conflict of Interest: none declared.

References

Bielen,A. *et al*. (2009) The SGNH-hydrolase of *Streptomyces coelicolor* has (aryl)esterase and a true lipase activity. *Biochimie*, **91,** 390–400.

Chepyshko,H. *et al.* (2012) Multifunctionality and diversity of GDSL esterase/lipase gene family in rice (*Oryza sativa* L. *japonica*) genome: new insights from bioinformatics analysis. *BMC Genomics*, **13,** 1471–2164.

Eyers,P.A. and Murphy,J.M. (2016) The evolving world of pseudoenzymes: proteins, prejudice and zombies. *BMC Biol.,* **14,** 98.

Furnham,N. *et al.* (2009) Missing in action: enzyme functional annotation in biological databases. *Nat. Chem. Biol.,* **5,** 521–525.

Galperin,M.Y. and Frishman,D. (1999) Towards Automated Prediction of Protein Function from Microbial Genomic Sequences. *Method. Microbiol.,* **28,** 245–263.

Goldstein,P. *et al.* (2009) Clustering of protein domains for functional and evolutionary studies. *BMC Bioinformatics,* **10,** 335.

Kress,A. *et al.* (2018) PROBE: analysis and visualization of protein block-level evolution. *Bioinformatics,* **19,** 3390–3392.

Leščić Ašler,I. *et al.* (2017) Catalytic Dyad in the SGNH Hydrolase Superfamily: In-depth Insight into Structural Parameters Tuning the Catalytic Process of Extracellular Lipase from *Streptomyces rimosus*. *ACS Chem. Biol.,* **12,** 1928–1936

Schäffer,A.A. *et al.* (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29,** 2994–3005.

Seo,S. *et al.* (2018) DeepFam: deep learning based alignment-free method for protein family modeling and prediction. *Bioinformatics,* **34,** i254–i262.

Upton,C. and. Buckley,J.T. (1995) A new family of lipolytic enzymes? *Trends. Biochem. Sci*., 20 178 e179

Vujaklija,I. *et al.* (2016) An effective approach for annotation of protein families with low sequence similarity and conserved motifs: identifying GDSL hydrolases across the plant kingdom. *BMC Bioinformatics,* **18,** 17–91.