W470–W477 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* *Published online 29 April 2017*
doi: 10.1093/nar/gkx341

HH-MOTiF: *de novo* detection of short linear motifs in proteins by Hidden Markov Model comparisons

Roman Prytuliak[1], Michael Volkmer[1], Markus Meier[2] and Bianca H. Habermann[1,3,*]

[1]Computational Biology Group, Max Planck Institute of Biochemistry, Martinsried, Germany, [2]Research Group Quantitative Biology and Bioinformatics, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany and [3]Computational Biology Group, Developmental Biology Institute of Marseille (IBDM) UMR 7288, CNRS, Aix Marseille Université, Marseille 13288 Cedex 9, France

Received January 27, 2017; Revised April 10, 2017; Editorial Decision April 12, 2017; Accepted April 18, 2017

ABSTRACT

Short linear motifs (SLiMs) in proteins are self-sufficient functional sequences that specify interaction sites for other molecules and thus mediate a multitude of functions. Computational, as well as experimental biological research would significantly benefit, if SLiMs in proteins could be correctly predicted *de novo* with high sensitivity. However, *de novo* SLiM prediction is a difficult computational task. When considering recall and precision, the performances of published methods indicate remaining challenges in SLiM discovery. We have developed HH-MOTiF, a web-based method for SLiM discovery in sets of mainly unrelated proteins. HH-MOTiF makes use of evolutionary information by creating Hidden Markov Models (HMMs) for each input sequence and its closely related orthologs. HMMs are compared against each other to retrieve short stretches of homology that represent potential SLiMs. These are transformed to hierarchical structures, which we refer to as motif trees, for further processing and evaluation. Our approach allows us to identify degenerate SLiMs, while still maintaining a reasonably high precision. When considering a balanced measure for recall and precision, HH-MOTiF performs better on test data compared to other SLiM discovery methods. HH-MOTiF is freely available as a web-server at http://hh-motif.biochem.mpg.de.

INTRODUCTION

Short linear motifs (SLiMs) are small, context-independent, functional motifs of three to ∼20 amino acids within proteins that are sufficient to fulfill certain functions. Their best characterized activities include: binding to other (macro-)molecules such as nucleic acids, proteins, lipids or other small chemicals; serving as spots for protein modifications; encoding cleavage signals or being required for proper protein localization (1,2). Both, computational, as well as wet lab biological research would considerably profit, if we could reliably predict all relevant SLiMs in proteins *de novo*. Bench scientists typically want to know SLiMs in a small set of proteins for further experimental testing, addressing questions like protein localization, modification, or interaction with (macro-)molecules. In computational biology, especially in the research field of network biology, comprehensive knowledge of functional SLiMs would for instance allow us to better understand and represent dynamical processes in protein interaction networks by identifying mutually exclusive binding partners of hub proteins. However, *de novo* SLiM prediction is computationally difficult, due to their shortness and their typically very poor conservation (3). The fact that short, recurrent sequence motifs may play a role in the structural maintenance and stability of structurally unrelated proteins (4) is an additional difficulty. It necessitates the discrimination between short sequence stretches that are relevant for a particular function (such as binding to another molecule) and those that are needed to maintain the overall fold of a protein. As a consequence, one has to anticipate a high number of false positive predictions when searching for SLiMs *de novo*.

A SLiM is not perfectly conserved between proteins but rather represents the set of its evolutionary possible, still functional variants. Consequently, simplified models to characterize motifs exist. The regular expression (regex) is the simplest form of representing and working with sequence motifs. However, a regex represents only highly conserved positions well and is not able to capture positions with a low, but still significant conservation. Profile-based approaches such as weighted regexes (5) or Hidden Markov Models (HMMs, (6)) overcome these limitations and have more recently been used in *de novo* SLiM prediction (7–10).

Several methods have been published that offer *de novo* prediction of SLiMs using either regexes or profile-based methods. Regex-based tools include DILIMOT (11), SLiMFinder (12) or MotifHound (8). The popular MEME

*To whom correspondence should be addressed. Tel: +33 4 91 26 92 36; Email: bianca.habermann@univ-amu.fr

suite (13), which includes MEME and GLAM2 (14) for SLiM discovery, uses position weight matrices and Gibbs sampling, respectively. Several algorithms based on HMMs were also reported (NestedMICA (7), whmm (10) and dhmm (9)). However, the latter three all lack web-server access and are more difficult to use for bench scientists. More recently, de Brujin graphs were tested for SLiM discovery in proteins (15).

De novo SLiM search methods can also be classified as either non-discriminative, which only require a set of putative SLiM-containing proteins as input (these include SLiMFinder, MEME, DILIMOT and whmm); or discriminative, which in addition require a negative dataset of proteins that do not contain the putative SLiM. Therefore, additional biological knowledge is necessary for these methods in order to define a negative dataset for the sought-after SLiM. However, this knowledge does not always exist. MotifHound and dhmm are examples of discriminative *de novo* SLiM predictors. Several papers and reviews provide a comprehensive overview on SLiM discovery methods, as well as the inherent problems in finding novel SLiMs (8,16,17).

As SLiM discovery is computationally a difficult problem, none of the *de novo* SLiM search methods reported to date are able to discover SLiMs in proteins with reasonably good recall and precision. In fact, in *de novo* SLiM prediction, one has to typically trade off one for the other, reaching either high recall (e.g. GLAM2 with default settings) or high precision (e.g. SLiMFinder with default settings). It is therefore evident that finding novel SLiMs in proteins remains an important computational challenge.

We have developed HH-MOTiF (for HH-MOtif-TreeFinder), a web-server for finding novel SLiMs in sets of mainly unrelated proteins. HH-MOTiF makes use of evolutionary information by creating HMMs for each input sequence and its orthologs. We then combine HMM–HMM (HH–) comparisons using a customized version of HH-suite (18) with a hierarchical motif representation, which we refer to as motif trees. We evaluate identified motif trees prior to assembly at several levels including its surface accessibility and apply a novel algorithm for correcting for conserved domains or larger homologous regions in SLiM detection. HMMs are restricted to closely related orthologs, ensuring the presence of the relevant SLiM in the HMM. HH-MOTiF works non-discriminatively, thus the only input required is a set of—ideally unrelated—protein sequences that should share one functional feature characterized by a common, sought-after SLiM. The web-server version of HH-MOTiF was designed for datasets >50 proteins, coming for instance from wet-lab studies on protein interaction or localization.

The HH-MOTiF workflow

The workflow of HH-MOTiF is summarized in Figure 1 A.

The input of an HH-MOTiF search is a set of FASTA formatted protein sequences. For each sequence, close orthologs are first searched; then a multiple sequence alignment and the HMM of selected orthologs is computed. All HMMs are compared against each other with an adapted version of HH-suite. As high-scoring alignments reflect overall homology between input sequences, only short alignment hits are further evaluated. Overlapping alignment hits are integrated using a model to which we refer as motif trees (Figure 1 B). These are evaluated and if selected, they are used for further regex-based motif definition and evaluation. Finally, SLiMs that pass all quality criteria are reported to the user. Details of individual steps are as follows.

Selection of closely related orthologs. As SLiMs can either be lost, gained or move along the sequence in evolution, we decided to only include closely related orthologs to the queries for building HMMs. BLAST searches (19) against the NCBI non-redundant (nr-) protein database are carried out to identify close homologs of each input sequence that fulfill the following criteria: *e*-value ≤ 1e–10; identity ≥ 70% and ≤ 95%; coverage ≥ 90%. These settings exclude too similar, as well as too distant orthology candidates. For candidates fulfilling these criteria, reciprocal BLASTs are used to verify orthology relationships; we consider all isoforms of the query for verification of orthologs. In ***advanced mode***, users can provide their own lists of orthologs for further processing.

Residue masking. As motifs are expected to be on a protein's surface, only surface residues are considered for motif prediction. Surface accessibility is computed using NetSurfP (20). Residues with a relative solvent accessibility (RSA) of at least 0.16 are considered exposed (21); all other residues are masked. To allow motif discovery in buried regions, residue masking is optional and can be switched off in ***advanced mode***. Alternatively, users can activate disorder masking using IUPred (22) with the option 'short', as many types of SLiMs are located predominantly in disordered regions (23). As in SLiMFinder, which uses IUPred for disorder masking, residues below the threshold of 0.20 are considered ordered. Users can furthermore specify their own regions of interest by provide a masking file, which is merged with surface accessibility and/or disorder masking, if the checkboxes of the latter are activated. If both checkboxes are deactivated and no file is provided, motif prediction proceeds with unmasked sequences.

Hidden Markov Model creation and comparison. At the core of HH-MOTiF is the comparison of Hidden Markov Models realized with the HH-suite. First, a multiple sequence alignment (MSA) is constructed for each input sequence and its selected orthologs using MAFFT (24); then a HMM for each query is created using *hhmake*. An all-against-all, pairwise HH-comparison is carried out using *hhalign* from the HH-suite. Reporting multiple, also suboptimal hits is allowed by using the '-smin 0 –alt 100′ option in *hhalign.* Furthermore, the '-template_excl' option was added to *hhalign* to permit exclusion of masked residues in both HMMs of a pair. For each HMM pair, the four best hits with a Viterbi score ≥11.0 and ≤40.0 and number of columns ≥ 3 and ≤30 are retained for further evaluation. These hits are used to create the motif trees in the next step (Figure 1B). Longer alignment hits and those with a Viterbi score >40.0 are considered to reflect sequence homology and are therefore not relevant for SLiM detection.

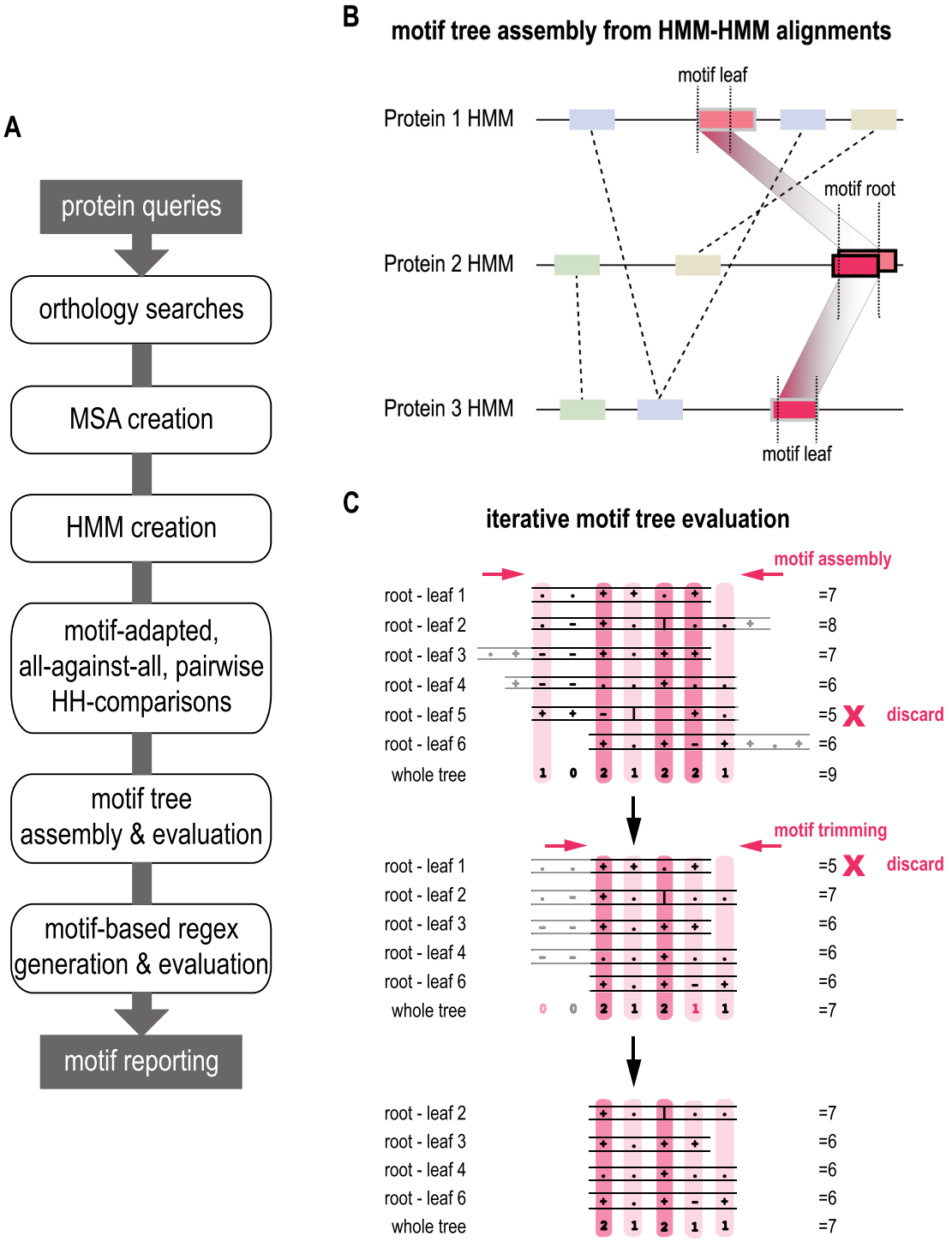

Figure 1. (**A**) Workflow of HH-MOTiF. Starting from a set of input queries, HH-MOTiF first searches for closely related orthologs, builds HMMs and performs an all against all, short linear motif-adapted HH-comparison. Identified motif trees are further evaluated and trimmed prior to reporting (for details, see main text). (**B**) Motif tree assembly from HMM–HMM alignments. Overlapping alignment hits (red-shaded boxes) are joined into hierarchical motif trees. Each tree has a root (overlapping part of black-framed boxes) and leaves (corresponding aligned parts of light gray-framed boxes). Motif leaves are independent. Alignment hits that fail to show a sufficiently strong overlap (non-framed boxes) are ignored. (**C**) Motif tree evaluation. Shown is the iterative process of motif tree evaluation and trimming based on an example of a motif tree with initial 6 leaves in different proteins, assuming $N_{min} = 3$. The score of each position in the whole tree is derived from the alignment sign in *hhalign* ('+','|' 2 points; '.' 1 point; '-',' = ', or gap 0 points). The score at a given position in a leaf cannot be higher than the respective overall position score in the whole tree. Leaves with a score <6 are removed, after which each position is re-evaluated for N_{min}, and if necessary, removed from the motif tree. In the given example, discarding leaf 5 leads to removal of one position and re-assignment of the score for another position in the motif. Consequently, leaf 1 does not fulfill the minimal score requirements and is eliminated from the motif tree. The motif is trimmed to the last conserved position at each of its borders.

Motif tree assembly and evaluation. After all-against-all, pairwise HH-comparisons, we first define so-called motif trees (Figure 1B) as follows: each HH-pair has a maximum set of four retained alignment hits from *hhalign*. If multiple alignment hits overlap by at least three residues, they are joined in a so-called motif root. Each motif root has a set of motif leaves, which are its alignment hits with other HMMs. Together the motif root and its leaves form a motif tree, which is a simplified representation of the underlying putative sequence motif. There can be multiple leaves in the same protein; in this case, the one with the higher score is used for further motif evaluation. To be considered further, a motif tree must be present in a minimum number (N_{min}) of HMMs (Figure 1B). N_{min} is computed on the basis of the dataset size using a dynamically estimated false positive rate (FPR) on negative data that lack a common motif: N_{min} is chosen such that the FPR is $<1\%$ for each set size (for details, see Supplementary Data and Supplementary Table S1). This low FPR is consistent with the $>99\%$ specificity that HH-MOTiF demonstrates on ELM data. However, a motif occupies only a small fraction of a protein's sequence. Therefore, owing to the false positive paradox (25), a high specificity in this case does not ensure an equally high precision.

Motif definition and evaluation. In the next step, positions with significant conservation in the motif tree are identified and evaluated. First, a score for each position in the motif tree must be calculated. We derive this position score from the conservation signs in the *hhalign* output between the motif root and its leaves: motif tree positions with at least $N_{min} - 1$ alignment hits of high conservation (indicated by 'l' or '+') score two points; whereas those, where this requirement is fulfilled by also considering moderate conservation (indicated by '.') score 1 point. 0 points are given, when less than $N_{min} - 1$ alignment hits in a position are conserved. The motif is trimmed to the borders defined by the first and last conserved position. The position scores are used for evaluating both, motif leaves, as well as the motif tree itself (Figure 1C). Weak motif leaves are discarded, the motif tree is iteratively re-evaluated and if necessary, the whole tree is trimmed or even discarded. Motif leaves are evaluated by the sum S of all their position scores. For a leaf to be accepted, its S must be at least 6 (corresponding to e.g. three highly conserved columns). S is also used to evaluate the motif tree itself: for a motif tree to persist, $S \geq 6$ and leaves in $N_{tree} - 1$ proteins must exist, where $N_{tree} \geq N_{min}$.

A motif tree can also have leaves localizing to a larger region of homology, which we can mark, as the corresponding alignment hits have an exceedingly high Viterbi score. Root-leaf pairs, which both locate to the same overall homology region are already discarded at an earlier stage. However, two leaves can still appear within a shared conserved domain or larger homologous region between two query proteins. In this case, only one of the two leaves will be used for scoring, but both will be reported in the results. Thus, the effective number of proteins corrected for homology N_{corr} is used for further calculations instead of the total number N_{tree} of proteins, which participate in a specific motif tree.

Regex generation and statistical evaluation. For motif trees that pass, a regex is generated from its conserved columns for further motif evaluation and final reporting to the user. For each regex, the probability to occur by chance within the submitted dataset is calculated. We have adapted the Šidák correction (26) for multiple testing. In brief, we construct all possible dimers D_{ij} separated by their exact linker lengths as found in all proteins (N_{tree}) that are part of the evaluated motif tree and correct for the product of the sums of the background counts of all D_{ij} in these proteins. This penalizes too vague motifs, low complexity regions, motif occurrences dependent on and reported in long proteins, as well as too long motifs with too many conserved positions, which are in fact rather conserved domains.

A more detailed description of the workflow can be found in Supplementary Data.

The HH-MOTiF web-server

HH-MOTiF is freely available at http://hh-motif.biochem.mpg.de.

For starting an HH-MOTiF search, the user can choose between ***standard*** (Supplementary Figure S1A) and ***advanced mode*** (Supplementary Figure S1B). In ***standard mode***, the input is a set of FASTA-formatted protein queries. Providing an e-mail address is optional. The ***advanced mode*** preferably takes as an input a set of FASTA-formatted protein sequence files in a zip-archive; submission of a single FASTA-file is also possible. Sequences can be submitted with or without orthologs. In the latter case, the orthology search should be activated. The user can provide information on the region of the SLiM, if it has been identified in one of the input proteins. It should be noted at this point that prior knowledge on the approximate localization of a SLiM in a protein sequence—as for instance determined by deletion studies—will greatly enhance the chance to detect the wanted SLiM. Other parameters that can be adapted include restriction of gap length, surface accessibility prediction, disorder masking, homology filtering, as well as the maximal p-value for the regex evaluation (regex p-value). Again, providing an e-mail address is optional, however recommended due to long processing times, especially when orthology searches are activated. The ***proteome-wide search*** (Supplementary Figure S1C) allows users to search for known SLiMs in selected proteomes. A multiple FASTA-file of the SLiM is required as input. The proteome-wide search launches an HMM-to-sequence comparison against the entire proteome.

After submission, the user is forwarded to the results page, which should be bookmarked for future viewing of results. Results are saved for seven days prior to deletion.

The output of an HH-MOTiF search is shown in Figure 2. All identified motif roots are displayed at the top of the page with its associated protein query, as well as the position within the query. This is followed by the full-length sequences of all input queries with the identified motif roots highlighted in red. Corresponding motif leaves, as are found in our chosen example, are highlighted in pink. All elements of a motif tree are linked via a dashed line upon selection of one element. At the right-hand side of the input query with the selected motif, the WebLogo (27), the regular ex-

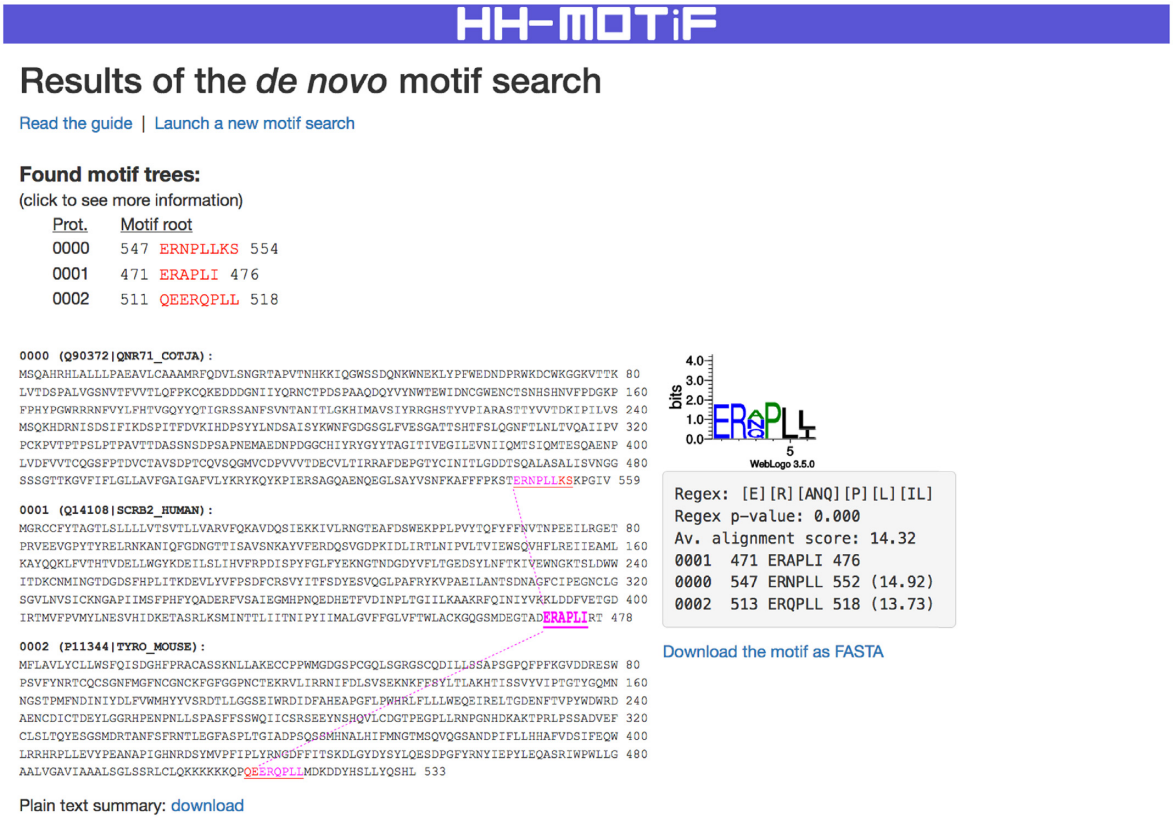

Figure 2. Output web page of HH-MOTiF. Identified SLiMs are reported in association with their input query and their position within the query. Motif trees are highlighted in red in the full-length sequences. Upon selection, a motif is connected to its tree. Next to the full-length sequences, the sequence logo, as well as the Pseudo-MSA of the selected motif is displayed. Results can also be downloaded in FASTA-format.

pression (regex) as well as the pseudo-MSA of the motif are displayed.

To demonstrate the functionality of our web-server, we chose the LysEnd_APsAcLL signal from the TRG class, which is a lysosomal–endosomal targeting signal found in the C-terminus of proteins (28). HH-MOTiF correctly identifies this motif in the three sample proteins QNR-71, SCRAB2 and Tyrosinase and finds no additional shared motif. Next to HH-MOTiF, GLAM2 and SLiMFinder were able to also predict this targeting signal correctly (see Supplementary Data for details).

Optimization and evaluation of HH-MOTiF and comparison with other *de novo* SLiM search tools

First, we used all experimentally verified SLiMs from the ELM database (29), which occur in at least three proteins to optimize HH-MOTiF. These included 176 motifs (classes) grouped into six types. The types CLV and DEG were used as training set; all other types (DOC, LIG, MOD and TRG) were used as test set.

It would be tempting to introduce a simple evaluation protocol, where each annotated SLiM is either 'rediscovered' (true positive) or 'missed' (false negative), as well as each predicted motif is either 'correct' (true positive) or 'incorrect' (false positive). However, in reality, predicted motifs are usually correct only to some extent, as they contain true positive residues or sequence stretches with varying degrees of additional false negatives and false positives. Therefore, we did not rely on a binary classification on motif-level for performance evaluation. Instead, we used performance measures calculated residue-wise and site-wise for all selected 176 SLiMs in the ELM database. We primarily used the balanced F1-score (F1) for evaluating performance, which offers a balanced measure between sensitivity and specificity. We calculated an overall F1 based on simple averaging across all 176 SLiMs from the ELM database. As an approximate binary classification on motif-level, we also counted how many ELM classes out of the 176 reached a residue-wise F1 of at least 0.5. To allow comparison with other statistical evaluations, we also provide data on balanced accuracy (BA) and the performance coefficient (PC) for all tested methods in Supplementary Data, where readers can also find the details on calculating F1, BA and PC.

Being fairly balanced between recall and precision, HH-MOTiF reached a site-based F1 of 0.333 and a residue-based F1 of 0.280 (Table 1 top row, Figure 3 and Supplementary Tables S2, and S3). We used the same dataset to compare our method to other, published *de novo* SLiM search tools. We focused on methods that work non-discriminatively and which provide a stand-alone version for local usage. Software packages considered included MEME (v4.0), GLAM2 (v4.11.1) and SLiMFinder (v5.2.3). The downloadable version of the HMM-based method whmm did not work in our hands. We could there-

Table 1. Performance measures of *de novo* SLiM prediction methods. For details, see main text and Supplementary Tables S2–S6

	Site-based			Residue-based		
	Recall	Precision	F1	Recall	Precision	F1
HH-MOTiF	0.236	0.564	0.333	0.210	0.420	0.280
MEME	0.249	0.099	0.142	0.219	0.061	0.095
GLAM2	0.413	0.164	0.235	0.380	0.073	0.123
SLiMFinder	0.272	0.389	0.320	0.203	0.350	0.257

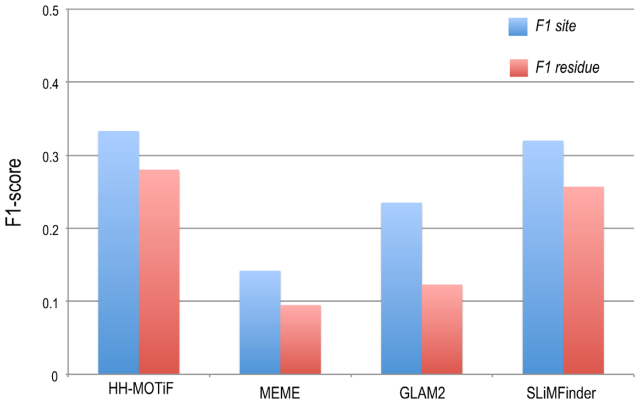

Figure 3. Performance as measured by F1 of HH-MOTiF and other tested *de novo* SLiM search methods. HH-MOTiF was compared against MEME, GLAM2 and SLiMFinder. Both, site- (blue) and residue-based (red) F1 were calculated based on recall and precision of the software suites in discovering SLiMs from the SLiM collection of the ELM database. For details, see Supplementary Tables S2–S6.

fore only compare our results to the originally published data ((10), see Supplementary Data). We tested different parameter values for all selected tools and chose those settings, which yielded the highest F1 (performance measures for all selected settings are available in Supplementary Table S2; performance dependencies of HH-MOTiF on several parameters are discussed in Supplementary Data and are shown in Supplementary Figure S2).

HH-MOTiF had the best F1 compared to all other tested tools, closely followed by SLiMFinder (see Table 1, Figure 3 and Supplementary Tables S2–S6). Our method reached a reasonable recall with a fairly good precision. For SLiMFinder, we used settings that turned the tool more sensitive, at the cost of its otherwise high precision with standard settings (see Supplementary Table S2). GLAM2, on the other hand, scored highest of all in recall, however performed poorly in precision. HH-MOTiF scored also better in site-wise PC than others, while GLAM2 performed better in BA. SLiMFinder had the best residue-wise PC, which could be explained by the fact that it tends to predict SLiMs that are shorter than the ELM annotation and no false positives due to flanking residues are produced. We also observed a dependency of F1 on the size of the dataset for some tools (Supplementary Table S7). HH-MOTiF showed no strong dependency on the set size. SLiMFinder, on the other hand, performed only moderately on small set sizes, however notably outperformed all other tested methods on motif sets containing 11–15 proteins.

Motif sets in ELM are highly variable. They have different lengths, they occur in many or only a few proteins, or they occur more than once in the same protein, representing so-called tandem repeats. These factors could influence the performance of *de novo* motif predictors. Therefore, we also calculated weighted performance measures for all tested tools (Supplementary Table S8). In general, introducing weights for either the number of proteins, the number of sites or the number of residues increased the performance measures for all tools. HH-MOTiF showed a slight bias towards the number of sites; weighting the number of residues per motif exhibited strongest influence on MEME; finally, consistent with our observation that SLiMFinder displayed varying performance on different set sizes, weighting performance measures based on set sizes showed the largest positive influence on SLiMFinder. These data indicate that SLiMFinder performs best on more abundant motifs, MEME on the longest ones, and HH-MOTiF on repeated motifs. Nevertheless, we think that simple averaging is the most useful approach for performance evaluation, as it is reasonably unbiased: it does not allow for 'easy' cases (long, abundant and protein tandem repeats) to outweigh the 'hard' ones (short and less frequent motifs).

DISCUSSION

Our tool combines the to-date most sensitive sequence similarity search method, HH-comparisons, with a representation of SLiMs as motif trees.

HMMs can capture the conservation profile of SLiMs more comprehensively than regexes and outperform pure sequence-based methods in the twilight zone of sequence similarity (18), in which functional SLiMs are to be expected. Moreover, we restrict our HMMs to closely related orthologs. This ensures that the function of the selected orthologs is maintained and that the relevant SLiM is conserved and at the same position in the included sequences.

Treating SLiMs as hierarchical motif trees has two advantages: first, motif trees allow a higher degree of degeneration of SLiMs. While the conservation of the motif-root to each motif-leaf must be over a certain threshold, the conservation between leaves is less critical: a lower conservation between motif-leaves does not disqualify the entire motif tree. Second, the motif-tree structure also allows us to consider flanking residues to a higher degree, even though they will not appear in the reported SLiM. The final SLiM is scored based on the initial pairwise alignment scores, not only on the regions of the SLiM, which is conserved in the minimum set of sequence queries. As a result, flanking regions contribute substantially to identifying SLiMs in HH-MOTiF. Finally, HH-MOTiF can detect several independent motif trees that occur in independent, possibly overlapping subsets of the provided input sequences (data not shown).

HH-MOTiF does not filter full-length sequences for homology, but rather candidate SLiMs at the level of their HMM-alignments. Therefore, it allows for graceful handling of homologous regions, conserved domains, and low complexity regions in the input proteins. As an example, due to extended low complexity regions of proteins containing the ELM motif LIG_EF_ALG2_ABM_2, SLiMFinder classified the whole dataset as too homologous and returned no results, while GLAM2 with default settings returned excessive putative positives, resulting in a precision <1%; with our optimized settings, it failed to find this SLiM. HH-MOTiF on the other hand correctly identified this SLiM as the only hit in the dataset. Consequently, users must not remove too closely related sequences prior to submission to the HH-MOTiF web-server. Furthermore, the fact that we do not explicitly filter for low complexity regions enables HH-MOTiF to distinguish between low complexity SLiMs and unrelated low complexity regions (see exemplary motifs LIG_SH3_3 and LIG_AP_GAE_1 on the *Tests* site of our web-server).

It becomes evident from our tests of different motif discovery tools including our own that their performance depends greatly on the chosen parameter settings, leading either to higher recall or higher precision. A user must therefore carefully evaluate, which settings to choose. Which performance measure is more important might depend on the availability of experimental assays for further verification: if a large-scale assay for testing motif function exists, one might choose a higher recall. If only a very time-consuming assay is at hand, which cannot be scaled up, a higher precision might be desirable.

None of the currently existing SLiM predictors reach an accuracy of more than 35%, including our own method, which again reflects the difficulty of discovering novel SLiMs in proteins and is perhaps inherent to the problem itself. Even unrelated proteins with no functional similarity may share similar motifs (4) and our knowledge on the function of many proteins – and thus the SLiMs they may harbor – is still incomplete: a presumable false positive prediction in the ELM dataset might in fact not be 'false positive'. It is therefore important to note that *de novo* predicted SLiMs should be experimentally verified, which make them difficult to use for purely *in silico* purposes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by the Max Planck Society and the CNRS. We thank Friedhelm Pfeiffer, Frank Schnorrer and Edlira Nano for critical reading of the manuscript.

FUNDING

Funding for open access charge: Max Planck Institute of Biochemistry, Computational Biology Group (Max Planck Society).
Conflict of interest statement. None declared.

REFERENCES

1. Davey,N.E., Cyert,M.S. and Moses,A.M. (2015) Short linear motifs - ex nihilo evolution of protein regulation. *Cell Commun. Signal.*, **13**, 43.
2. Diella,F., Haslam,N., Chica,C., Budd,A., Michael,S., Brown,N.P., Trave,G. and Gibson,T.J. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front. Biosci.*, **13**, 6580–6603.
3. Gould,C.M., Diella,F., Via,A., Puntervoll,P., Gemund,C., Chabanis-Davidson,S., Michael,S., Sayadi,A., Bryne,J.C., Chica,C. *et al.* (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.*, **38**, D167–D180.
4. Johansson,M.U., Zoete,V. and Guex,N. (2013) Recurrent structural motifs in non-homologous protein structures. *Int. J. Mol. Sci.*, **14**, 7795–7814.
5. Prieto,G., Fullaondo,A. and Rodriguez,J.A. (2014) Prediction of nuclear export signals using weighted regular expressions (Wregex). *Bioinformatics (Oxford, England)*, **30**, 1220–1227.
6. Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.
7. Dogruel,M., Down,T.A. and Hubbard,T.J. (2008) NestedMICA as an ab initio protein motif discovery tool. *BMC Bioinformatics*, **9**, 19.
8. Kelil,A., Dubreuil,B., Levy,E.D. and Michnick,S.W. (2014) Fast and accurate discovery of degenerate linear motifs in protein sequences. *PLoS One*, **9**, e106081.
9. Song,T., Bu,X. and Gu,H. (2015) Combining intrinsic disorder prediction and augmented training of hidden Markov models improves discriminative motif discovery. *Chem. Phys. Lett.*, **634**, 243–248.
10. Song,T. and Gu,H. (2015) Discovering short linear protein motif based on selective training of profile hidden Markov models. *J. Theor. Biol.*, **377**, 75–84.
11. Neduva,V. and Russell,R.B. (2006) DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res.*, **34**, W350–W355.
12. Edwards,R.J., Davey,N.E. and Shields,D.C. (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One*, **2**, e967.
13. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME Suite. *Nucleic Acids Res.*, **43**, W39–W49.
14. Frith,M.C., Saunders,N.F., Kobe,B. and Bailey,T.L. (2008) Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol.*, **4**, e1000071.
15. Czeizler,E., Hirvola,T. and Karhu,K. (2015) A graph-theoretical approach for motif discovery in protein sequences. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, doi:10.1109/TCBB.2015.2511750.
16. Bhowmick,P., Guharoy,M. and Tompa,P. (2015) Bioinformatics approaches for predicting disordered protein motifs. *Adv. Exp. Med. Biol.*, **870**, 291–318.
17. Edwards,R.J. and Palopoli,N. (2015) Computational prediction of short linear motifs from protein sequences. *Methods Mol. Biol.*, **1268**, 89–141.
18. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics (Oxford, England)*, **21**, 951–960.
19. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
20. Petersen,B., Petersen,T.N., Andersen,P., Nielsen,M. and Lundegaard,C. (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.*, **9**, 51.
21. Momen-Roknabadi,A., Sadeghi,M., Pezeshk,H. and Marashi,S.A. (2008) Impact of residue accessible surface area on the prediction of protein secondary structures. *BMC Bioinformatics*, **9**, 357.
22. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
23. Stavropoulos,I., Khaldi,N., Davey,N.E., O'Brien,K., Martin,F. and Shields,D.C. (2012) Protein disorder and short conserved motifs in disordered regions are enriched near the cytoplasmic side of single-pass transmembrane proteins. *PLoS One*, **7**, e44389.

24. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
25. Rheinfurth,M. and Howell,L.W. (1998) Probability and statistics in aerospace engineering. *National Aeronautics and Space Administration, Marshall Space Flight Center, National Technical Information Service*. Huntsville,Ala., Springfield, Vol. **16**, https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19980045313.pdf.
26. Wright,S.P. (1992) Adjusted P-values for simultaneous inference. *Biometrics*, **48**, 1005–1013.
27. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
28. Letourneur,F. and Klausner,R.D. (1992) A novel di-leucine motif and a tyrosine-based motif independently mediate lysosomal targeting and endocytosis of CD3 chains. *Cell*, **69**, 1143–1157.
29. Dinkel,H., Michael,S., Weatheritt,R.J., Davey,N.E., Van Roey,K., Altenberg,B., Toedt,G., Uyar,B., Seiler,M., Budd,A. *et al.* (2012) ELM–the database of eukaryotic linear motifs. *Nucleic Acids Res.*, **40**, D242–D251.