NAR Genomics and Bioinformatics, 2024, **6**, lqae103
https://doi.org/10.1093/nargab/lqae103
Advance access publication date: 21 August 2024
Methods Article

FaSTPACE: a fast and scalable tool for peptide alignment and consensus extraction

Hazem M. Kotb and Norman E. Davey*

Division of Cancer Biology, The Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, UK
*To whom correspondence should be addressed. Tel: +44 203 437 7662; Email: norman.davey@icr.ac.uk

Abstract

Several novel high-throughput experimental techniques have been developed in recent years that generate large datasets of putative biologically functional peptides. However, many of the computational tools required to process these datasets have not yet been created. In this study, we introduce FaSTPACE, a fast and scalable computational tool to rapidly align short peptides and extract enriched specificity determinants. The tool aligns peptides in a pairwise manner to produce a position-specific global similarity matrix for each peptide. Peptides are realigned in an iterative manner scoring the updated alignment based on the global similarity matrices of the peptides and updating the global similarity matrices based on the new alignment. The method then iterates until the global similarity matrices converge. Finally, an alignment and consensus motif are extracted from the resulting global similarity matrices. The tool is the first to support custom weighting for the input peptides to satisfy the pressing need to include experimental attributes encoding peptide confidence in specificity determinant extraction. FaSTPACE exhibited state-of-the-art performance and accuracy when benchmarked against similar tools on motif datasets generated using curated peptides and high-throughput data from proteomic peptide phage display. FaSTPACE is available as an open-source Python package and a web server.

Introduction

When mentioned in the context of biological sequences, motifs refer to short nucleic acid or protein sequence patterns encoding structural or functional characteristics (1). Short linear motifs (SLiMs) are a common class of short protein motifs that encode a wide range of functions, from directing protein localization to determining protein stability (2). SLiMs are usually found in the intrinsically disordered regions of the proteome, where they form low-affinity transient binding sites. Most SLiM-containing peptides contain only three to four key residues that encode the major specificity and affinity determinants. These attributes make SLiMs difficult to discover both experimentally and computationally. Classical low-throughput experimental interactomics approaches have been employed to characterize the majority of validated SLiMs. Most investigations have adopted mutagenesis-based methods, such as deleting or mutating motifs, to elucidate the interaction or function of specific individual instances (3–5). Biophysical approaches such as isothermal titration calorimetry, surface plasmon resonance and fluorescence polarization are regularly applied biophysical methods to characterize the binding strength of motif-mediated interactions (6). Crystallography has been a prevalent technique in numerous studies, offering structural insights into motifs bound to a globular binding partner (7,8). Although these techniques are accurate and reliable approaches for locating motifs, they are low-throughput methods regarding the number of motifs studied in parallel with limited scalability. Furthermore, most of these studies depend on *a priori* information suggesting the presence of a putative motif, generally coming from a known consensus, conservation or laborious truncation analyses. Consequently, this makes those experimental methods more suitable for validating motifs rather than discovering them.

In recent years, several new high-throughput approaches have been developed to experimentally discover novel motif-containing peptides in an unbiased manner (9), producing thousands of peptides in a single experiment. Of these approaches, proteomic peptide phage display (ProP-PD) has been applied most extensively to identify novel SLiMs (10–12), and other approaches, such as yeast surface display and bacterial surface display, are similarly scalable (9). The explosion in peptide data has ushered in a new era of interactomics, promising deeper insights into the motif-mediated interactome. However, the exponential growth in data has outstripped the development of tools capable of handling such massive datasets. This necessitates the development of novel methods to analyse these data. A key processing step for peptide datasets is the discovery of specificity determinants in the returned peptides. Enriched specificity determinants indicate that specific peptides have been identified, verifying that the experiment was successful. These specificity determinants can then be used for downstream processing, such as filtering and peptide ranking. The novel high-throughput SLiM-based interactomics approaches have created a requirement for specificity determinant extraction methods with a particular focus on scalability in terms of both the number of screens and, more importantly, the number of peptides. Moreover, for most of these methods, the returned peptides hold distinct levels of confidence encoded, dependent on the experimental approach, by attributes such as sequencing counts and presence in

Received: January 22, 2024. **Revised:** July 4, 2024. **Editorial Decision:** August 1, 2024. **Accepted:** August 5, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

experimental replicates. Therefore, future tools should integrate this information to discriminate between noise and meaningful signals, ensuring that the most robust and biologically relevant information is highlighted.

Several approaches have been developed to extract motifs from datasets of functionally related peptides. SLiMFinder (13) is an enumerative approach that identifies the optimal motifs at the expense of computationally expensive algorithms and represents motifs as regular expressions, which can result in the loss of important quantitative information. MEME (14) is a probabilistic algorithm for discovering motifs by performing expectation maximization to fit a mixture model. It predicts fixed-length ungapped motifs in a given range of motif window lengths. MEME Suite (15) is complemented by the GLAM2 algorithm (16) to allow the discovery of gapped motifs. However, to avoid enumerating the massive number of possible alignments, GLAM2 uses simulated annealing as an optimization method to sample stochastic insertions and deletions while adjusting the alignment of sequences and finding a motif. MUSI (17) detects specificity in a large dataset of short biological sequences by fitting a mixture model with maximum likelihood estimation to generate position-specific scoring matrices (PSSMs). GibbsCluster (18,19) groups the dataset into clusters, maximizing the intracluster alignment fit while minimizing the intercluster similarity. HH-MOTiF (20) is a web-based tool that compares Hidden Markov Model (HMM) profiles of closely related orthologues collected for each query sequence to find SLiMs. However, it favours high recall at the cost of low precision and uses many untunable parameters. Other discriminative methods, which require a negative set in addition to the positive set, are presented to solve the motif-finding problem (21–24). Likewise, they suffer from similar problems: DEME (21), for example, in addition to the requirement of *a priori* negative dataset, requires determining a fixed window length for its substring search space.

Multiple sequence alignment programmes are available with a diversity of algorithms and features. Though there are many approaches to motif discovery, one commonly applied approach is tackling motif discovery as an alignment task to generate the best alignment of a set of motif-containing peptides. Although it will still need additional steps to identify *ab initio* motifs, motif extraction is a more straightforward question with many viable solutions. ClustalW (25) is one of the early general-purpose multiple sequence alignment programmes. It follows the branching order in a guide tree to progressively build up multiple alignments from a series of pairwise alignments. T-Coffee (26) uses a pairwise progressive strategy optimized for speed and simplicity by considering the alignments between all the pairs while carrying out each step. ProbCons (27) introduces probabilistic consistency, incorporating a simple sequence similarity model (a three-state pair HMM) into traditional progressive sum-of-pairs scoring, while Clustal Omega (28) makes the progressive alignment approach scale by enhancing the creation of the guide trees. MAFFT (29) converts sequences to volume and polarity values based on their amino acid composition. Then, it uses the fast Fourier transform algorithm and normalized similarity matrix to align sequences by a progressive method and an iterative refinement method directed by formed guide trees. MUSCLE (30) uses an alignment scoring function called the log-expectation score. It adapts a strategy of fast distance estimation, progressive alignment and refinement techniques. Opal (31) conducts a form-and-polish strategy and employs normalized alignment costs to estimate distances between sequences and a stochastic 3-cut approach to polish produced alignments. Although there are many other tools for identifying motifs computationally (32), not all of them are suitable for proteins generally and SLiMs specifically.

In this paper, we introduce the FaSTPACE (fast and scalable tool for peptide alignment and consensus extraction) tool, designed specifically to tackle the new wave of high-throughput SLiM-based interactomics approaches by focusing on accurate and fast peptide alignment and enriched motif extraction on large peptide datasets. FaSTPACE also allows weights to be integrated into motif identification, thereby allowing experimental confidence to be encoded during motif discovery.

Materials and methods

FaSTPACE is a novel peptide analysis tool for peptide alignment and motif extraction. The tool takes as input a set of peptides and, optionally, a set of positive peptide weights. The output of FaSTPACE is a peptide alignment and the most enriched motif, the motif consensus that is most overrepresented compared to its expected number of occurrences in the set of peptides. The enriched motif is returned with a set of metrics describing the level of enrichment of the motif. The core of the algorithm produces a global similarity matrix (GSM), a probabilistic representation of the similarity of a given peptide to all other peptides in the dataset, that is optimized through multiple iterations. The optimal alignment can then be defined based on the highest scoring peptide GSM, which can be considered the most representative peptide with highest level of similarity to other peptides in the dataset. From each peptide GSM, a motif is extracted, and the most significantly enriched motif in the input peptide set is returned.

Algorithm steps

The method consists of three steps (Figure 1): initiation, refinement and post-processing of the GSM.

Initiation

The initiation step of the pipeline produces an initial GSM for each peptide in the input dataset. The initiation step includes four stages:

1. *Perform an all-by-all peptide alignment of the dataset*: The first step of the FaSTPACE pipeline performs ungapped global alignments between all pairs of sequences (external gaps are permitted, but no gaps are added within either peptide). Motif-containing peptide alignment differs from classical alignment as the expectation in an alignment of motif-containing peptides is that the residues outside the consensus positions of the motif will share limited similarity. FaSTPACE considers this by scoring sequence alignments using a BLOSUM62 matrix with negative values capped at zero, thereby positively scoring similarities and allowing any dissimilarity with no penalty. This overcomes the issue of mismatch intolerance of global alignments for motif-containing peptides where unaligned negatively scoring non-motif positions can have an excessive influence resulting in misalignment. The query peptide is moved across the comparison peptide one residue at a time and an alignment score A_B, the sum of the non-negative BLOSUM62 values for the corresponding residue pairs

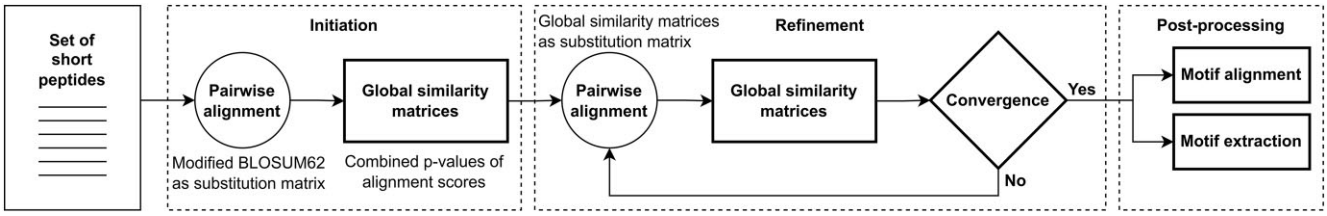

Figure 1. Schema of the three steps in the FaSTPACE tool: (i) *initiation* starts by calculating *P*-values of alignment scores based on a modified BLOSUM62 matrix and uses them to generate per-peptide GSMs; (ii) *refinement* uses the GSMs as substitution matrices to produce new GSMs iteratively; and (iii) *post-processing* uses refined GSMs to align peptides and extract motifs.

in the alignment, is calculated for each peptide pair comparison. The highest scoring comparison is used as the aligned peptide pair.

2. *Estimate statistical significance of alignment scores*: The statistical significance of an alignment score A_B can be represented as the probability of observing that alignment score or higher by chance (33). Therefore, a null model is used to run N simulations, and the P-value of a score A_B is estimated from the number of runs M with a score greater than or equal to A_B (33):

$$P(A_B) = \frac{M}{N}.$$

Although null models to find SLiMs can be obtained by reversing or randomly shuffling dataset sequences (34), these approaches can introduce biases into the significance calculations. Instead, FaSTPACE uses the original dataset sequences as the null model, and P-values are estimated based on best alignment score for each of the peptide pairs obtained from *all-by-all* pairwise alignments in the given dataset. This way, the corrected score $\hat{A}_B$ reflects how unexpected the score A_B is from the dataset alignment score distribution:

$$\hat{A}_B = -\log(P(A_B)).$$

3. *For each peptide pair, calculate a pairwise similarity matrix (PSM)*: Next, PSMs are generated from peptide pairs. Each peptide in the dataset (Figure 2A) is considered a query peptide one at a time and is aligned with all the other peptides one target peptide at a time. For each pairwise alignment, the FaSTPACE scoring algorithm analyses the query sequence with length q by sliding the target sequence across it one position at a time and calculating the alignment score $\hat{A}_B$ each time (Figure 2B). The alignment score $\hat{A}_B$ of the best sliding alignment between the two sequences is stored for each query peptide position p involved in the alignment along with the aligned target peptide residue r in a corresponding matrix cell PSM_{rp}. Therefore, the end result of a pairwise alignment is a $20 \times q$ PSM with one or no non-zero score existing per column p (Figure 2C).
4. *For each peptide, produce an initial GSM*: After considering each peptide in the dataset in turn and aligning the query peptide to all other peptides in the dataset in a pairwise manner, $n - 1$ PSMs have been created for each query peptide (Figure 2D). These matrices are combined for each query peptide by summing the scores in the corresponding matrix cells (Figure 2E and Supplementary Figure S1). Consequently, FaSTPACE generates a GSM containing residue-level scores for each peptide within the dataset. Residues within an enriched consensus exhibit higher GSM_{rp} scores in this matrix (Figure 2F and Supplementary Figure S2).

$$GSM_{rp} = \sum_{k=1}^{n-1} PSM^k{}_{rp}.$$

The combined residue score GSM_{rp} inversely correlates with the probability of observing position p aligned with residue r in a motif by chance.

Refinement

The refinement step of the pipeline iteratively refines the GSM to separate the enriched motif positions from the noise in the initial per-peptide GSMs of the pairwise alignments. The refinement is accomplished by iteratively repeating the initiation step with one major difference. In the refinement step, the normalized initial GSMs of both the query and target peptides replace the non-negative BLOSUM62 matrix when scoring the pairwise alignments. Each iteration of the refinement process starts by normalizing the GSM for each peptide by the Frobenius norm of all GSMs in the dataset (the square root of the sum of the squares of the elements of all the matrices). Next, FaSTPACE performs an all-by-all pairwise alignment resulting in an updated PSM for each peptide pair. The normalized GSMs (for a query peptide i and a target peptide j) function as substitution scores in the computation of the best alignment score A_G during the creation of the updated PSM. Each aligned position p_i involved in the best alignment with a positive score is assigned the alignment score A_G in the updated PSM. These refined PSMs are then cumulatively integrated to form new intermediate per-peptide GSMs.

$$A_{G_{p_i}} = GSM^i{}_{r_i p_i} + GSM^j{}_{r_i p_j},$$

$$A_G = \sum_{p_i=1}^{q} A_{G_{p_i}},$$

$$\forall\ p_i \text{ with } A_{p_i} > 0 :\ PSM^i{}_{r_i p_i} = A_G,$$

$$GSM_{rp} = \sum_{k=1}^{n-1} PSM^k{}_{rp}.$$

The refinement process is performed multiple times in iterations and stops when the average change in per-residue scores GSM_p (position in a peptide in the dataset) falls below a cutoff (default: 1×10^{-6}) or the algorithm hits the maximum number of iterations without convergence (default: 100).

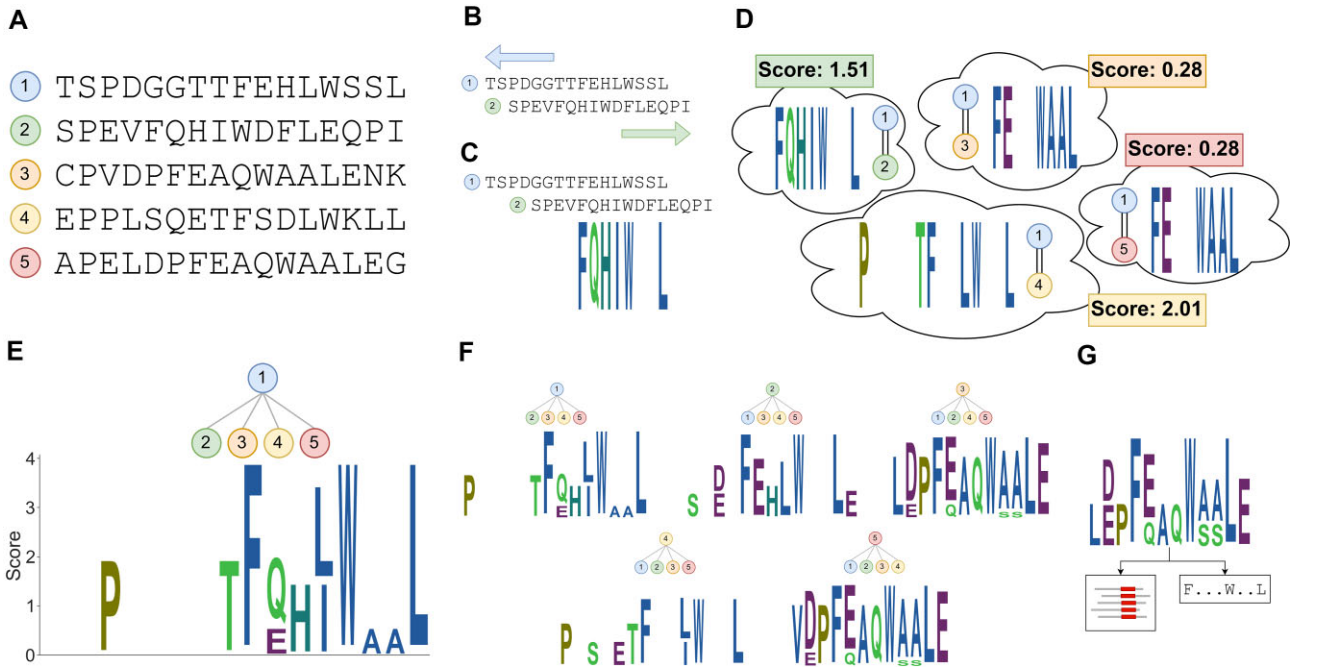

Figure 2. Illustration depicting the step-by-step process in the generation of GSMs using an example dataset. (**A**) An illustrative peptide dataset for the MDM2 SWIB domain with the consensus FxxxWxxL. The dataset consists of five motif-containing peptides (positive peptides) with a length of 16 amino acids without any noise (negative peptides). (**B**) Sliding the first and second peptides relative to each other by one position at a time and assessing the alignment to assign the best aligned regions. (**C**) The alignment that yields the highest score obtained through the sliding process of the first two peptides against each other. (**D**) The alignments with the highest scores derived from the sliding process of the first peptide against each of the other peptides in the dataset. (**E**) The aggregation of all PSMs of the first peptide into a unified GSM. (**F**) GSMs for all peptides in the dataset. (**G**) Generating peptide alignments and extracting motif consensus based on the produced global matrices.

Post-processing

The post-processing step converts the refined GSMs into peptide alignments and enriched motifs (Figure 2G).

Motif alignment

An alignment is generated for each peptide set based on the final GSMs. FaSTPACE picks the highest residue score across all GSMs in the dataset and regenerates the pairwise alignments that lead to that high score by aligning the peptide with the highest score (template peptide) against all other peptides using their similarity matrices for alignment scoring as described in the 'Refinement' section. Therefore, the resulting multiple sequence alignment is a snapshot of the pairwise alignments of all peptides with the most representative peptide in the dataset after the final convergence of the algorithm.

Motif extraction and scoring

Unlike in the alignment generation that has one alignment template peptide, each peptide is a motif template during motif extraction, so a motif is extracted from each peptide based on its GSM. The best motif in the dataset can come from any peptide in the dataset. For each peptide in the dataset, the highest scoring residue in the GSM is considered as a seed for a putative motif. Residues from the GSM are added to the putative motif one at a time in descending order from the maximum score in the GSM. The *P*-value of the extracted motif is calculated with each additional position using a variant of SLiMChance (13). The probability of the motif happening by chance in the dataset is computed using amino acid frequencies of the complete peptide dataset. This computed probability defines the likelihood of the motif appearing at least as often as observed in the dataset. FaSTPACE returns the putative motif for a given peptide when the *P*-value of the formed motif cannot be improved by adding additional positions. Then, a significance score is computed for that motif to adjust the *P*-value for multiple experiments. The maximum number of trials is conservatively estimated by multiplying 20, the number of amino acids in the peptide search space, by the average peptide length and the number of peptides in the dataset. Consequently, FaSTPACE determines the significance of a motif with a *P*-value p_v similar to SLiMChance (13) by applying the binomial distribution to calculate the probability of obtaining one or more successful outcomes within the specified number of trials, each with a probability p_v.

Scoring peptides

FaSTPACE returns three metrics for scoring a peptide based on the presence of a motif: (i) utilizing the *P*-value *(similarity P-value)* and the significance (*similarity significance*) of the motif obtained from the peptide's GSM; (ii) using the motif *P*-value associated with the most significant motif extracted from any peptide in the dataset that matches the peptide (*matched P-value*), along with its significance (*matched significance*); and (iii) employing the peptide's GSM as a scoring matrix by summing the scores attributed to each amino acid at its respective position within the GSM (*similarity score*).

Peptide weighting

Positive peptide weights can easily be integrated into the FaSTPACE algorithm to encode experimental information that reflects peptide confidence into the peptide alignment and motif extraction. Peptides are paired with one another, and align-

ments are produced and subsequently collapsed. Weights can be applied in the aligning and collapsing phase, contributing to the overall value. The introduced weights x should be corrected to make the similarity scores from the GSMs comparable by aligning to similar dataset sizes (the dataset after removing the query peptide from it). FaSTPACE normalizes the weights of the target peptide j after removing the query peptide weight x_i from the dataset (total dataset weight w_t) and multiplies the residue scores PSM_{rp} by the corrected weight w_j while accumulating PSM score to produce GSM scores.

$$w_j = \frac{x_j \times w_t}{w_t - x_i},$$

$$GSM_{rp} = \sum_{k=1}^{n-1} w_k \times PSM^k{}_{rp}.$$

Benchmarking dataset

FaSTPACE's ability to align and extract motifs was benchmarked on the *Motif Extraction from PeptidesBenchmarking* (MEP-Bench) datasets (dataset: https://zenodo.org/records/10467208; bioRxiv: 10.1101/2024.01.12.574168). The MEP-Bench test sets were sampled from annotated motif instances for a given ELM class in the ELM database (35). Peptide sets were built with different attributes to characterize the ability of the method to perform on a range of real-world applications. The peptide set attributes tested were peptide length (12, 16, 20, 24, 28, 32), number of positive (motif-containing) peptides (5, 10, 15, 20, 30, 40, 50) and percentages of noise using negative peptides sampled from other ELM classes (0–800%). Each test case contained numerous peptide sets (100 replicates per attribute) to produce 653 846 peptide sets in total. A series of motif discovery tools (SLiMFinder, MUSI, GibbsCluster, MEME) and alignment tools (OPAL, Clustal Omega, MUSCLE, ClustalW, ProbCons, MAFFT, T-Coffee) were evaluated on the peptide sets. All tools were used with default parameters. It should be noted that the quality of the results of each method could potentially be improved by non-default parameters, particularly the multiple sequence alignment-based tools that are not optimized for short peptide alignment. The results of all methods were converted to peptide alignments to allow direct comparison of the disparate output format. The percentage of correctly aligned motifs in the positive peptides was used as the evaluation metric. The number of correctly aligned peptides represents the highest count of positive peptides, wherein the initial motif position is aligned in each of them. Subsequently, the percentage of this count is computed based on the total number of positive peptides within the provided dataset.

Application to ProP-PD data

FaSTPACE was evaluated through benchmarking on ProP-PD as a high-throughput technique generating real-world data. A ProP-PD experiment is a protein–protein interaction assay between a bait protein and a ProP-PD library. A *screen* is a set of replicate ProP-PD experiments for a given bait against a single ProP-PD library. The enriched phages for each replicate are collected and sequenced through several phage display *selections* on sequential days. Hence, next generation sequencing-derived peptide counts for each unique peptide that binds a specific screened bait using the library are obtained. These counts are normalized to the proportion of observed peptides in a single selection to produce normalized peptide counts and combined into a single replicate value for each peptide by calculating the mean. After that, the per-peptide replicate mean counts are combined into a single screen value by retaking the mean again. Intuitively, motif-containing biologically relevant peptides should bind their baits with higher mean normalized peptide counts, be found in more replicates and be more specific by showing less promiscuity through different baits. Therefore, we tested FaSTPACE on a ProP-PD dataset of 40 screens (12). The FaSTPACE evaluation process involves three distinct runs. In the first two, the tool is executed twice on each screen's peptides, once with mean normalized counts as weights and once without any weights. In the third run, FaSTPACE is evaluated on a filtered list of peptides per screen. This filtered list comprises peptides identified in more than one replicate and those overlapping with at least one other peptide in the screen. Promiscuous peptides are considered by removing peptides from the filtered list that have <20% of all the counts across all screens in a given screen. To provide a basis for comparison, SLiMFinder is also run on this filtered list. For each of the four experiments, consensus information is extracted from every screen. Subsequently, CompariMotif (36) normalized shared information content scores are computed to quantify the similarity of the extracted motif for the selection to the expected motif for the bait protein as defined by the ELM database. This comprehensive approach allows for a thorough evaluation of FaSTPACE's performance across various conditions and a comparison with SLiMFinder.

Results

Evaluating FaSTPACE on artificially generated datasets

FaSTPACE is evaluated on datasets (dataset: https://zenodo.org/records/10467208; bioRxiv: 10.1101/2024.01.12.574168) generated from validated motif instances from the ELM database (see the 'Materials and methods' section) (Figure 3A). The tool was tested against 11 other tools that can be roughly classified into two groups: motif discovery tools (SLiMFinder, MUSI, GibbsCluster, MEME) and alignment tools (OPAL, Clustal Omega, MUSCLE, ClustalW, ProbCons, MAFFT, T-Coffee). FaSTPACE aligns the given sequences based on their GSMs instead of the found motif, so similar to alignment tools, it always manages to align the given data (percentage runs completed: FaSTPACE 100%, SLiMFinder 93%, MUSI 96%, GibbsCluster 99%, MEME 100%, Opal 100%, Clustal Omega 100%, MUSCLE 100%, ClustalW 100%, ProbCons 100%, MAFFT 100%, T-Coffee 100%) (Figure 3B). Average running times and alignment accuracies are calculated per tool on groups of test sets with the same number of positive peptides, peptide length and noise percentage (Supplementary Table S1). FaSTPACE showed higher performance than all tested motif discovery tools and most of the alignment tools with a median of average running times of <1 s (FaSTPACE with no refinement 0.11 s, FaSTPACE 0.25 s, SLiMFinder 73.9 s, MUSI 8.6 s, GibbsCluster 2.2 s, MEME 2.2 s, Opal 3.3 s, Clustal Omega 13.2 s, MUSCLE 0.10 s, ClustalW 0.06 s, ProbCons 0.36 s, MAFFT 0.87 s, T-Coffee 1.6 s) (Figure 3C). The running time of FaSTPACE is influenced by both the number and length of peptides in the datasets. This correlation becomes more apparent when the effect of

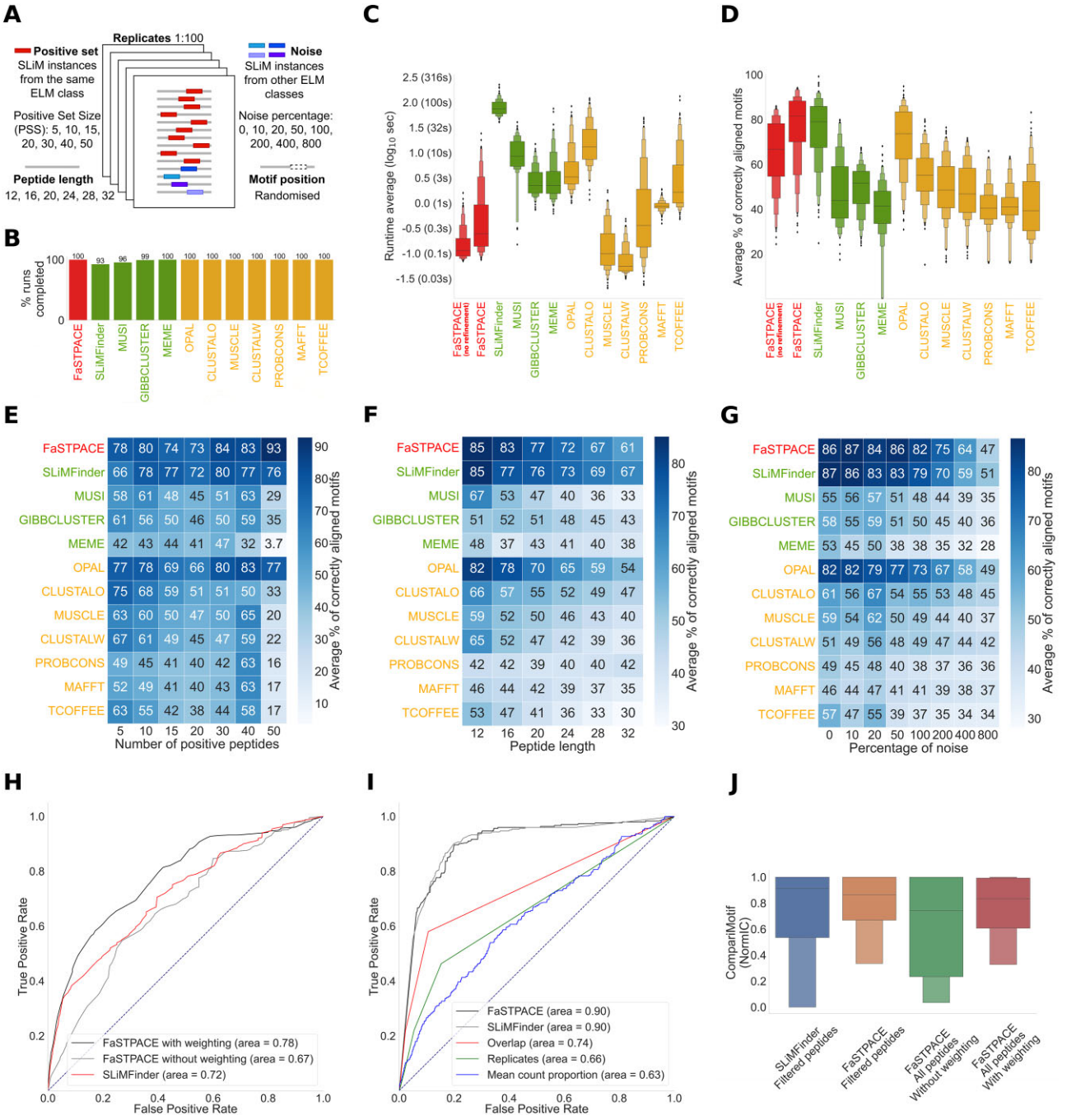

Figure 3. The evaluation of FaSTPACE on artificial data and ProP-PD data. (**A**) Overview of the generation of the benchmarking test sets from the ELM resource. The dataset includes 653 846 artificial sets of motif-containing peptides representing 133 groups based on the number of motif-containing peptides, peptide length and noise percentage. (**B**) A bar plot of the percentage of aligned test sets that return peptide alignment and/or motif extraction data from the benchmarking test sets. (**C**) A boxen plot of the average running time of each group of peptide sets for the different tools on the benchmarking test sets. (**D**) A boxen plot of the average percentage of correctly aligned motifs for each group for the different tools on the benchmarking test sets. (**E**) Heatmap of the percentage of correctly aligned peptides in groups with different numbers of positive peptides. (**F**) Heatmap of the percentage of correctly aligned peptides in groups with different peptide lengths. (**G**) Heatmap of the percentage of correctly aligned peptides in groups with different noise percentages. (**H**) Average receiver operating characteristic (ROC) curves comparing FaSTPACE (with/without weighting) and SLiMFinder in discriminating curated motif-containing peptides from others in ProP-PD screens. (**I**) ROC curves comparing different peptide scoring methods in discriminating curated motif-containing peptides from others in ProP-PD screens. (**J**) A boxen plot of CompariMotif scores comparing the expected motif in the ELM database for each bait protein of the 34 baits with ELM motifs versus the motif discovered from the ProP-PD screen by SLiMFinder and FaSTPACE.

the added noise to datasets is eliminated, as demonstrated by assessing solely the running time of the initiation step (Supplementary Figure S3). Specifically, the running time of the algorithm is found to be quadratically correlated with the total dataset size (Supplementary Figure S4A), with convergence occurring more rapidly and with less noise when the dataset contains an enriched motif, thus limiting the number of refinement iterations. A significant advantage of FaSTPACE lies in its memory consumption, which grows linearly with the dataset size regardless of the noise present in the dataset (Supplementary Figure S4B). FaSTPACE outperforms all benchmarked tools in the accuracy of generated alignments with a median of the average percentage of correctly aligned motifs of 82% (FaSTPACE with no refinement 67%, FaSTPACE 82%, SLiMFinder 79%, MUSI 44%, GibbsCluster 52%, MEME 42%, Opal 74%, Clustal Omega 55%, MUSCLE 49%, ClustalW 47%, ProbCons 41%, MAFFT 41%, T-Coffee 39%) (Figure 3D). Removing the penalization of dissimilar residues in the alignment, by ignoring negative values in the BLOSUM matrix, improved results compared to application of the standard BLOSUM matrix (FaSTPACE 82% aligned motifs, FaSTPACE with standard BLOSUM 77% aligned motifs) (Supplementary Figure S5). FaSTPACE is significantly better than any other tool when the dataset contains a large number of motif-containing peptides (Figure 3E), and gives better performance when the peptides in the dataset are shorter (Figure 3F). It is also noted that the number of positive peptides and their length in a dataset are more critical factors in aligning motifs correctly than the noise percentage added to the dataset (Figure 3G).

Evaluating FaSTPACE on ProP-PD benchmarking data

We evaluated FaSTPACE on a ProP-PD dataset (12) and compared its scores (matching motif *P*-value) against other methods of scoring peptides for containing a motif (mean normalized counts, replicate score, overlapping score and SLiMFinder score) (Supplementary Table S2). In Figure 3H, each motif-containing screen's ROC curves are calculated, and their average is plotted. FaSTPACE matching motif *P*-value scores outperform all the other scores in discriminating motif-containing peptides (Area under the ROC Curve (AUC): 0.78). For weighting peptides, FaSTPACE uses mean normalized counts with a modest discrimination capability (AUC: 0.78). SLiMFinder *P*-value score comes next after FaSTPACE scores (AUC: 0.72). SLiMFinder is used to find motifs in replicated peptides (peptides that appear in more than one replicate of a screen) and overlapping peptides (peptides with other peptides overlapping them in the screen). SLiMFinder extracted motif aligns those peptides, and a PSSM is built from them. Then, PSSMSearch scores all screen peptides based on the built PSSM. It gives the SLiMFinder-based score an edge as it runs on clean data of replicated and overlapping peptides with descent discrimination power (replicate AUC: 0.59; overlap AUC: 0.63) (Figure 3I). Despite that, FaSTPACE outperforms SLiMFinder-based scores by a wide margin. In Figure 3I, all the peptides in all the ProP-PD screens are used in building a single plot of ROC curves. To account for the differences in the number of peptides in each screen, FaSTPACE is run with a normalization factor of 1000 (all the screen sizes are corrected to be 1000). Given those settings, FaSTPACE and SLiMFinder reach similar results (AUC: 0.90). FaSTPACE extracted motifs match the expected motifs from the ELM database and the SLiMFinder motif in most screens, as evidenced by the comparisons made using CompariMotif (Supplementary Table S3 and Figure 3J). It suggests a high potential for improving the FaSTPACE results by integrating more experimental features into the weighting schemes.

The FaSTPACE web server

The FaSTPACE tool is now accessible as a web server, which can be found at https://slim.icr.ac.uk/projects/fastpace. Users can input peptides into the tool via the web server, and it will provide comprehensive information about the best-extracted motif and alignment. The 'Motif' section of the results presents a sequence logo of the peptide with the highest residue score in its similarity matrix, the regular expression of the best *P*-value motif extracted from a peptide's similarity matrix and the scores associated with the best-extracted motif. In the 'Alignment' section, users will find a table displaying aligned peptides, along with information extracted from the similarity matrix of each peptide.

Discussion

Identifying overrepresented patterns in datasets of short peptides is crucial for many biological problems and has been a long-standing challenge for several years. Numerous tools have emerged to tackle this problem by extracting specificity determinants, such as position-specific scoring matrices or regular expressions, from sets of biologically related peptides. A common problem of all tools is the trade-off between speed and accuracy. The available methods can be grouped as accurate enumerative approaches with exponential running time, or as fast but suboptimal heuristic methods. FaSTPACE is the first tool with both speed and accuracy. The FaSTPACE approach uses per-residue motif-oriented alignment scores from all input peptides to reach local optimality, followed by iterative refinement to achieve global optimality. FaSTPACE can also integrate peptide-specific experimental information as peptide weights into this motif extraction process, eliminating the need to pre-filter data that can lead to the loss of relevant information. FaSTPACE has a polynomial running time that increases quadratically with the dataset size, enabling the tool to be applied to datasets of thousands of peptides. This scalability and the algorithm's weighting capabilities enabled FaSTPACE to extract specificity determinants from large datasets of unprocessed peptides, such as the phage display examples presented.

The major limitation of FaSTPACE is the inability to extract variable-length gapped motifs. The addition of variable-length gaps would lead to an exponential growth in the running time and would likely introduce more noise to the algorithm's results. FaSTPACE reaches state-of-the-art accuracy in motif extraction in the benchmarking results of the current internal-gap-free algorithm, indicating that the fixed-length submotifs extracted in these cases describe the specificity determinants of the data sufficiently. Consequently, we believe that the addition of variable-length gaps would not provide significant benefits compared to the additional complexity it would introduce to the algorithm and that it is not a worthwhile trade-off for the suboptimal speed and accuracy of fixed motif extraction. Furthermore, tools such as SLiMFinder allow variable-length gaps and provide high-accuracy motif extraction;

however, the field would benefit from further research in this area to allow variable-length gapped motifs with speed, accuracy and at scale. An additional potential improvement in future releases is the optimization of the similarity matrix applied in the initiation step. FaSTPACE currently utilizes the BLOSUM matrix to quantify similarity between peptides. Several additional similarity matrices exist that may yield more accurate results. However, intuitively, the development of motif alignment-specific similarity matrices based on characterized motif instances would provide the optimal solution.

In summary, we have developed FaSTPACE, a fast and scalable tool to align and extract motifs from sets of short peptides that allows peptide weighting. FaSTPACE is an important addition to the set of tools tackling the current explosion of peptide-centric biological data.

Data availability

FaSTPACE is implemented as an open-source Python C++ extension combining the speed and performance of C++ and the usability of Python. Therefore, it can be installed as a Python PyPI package. The source code is available at GitHub (https://github.com/hkotb/fastpace) and figshare (https://doi.org/10.6084/m9.figshare.24996518.v2). The FaSTPACE tool is also accessible as a web server at https://slim.icr.ac.uk/projects/fastpace. The open-source version of FaSTPACE has additional functionality that allows users to rerun the algorithm multiple times, each time masking motifs that were identified in previous runs. This iterative approach allows the algorithm to detect less populated motif classes that might remain undetected.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

The authors express their gratitude to Ylva Ivarsson and Leandro Simonetti for graciously hosting the project during its initial development and for their assistance in supplying the ELM-based artificially generated data and scripts in the early stages. Special thanks are extended to Izabella Krystkowiak for her support in setting up the FaSTPACE server.

Funding

HORIZON EUROPE Marie Sklodowska-Curie Actions [860517 to H.M.K.]; Cancer Research UK [C68484/A28159 to N.E.D.].

Conflict of interest statement

None declared.

References

1. Mohamed,S.A.E.H., Elloumi,M. and Thompson,J.D. (2016) Motif discovery in protein sequences. In: Ramakrishnan,S. (ed.) *Pattern Recognition: Analysis and Applications*. InTech, London, UK.
2. Van Roey,K., Uyar,B., Weatheritt,R.J., Dinkel,H., Seiler,M., Budd,A., Gibson,T.J. and Davey,N.E. (2014) Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem. Rev.*, **114**, 6733–6778.
3. Zhang,C., Li,X., Adelmant,G., Dobbins,J., Geisen,C., Oser,M.G., Wucherpfenning,K.W., Marto,J.A. and Kaelin,W.G. Jr (2015) Peptidic degron in EID1 is recognized by an SCF E3 ligase complex containing the orphan F-box protein FBXO21. *Proc. Natl Acad. Sci. U.S.A.*, **112**, 15372–15377.
4. Clijsters,L., Hoencamp,C., Calis,J.J.A., Marzio,A., Handgraaf,S.M., Cuitino,M.C., Rosenberg,B.R., Leone,G. and Pagano,M. (2019) Cyclin F controls cell-cycle transcriptional outputs by directing the degradation of the three activator E2Fs. *Mol. Cell*, **74**, 1264–1277.
5. Zhang,J., Bu,X., Wang,H., Zhu,Y., Geng,Y., Nihira,N.T., Tan,Y., Ci,Y., Wu,F., Dai,X., *et al.* (2017) Cyclin D–CDK4 kinase destabilizes PD-L1 via cullin 3–SPOP to control cancer immune surveillance. *Nature*, **553**, 91–95.
6. Blikstad,C. and Ivarsson,Y. (2015) High-throughput methods for identification of protein–protein interactions involving short linear motifs. *Cell Commun. Signal.*, **13**, 38.
7. Chen,Z., Wasney,G.A., Picaud,S., Filippakopoulos,P., Vedadi,M., D'Angiolella,V. and Bullock,A.N. (2020) Identification of a PGXPP degron motif in dishevelled and structural basis for its binding to the E3 ligase KLHL12. *Open Biol.*, **10**, 200041.
8. Yan,X., Wang,X., Li,Y., Zhou,M., Li,Y., Song,L., Mi,W., Min,J. and Dong,C. (2021) Molecular basis for ubiquitin ligase CRL2FEM1C-mediated recognition of C-degron. *Nat. Chem. Biol.*, **17**, 263–271.
9. Davey,N.E., Simonetti,L. and Ivarsson,Y. (2023) The next wave of interactomics: mapping the SLiM-based interactions of the intrinsically disordered proteome. *Curr. Opin. Struct. Biol.*, **80**, 102593.
10. Ivarsson,Y., Arnold,R., McLaughlin,M., Nim,S., Joshi,R., Ray,D., Liu,B., Teyra,J., Pawson,T., Moffat,J., *et al.* (2014) Large-scale interaction profiling of PDZ domains through proteomic peptide-phage display using human and viral phage peptidomes. *Proc. Natl Acad. Sci. U.S.A.*, **111**, 2542–2547.
11. Davey,N.E., Seo,M.-H., Yadav,V.K., Jeon,J., Nim,S., Krystkowiak,I., Blikstad,C., Dong,D., Markova,N., Kim,P.M., *et al.* (2017) Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome. *FEBS J.*, **284**, 485–498.
12. Benz,C., Ali,M., Krystkowiak,I., Simonetti,L., Sayadi,A., Mihalic,F., Kliche,J., Andersson,E., Jemth,P., Davey,N.E., *et al.* (2022) Proteome-scale mapping of binding sites in the unstructured regions of the human proteome. *Mol. Syst. Biol.*, **18**, e10584.
13. 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.
14. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
15. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
16. Frith,M.C., Saunders,N.F.W., Kobe,B. and Bailey,T.L. (2008) Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol.*, **4**, e1000071.
17. Kim,T., Tyndel,M.S., Huang,H., Sidhu,S.S., Bader,G.D., Gfeller,D. and Kim,P.M. (2012) MUSI: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets. *Nucleic Acids Res.*, **40**, e47.
18. Andreatta,M., Lund,O. and Nielsen,M. (2013) Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics*, **29**, 8–14.
19. Andreatta,M., Alvarez,B. and Nielsen,M. (2017) GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res.*, **45**, W458–W463.
20. Prytuliak,R., Volkmer,M., Meier,M. and Habermann,B.H. (2017) HH-MOTiF: *de novo* detection of short linear motifs in proteins

by hidden Markov model comparisons. *Nucleic Acids Res.*, **45**, 10921.
21. Redhead,E. and Bailey,T.L. (2007) Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics*, **8**, 385.
22. 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.
23. Mehdi,A.M., Sehgal,M.S.B., Kobe,B., Bailey,T.L. and Bodén,M. (2013) DLocalMotif: a discriminative approach for discovering local motifs in protein sequences. *Bioinformatics*, **29**, 39–46.
24. Asgari,E., McHardy,A.C. and Mofrad,M.R.K. (2019) Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX). *Sci. Rep.*, **9**, 3577.
25. 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.
26. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
27. Do,C.B., Mahabhashyam,M.S.P., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
28. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
29. Katoh,K., Misawa,K., Kuma,K.-I. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
30. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
31. Wheeler,T.J. and Kececioglu,J.D. (2007) Multiple alignment by aligning alignments. *Bioinformatics*, **23**, i559–i68.
32. Hashim,F.A., Mabrouk,M.S. and Al-Atabany,W. (2019) Review of different sequence motif finding algorithms. *Avicenna J. Med. Biotechnol.*, **11**, 130–148.
33. Mitrophanov,A.Y. and Borodovsky,M. (2006) Statistical significance in biological sequence analysis. *Brief. Bioinform.*, **7**, 2–24.
34. Krystkowiak,I., Manguy,J. and Davey,N.E. (2018) PSSMSearch: a server for modeling, visualization, proteome-wide discovery and annotation of protein motif specificity determinants. *Nucleic Acids Res.*, **46**, W235–W241.
35. Kumar,M., Michael,S., Alvarado-Valverde,J., Mészáros,B., Sámano-Sánchez,H., Zeke,A., Dobson,L., Lazar,T., Örd,M., Nagpal,A., *et al.* (2022) The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res.*, **50**, D497–D508.
36. Edwards,R.J., Davey,N.E. and Shields,D.C. (2008) CompariMotif: quick and easy comparisons of sequence motifs. *Bioinformatics*, **24**, 1307–1309.

Received: January 22, 2024. **Revised:** July 4, 2024. **Editorial Decision:** August 1, 2024. **Accepted:** August 5, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.