, , pp. 1–10

doi:
Paper

PAPER

GBSC: Graph-Based Sequence Clustering method for similar short tandem repeats in protein sequences

Patryk Jarnot,[1] Joanna Ziemska-Legiecka,[2] Marcin Grynberg,[2] Vasilis J. Promponas[3,*] and Aleksandra Gruca[1,*]

[1]Department of Computer Networks and Systems, Silesian University of Technology, Poland, [2]Institute of Biochemistry and Biophysics, PAS, Poland and [3]Department of Biological Sciences, University of Cyprus, Cyprus
*Corresponding author. vprobon@ucy.ac.cy or aleksandra.gruca@polsl.pl
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Motivation: Short tandem repeats (STRs) are abundant in protein sequences and play an important role in determining their structures and functions. Strikingly, the unusual compositional characteristics of tandem repeats break classical sequence analysis tools. **Results**: We here establish the first algorithm to effectively identify and cluster STRs: Graph-Based Sequence Clustering (GBSC) features linear time complexity, and clusters protein sequence fragments based on their STRs, while allowing for insertions and mutations, supporting an analysis of imperfect or cryptic repeats. Due to its computational efficacy, our algorithm can be used to systematically scan for patterns in large datasets. We compare our method both with state-of-the-art methods for identifying STRs in proteins and alternative clustering approaches. Unlike existing STR analysis methods, GBSC clusters repeat patterns rather than raw sequences, operating at the level of structural repeat identity, while tolerating biological variations and preventing erroneous merging of structurally and functionally distinct motifs. Whereas functional annotation is typically only available at the protein level, the functions of individual STRs and sequences of adjacent STRs remains largely unknown. On a challenging use-case we here demonstrate and discuss how our method can be used to associate previously unannotated repetitive protein fragments with similar ones, allowing a transfer of annotation by similarity. For the first time, GBSC offers a tool that systematically extends this fundamental bioinformatics principle to low-complexity regions across large datasets. **Availability and Implementation**: GBSC is available at GitHub https://github.com/patryk-jarnot/GBSC and https://doi.org/10.5281/zenodo.18965247. The data and scripts to reproduce the analysis are available at https://doi.org/10.5281/zenodo.16906653.

Key words: Protein Sequences, Short Tandem Repeats, Identification, Clustering, Functional Analysis

Introduction

Protein sequence clustering methods are fundamental in the discovery of protein properties, as they allow to identify groups of potentially orthologous proteins by way of their sequence similarities and, based on the likely orthologous relationship, infer functions of so-far uncharacterized proteins from similar proteins that are better annotated. Existing methods fall into two broad categories: alignment-based and alignment-free approaches. Alignment-based tools such as CD-HIT [Li and Godzik, 2006], UCLUST [Edgar, 2010] and MMseqs2 [Steinegger and Söding, 2017] group sequences by computing pairwise or representative-sequence similarity using local or global alignments, accepting sequences into a cluster when they meet a predefined identity threshold [Wei et al., 2023]. While highly effective for redundancy reduction and homology-based grouping of full-length protein sequences [Suzek et al., 2015, Feldbauer et al., 2020], these methods assume positional independence of residues, which is a critical assumption known to be strongly violated in short tandem repeats (STRs). Consequently they are fundamentally ill-suited for clustering such regions as short tandem repeats are inherently repetitive, low-complexity sequences whose alignment scores are dominated by repeat copy number and unit length rather than by the biological identity of the underlying pattern, causing structurally distinct STRs to be erroneously merged. Alignment-free approaches, including k-mer frequency profiling [Moeckel et al., 2024], feature vector methods based on physicochemical properties [He et al., 2017], and MinHash-based sketching [Abnousi et al., 2018], appear to offer a more flexible alternative, yet they inherit the same conceptual limitation: by reducing sequences to aggregate compositional or statistical descriptors, they discard the periodic structural information intrinsic to tandem repeats, rendering them equally unsuitable for pattern-based STR discrimination. More recently, representations derived

from protein language models (PLMs) such as ESM-2 [Lin et al., 2023] and ProtTrans [Elnaggar et al., 2021] have been proposed as the basis for sequence clustering, with embeddings capturing contextual residue relationships [Pereira et al., 2025]. However, PLM embeddings are trained primarily on globular, non-repetitive sequences, and while internal attention mechanisms can reliably detect periodic structure in ideal repeats, PML performance has been shown to be low and unstable specifically for repeats containing insertions and deletions [Kesten-Pomeranz et al., 2026], a structural feature common in imperfect STRs, making PLM-based clustering unreliable for pattern-based STR discrimination.

As a result, functional low-complexity regions have to date largely been identified by dedicated tools for specific instances, typically exploiting instance specific properties. For example, in an earlier study, we found that within the intrinsically disordered region of the delta subunit of bacterial RNA polymerase, that includes a positively charged stretch followed by a negatively charged one. These stretches fold back on each other, compacting the overall structure while maintaining its flexibility, which seems to be essential for the subunit function. Searching for similar charge-patterned regions subsequently revealed other RNA-binding proteins [Kubáň et al., 2019]. Similarly, to identify such regions, scientists used regular expressions for collagen-like repetitions [Teuscher et al., 2019] or sequence profiles to identify HEAT repeats [Kamel et al., 2021]. While these approaches are already used for the automatic annotation of protein databases, low-complexity regions remain functionally uncharacterized [Consortium, 2023, Sigrist et al., 2012].

We here introduce a complementary and generic approach to address this gap by providing a new method for investigating the likely functions of unannotated short tandem repeats (STRs), a specific class of low-complexity regions. These are known to evolve by replication slippage and often confer function in line with their biochemical properties [Lin and Kussell, 2012]. In contrast to existing methods for protein sequence clustering, our Graph-Based Sequence Clustering (GBSC) method introduces a conceptually distinct paradigm: it first identifies STR patterns using a graph-based approach and then clusters those patterns, rather than the raw sequences themselves. This two-stage, pattern-centric strategy operates at the level of structural repeat identity rather than sequence-level statistics. Crucially, it retains biological flexibility by tolerating insertions and point mutations within sequences sharing the same pattern, while overcoming a critical limitation of similarity-based methods that causes the erroneous grouping of structurally and functionally distinct STRs into the same cluster. To our knowledge, GBSC is the first STR clustering method to explicitly decouple pattern recognition from sequence similarity scoring, offering a more principled and biologically interpretable framework for repeat characterization.

Each GBSC cluster provides a collection of protein fragments with highly similar STRs for further analysis. We demonstrate on an illustrative use-case how this enables inference the RNA/DNA-binding function of STRs that have no annotated function by combining protein-level functional annotation with STR similarity. This approach allows the transfer of protein-level annotation to higher-resolution knowledge about the STR itself, providing high-confidence candidates for experimental validation. By introducing the *s* measure for identifying the most functionally concise clusters we provide, for the first time, a tool [1] that allows the systematic transfer of functional annotations by similarity to STRs in large datasets.

Methods

In this section, we introduced the GBSC method and described its analysis. We began by outlining the identification process of GBSC, discussing its parameters, and providing an example of STR identification. We also illustrated how GBSC clusters STRs based on similarity. Then, we compared GBSC with other methods for clustering protein sequences. Finally, we enriched GBSC clusters with Gene Ontology annotations, validated their confidence and analysed the cluster containing RG repeats.

Identification

The identification and analysis of STRs requires a diverse toolkit, as these sequences often vary in purity and periodicity. Several computational tools have been developed to detect those patterns in protein sequences, each employing distinct algorithmic strategies. Among the most widely used detection tools, SEG identifies low-complexity segments based on local compositional complexity [Wootton and Federhen, 1993], while fLPS detects compositionally biased regions using a statistical framework sensitive to amino acid composition [Harrison, 2021]. LCD-Composer extends this approach by combining amino acid composition and the linear dispersion of residues for flexible LCR identification [Cascarina and Ross, 2022], and CAST offers an integrative compositional assessment of LCR structure [Promponas et al., 2000]. Beyond compositional bias, repeat-oriented tools address sequence periodicity: GBA provides a graph-based method for repeat analysis [Li and Kahveci, 2006], T-REKS employs a k-means clustering algorithm to identify periodic patterns [Jorda and Kajava, 2009], XSTREAM uses heuristic pattern matching to detect repeat units across varying period lengths [Newman and Cooper, 2007], and SIMPLE targets highly cryptic (non-tandem) repeats by flagging residues whose motif frequency exceeds that expected from randomized sequences [Albà et al., 2002]. Given the diversity of these approaches — and the fact that no single tool captures the full spectrum of STR types — we provide a comparative analysis across multiple tools to demonstrate that the STR identification method implemented by the GBSC algorithm is comparable with other established methods, collectively covering a broad range of repeat types, thus validating its utility as a robust and comprehensive approach for STR characterization. This comparison is included in Supplementary Material 1.

GBSC scans a sequence and constructs a directed graph of overlapping dimers to find repeating patterns. It marks fragments represented by cycles in the graph representation as repetitive regions. The method begins by dividing the sequence into dimers and iteratively processing them. For each dimer it creates a node if it is absent in the graph. Each node has a lifetime variable initially set to 0. This lifetime variable is incremented by 1 during each iteration; however, if a node is revisited, its lifetime resets to 0. Nodes represented by two neighborly dimers in the sequence are connected by edges. The weights of these edges indicate the number of occurrences of adjacent dimers, initially set to 1 and increased each time a transition between the same dimers occurs in the same

[1] https://github.com/agruca-polsl/gbsc_clusters_functional_analysis

	GBSC-strict	GBSC-relaxed
weight	4	4
lifetime	10	20
max-gap-len	2	3
max-node-count	6	6
include-orphan-nodes	no	yes

Table 1. *Strict* and *relaxed* parameter sets.

direction. If any disappearing edge has a weight exceeding the threshold, all edges in the current graph with weights above this threshold contribute to the graph representation of the STR. The boundaries of this STR in the sequence are determined by the first and last occurrence of nodes in the graph. Consequently, the method identifies STRs in protein fragments and provides a graph representation for them.

Identification - parameters

GBSC has several parameters that can be adjusted to identify STRs of interest. The *weight* threshold determines how many times an edge must be visited to mark corresponding residues as repeatable. Therefore, it determines a minimal number of repetitions of a pattern. The *lifetime* threshold affects both a maximal repeat pattern length and a maximal insertion length between pattern occurrences. However, for the same parameter value, different insertion lengths are accepted for different lengths of the repeating pattern. For instance, with the lifetime threshold of 4, the maximal insertion length between QA runs is 2 (e.g. QAQAQAEDQAQAQA), and the maximal length of a repeating pattern without such gaps is 4. To better manage these aspects, we introduced additional parameters to manage a maximal gap (*max-gap-len*) and pattern lengths in STRs (*max-node-count*), separately. Another parameter, *min-node-count*, allows researchers to filter out homopolymers, as they are often well-studied. For increased flexibility in detecting mutations, the *include-orphan-nodes* parameter can be utilized. This parameter enables the identification of single-node repetitions, which can be particularly useful for finding motifs such as KKX, essential for ribosome biogenesis.

In this article, we utilize two distinct sets of parameters, referred to as *strict* and *relaxed*. The *strict* parameter set identifies exact repeats and those with low mutation content, while *relaxed* set accepts more mutations in between repeats. The parameters for both sets are detailed in Table 1. In the subsequent sections, we refer to GBSC with *strict* and *relaxed* parameters as GBSC-strict and GBSC-relaxed.

Identification - example

Figure 1 illustrates how the GBSC method identifies RG repeats with a methionine insertion. In this example, we set the *lifetime* threshold to 5 and the *weight* threshold to 3. These repeats are derived from the coilin protein (P38432) and are part of the RG box, which is essential for interaction with SMN [Hebert et al., 2001]. The example presents the RG repeats along with surrounding residues to provide a comprehensive view of how GBSC identifies STRs. Panel **(A)** shows the state after the second iteration, where the method has already created two nodes from the first two dimers of the sequence. The numbers above the nodes indicate their lifetime, while the numbers above the edges represent their weight. Panel **(B)** displays the step where the method approached the RG repeats, encounters a single methionine insertion, and then returns to the RG repeats. Notably, nodes from previous steps that reached the

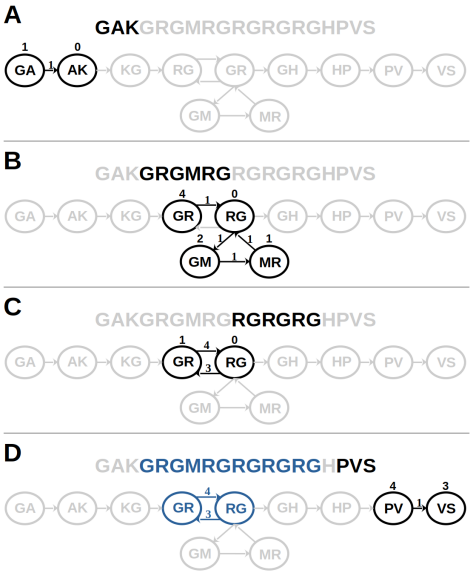


Fig. 1. Example identification of RG repeats with a methionine insertion. **(A)** shows algorithm state after processing two k-mers. In **(B)** the method handles insertion of a methionine. In **(C)** it increases weights in edges already visited. **(D)** shows how GBSC marks a repetitive region and selects its graph representation after leaving it. In this example, the *lifetime* threshold was set to 5, and the *weight* threshold was set to 3.

lifetime threshold have disappeared. In the final step of this panel, the method revisits the existing RG node, resetting its lifetime. Panel **(C)** demonstrates how the edge weights increase as the method processes repetitions by revisiting them. In **(D)** scanning did not recognize any new repeats on its way. Overall, nodes with connected edges that have weights equal to or greater than the weight threshold are marked as repetitive, along with the associated sequence fragment.

Clustering

GBSC utilizes the graph representation of identified sequences to assign them into relevant clusters. Traditional clustering methods, such as Markov Cluster (MCL), are designed and optimized for full-length protein sequences, often requiring substantial computational power and memory to calculate and store statistical similarities between these sequences [Enright et al., 2002]. Other tools attempt to mitigate this issue by minimizing the number of sequence comparisons. For instance, CD-HIT clusters sequences by comparing them only to representative sequences, assuming that remaining sequences are also similar to the each other [Li and Godzik, 2006]. However, this assumption can fail when comparing adjacent STRs of different types, leading to the inappropriate mixing of STRs in a single cluster [Jarnot et al., 2022]. In contrast, GBSC assigns STRs to clusters based on their graph representation established during the identification process. This method eliminates the need for a sequence comparison step, allowing sequences to be saved directly into their corresponding cluster files without occupying memory space. As a result, GBSC can efficiently manage large datasets, such as UniProtKB. Moreover, GBSC effectively addresses the challenge of adjacent STRs, where two or more STRs of different patterns are neighbors. The location of STRs within the sequence is known, enabling neighboring STRs to be assigned to clusters based on their types as well as to clusters representing their

combinations. For example, if a poly-Q STR is adjacent to a QA run STR, they are assigned to three distinct clusters: a cluster with poly-Q sequences, a cluster with QA runs, and finally a cluster containing adjacent STRs of poly-Q and QA runs in the same order as they appeared in the sequence. This approach is unique compared to most protein sequence clustering methods, which typically assign each sequence to only a single cluster.

Alphabet reduction

In GBSC, sequences can be identified and clustered using a reduced alphabet. Certain amino acids in protein sequences share similar physicochemical properties, allowing them to be grouped into categories of similar residues. Researchers have defined several alphabet reductions, demonstrating their effectiveness in uncovering various aspects of similarity between proteins [Liang et al., 2022, Steinegger and Söding, 2017]. The GBSC method supports alphabet reduction passed as a parameter, which is applied when creating a graph representation of STRs. As a result, the identified and clustered sequences remain unchanged. This capability expands the range of STR repeats that GBSC can analyze. For instance, if we reduce the alphabet to three residual groups, which are glycine, proline and others, then we can find collagen sequences with prolines at different positions in the triplet (GPX, GXP and GPP).

Analysis – clustering

GBSC is designed to cluster STRs according to similar, repeatable patterns, while state-of-the-art methods assess similarity between sequences using statistical approaches. To demonstrate the differences between these methodologies, we compared GBSC with CD-HIT v4.6 and MMseqs v15-6f452. The parameters used for STRs are detailed in Table 1. For all methods, we utilized STRs identified by GBSC in the UniProtKB/Swiss-Prot database version 2022_05 to focus solely on clustering solutions. We compared them quantitatively, providing statistics that describe the clusters and a Venn diagram showing the overlap between methods' results. Since overlap cannot be directly derived from clusters, we paired all similar sequences by combining all possible combinations in each cluster. A pair of similar sequences was considered overlapping with other methods if both sequences belonged to the same cluster in those methods. From each section of the diagram, we selected representative examples to demonstrate the types of similarity detected between the STRs by each method. Additionally, we summarized the number of clusters containing adjacent STRs and the total number of such sequences to illustrate the impact of GBSC's solutions. Figure 2 presents the workflow of the clustering analysis.

Analysis – functional evaluation of clustering results with the **C** measure

The GBSC method can be executed with various parameter settings, which can group functionally important sequences differently. Therefore, we are interested in finding a set of parameters that would allow the generation of the most functionally informative set of clusters as described by Gene Ontology. Such a set should include a high number of clusters that contain STRs from functionally annotated protein sequences, as the known annotations allow us to hypothesize about the functions of other unannotated sequences within the same cluster by transferring protein-level annotation to higher-resolution knowledge about the STR itself (Figure 3).

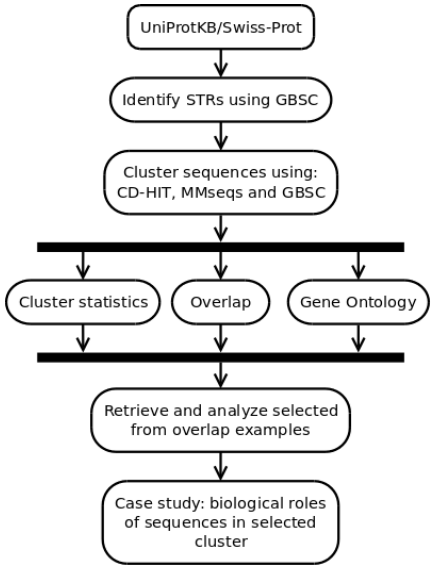


Fig. 2. The workflow used to compare clustering methods. GBSC is used only to identify STRs in UniProtKB/Swiss-Prot database. Next CD-HIT, MMseqs and GBSC are used to create clusters of similar sequences which are further analysed. Clusters are firstly analysed using exploratory approach, then by examining composition of selected examples and finally by looking at their biological roles.

The newly annotated STRs form high-confidence candidates for experimental validation. In addition, proteins with no unannotated function in the same STR cluster provide further candidates for functional validation. To streamline the analysis, we designed the simple but informative C measure that allows to find the best set of GBSC parameters that results in a high number of functionally annotated clusters. While evaluating the clustering results, the function takes into account the number of clusters annotated with at least one statistically significant GO term in the cluster, calculated with the hypergeometric test with the Benjamini - Hochberg multiplle FDR correction procedure [Benjamini and Hochberg, 1995] and normalized by the number of clusters.

In addition, we also provide the s measure that can be used to identify the most functionally concise clusters in the collection of clusters obtained using GBSC. This function takes into account the number of sequences in clusters that are annotated with the same, most frequent statically significant GO term in the cluster. We provide the descriptions the C measure and the s measure in the Supplementary Material S2. For the C measure, we also provide a detailed analysis of how the different GBSC parameter values influence its values, the length of sequences within clusters, and the sizes of the clusters.

In order to identify an interesting case study, we analysed the STRs in clusters obtained from GBSC analysis for all 216 parameter combinations, as described in Supplementary Material S2. After a thorough analysis, one type of cluster with the repetitive pattern RG/GR caught our attention. It consistently appeared in the results with a statistically significant Gene Ontology term assigned and high percentage of proteins with assigned Gene Ontology terms. From these clusters, we selected the one with the highest value of the s measure and we analysed functions of its STRs.

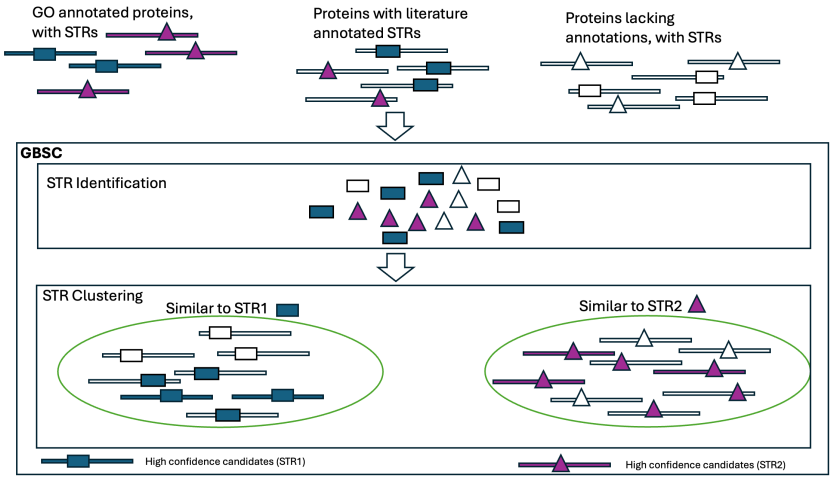


Fig. 3. A workflow presenting the design of the Gene Ontology analysis. Proteins containing STRs and annotated with GO terms (functional annotation across the full protein sequence) that cluster with proteins with similar STRs, for which functional annotation is available from the literature, represent high-confidence candidates for experimental validation.

Results

Statistics

As shown in Table 2, the clustering statistics for GBSC exhibited different trends compared to those of statistical-based methods when the parameters for STR identification were relaxed. MMseqs generated the highest number of clusters; however, many of these were singleton clusters, containing only one sequence. Interestingly, CD-HIT produced fewer clusters than GBSC for strict parameters but significantly more for relaxed parameters. All three methods resulted in an increase in singleton clusters when using relaxed GBSC identification parameters. Despite this, GBSC, on average, formed larger clusters with a greater standard deviation for relaxed parameters, while the other methods produced smaller clusters with lower standard deviations. The average number of sequences per cluster was highest for the CD-HIT method when using the strict set of parameters. However, for relaxed parameters, GBSC yielded the largest clusters on average. The standard deviations for the strict parameters were similar for GBSC and CD-HIT, in contrast to MMseqs. For relaxed parameters, the standard deviations for CD-HIT and MMseqs decreased significantly, while GBSC's standard deviation increased slightly. In summary, for GBSC, all analyzed metric values increased with relaxation of the parameters, whereas for CD-HIT and MMseqs, the number of clusters increased, but the average number of sequences and the standard deviation decreased. The clustering parameters from the MMseqs and CD-HIT methods are available in the Supplementary Material S3.

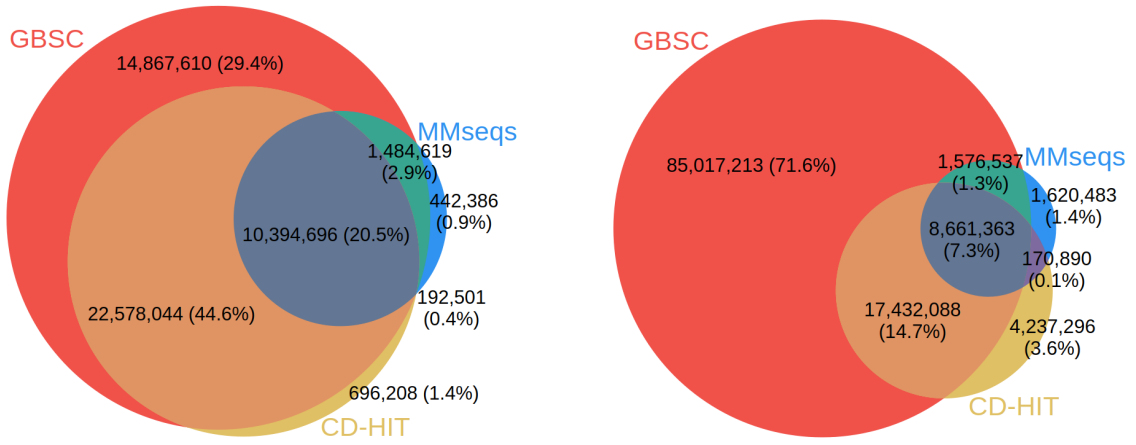
Similar pairs overlap

Clusters formed by protein sequence similarity analysis methods should contain sequences that are similar to each other. Therefore, we if we pick two randomly selected sequences from the same cluster, then these sequences should be similar. To examine the overlap between the results of different methods, we combined all pairs of sequences within the same cluster and created Venn diagrams to illustrate how the selected methods intersect.

Method	STR detection params	Cluster count	Singleton cluster count	Mean seq. count per cluster	Std dev. of seq. count per cluster
GBSC	Strict	2,913	1,985	13.6	183.6
	Relaxed	6,268	4,078	15.3	189.1
MMseqs	Strict	9,215	7,774	4.2	52.0
	Relaxed	31,689	23,727	2.8	27.5
CD-HIT	Strict	1,979	894	19.7	184.0
	Relaxed	14,158	7,275	6.2	65.4

Table 2. Alignment based methods have different cluster tendencies than GBSC after relaxing parameters. The input dataset for the clustering methods was prepared by identifying STRs in UniProt/Swiss-Prot using two parameter sets: strict and relaxed.

Table 3 presents the number of similar sequence pairs generated by all three methods. GBSC groups sequences with similar repetitive patterns regardless of insertions and mismatches, whereas statistical-based methods take these factors into account. For strict parameters, GBSC detected 97.4% of all similar pairs identified by all methods, with CD-HIT following at 66.8% of the total similar pairs. However, after relaxing the parameters, GBSC discovered over twice as many similarities, while CD-HIT clustered approximately 25.7% of total similarities. This increase in the number of similar pairs in the GBSC results is primarily due to large clusters containing homopolymers with various mutations. MMseqs found the fewest similar pairs, accounting for 24.7% and 10.1% for strict and relaxed parameters, respectively. The total number of similar pairs is increased for relaxed parameters compared to strict by 234%, to which mostly contribute GBSC. Specifically, GBSC results increased by 228%, while the other methods found fewer similar pairs, indicating a divergence in the results among the methods. For the strict set of parameters, all methods identified over 50 million similar pairs, while this number exceeded 118 million for the relaxed set. Overall, GBSC found significantly more similar pairs after relaxing the parameters, whereas the other methods identified fewer. It is important to note here that in the case of GBSC, a single sequence motif can be assigned to multiple clusters, a feature not available in the other methods.



(A) Input sequences: GBSC-strict (B) Input sequences: GBSC-relaxed

Fig. 4. Venn diagram showing overlap among GBSC, CD-HIT and MMseqs when clustering STRs identified by GBSC. (A) shows results identified with GBSC-strict, while (B) shows results for GBSC-relaxed.

	Strict STRs	Relaxed STRs
GBSC	49,324,969 (97.4%)	112,687,201 (94.9%)
MMseqs	12,514,202 (24.7%)	12,029,273 (10.1%)
CD-HIT	33,861,449 (66.8%)	30,501,637 (25.7%)
Total	50,656,064 (100%)	118,715,870 (100%)

Table 3. GBSC found more pairs of similar sequences after relaxing parameters while rest of the methods less.

Figure 4 shows Venn diagrams for STRs identified by GBSC-strict and GBSC-relaxed. For strict parameters, CD-HIT and MMseqs shared most of the results with GBSC. The unique results we collected using these methods represented only a small part of the total results, which were 0.9% and 1.4%, respectively. CD-HIT, which generated more similar pairs than MMseqs, covered approximately 84.6% of the MMseqs results, which means that these methods aligned STRs differently, producing unique results in almost exact repeats identified by GBSC-strict. GBSC had a high overlap with the other methods, and 68% of all similar pairs were found by this method with CD-HIT, MMseqs or both. 29.4% of all similar pairs were unique to GBSC, and only 2.6% were not detected by GBSC. For this set of parameters, we found 403 clusters containing 695 adjacent STRs, which partially characterize the red area of the diagram.

With relaxed parameters, the Venn diagram results showed significantly greater differences compared to those obtained with strict parameters. As previously noted, GBSC identified more similar pairs under relaxed conditions, while the other methods found fewer. Consequently, GBSC uncovered a much larger number of similar sequence pairs unique to this method. Specifically, 71.6% of the total similar pairs were identified solely by GBSC. This is justified by the fact that allowing more mutations within and between repetitions leads to more mismatches in alignments. GBSC shared 23.3% of similar pairs with the other methods, while 5.1% were found exclusively by the other methods. For this set of parameters, CD-HIT covered a smaller fraction of MMseqs results, that was 73.4%. The unique similar sequence pairs identified by CD-HIT and MMseqs represented 1.4% and 3.6% of all similar pairs, respectively, which was an increase compared to the strict parameter set. For relaxed parameters, we found 1067 cluster containing 8676 adjacent STRs.

Selected examples

Figure 5 presents sample sequences for all regions of the Venn diagram shown in Figure 4 **(A)**. Panels **(A)** and **(B)** of Figure 5 illustrate example pairs of potentially similar sequences clustered by CD-HIT and MMseqs. A closer examination of those cases reveals that both pairs were incorrectly clustered according to the repeating pattern of which they consist. The sequence pairs clustered by CD-HIT include an STR of alanine and proline alongside a poly-A fragment. Although both sequences are 50% identical, resulting in a high alignment score, this does not accurately reflect their similarity in terms of repeating patterns. A similar situation is observed with the two sequences in the MMseqs example, where the first sequence is an STR of the HD pattern, while the second consists of DY repeats.

Panel **(C)** shows two example sequence pairs containing repeating patterns according to GBSC. The first pair consists of two sequences made up of adjacent STRs, with the first part containing NS repeats and the second part being poly-N. These sequences differ in both length and the number of repeats of each type. The second example consists of sequences containing GN repeats. The first sequence contain a small number of mutations, while the second is highly degenerate and contain various mutations. Additionally, these two sequences also vary in length.

When GBSC overlaps with any of the other methods, the accuracy of STRs improves. Panels **(D)** and **(E)** illustrate examples where GBSC overlaps with CD-HIT and MMseqs, respectively. In the first case, where GBSC overlaps with CD-HIT, the sequences consist of varying numbers of SR runs. Both sequences contain several different mutations, but the predominance of exact repetitions enables both to be identified as similar by GBSC and the other methods. In the example where GBSC overlaps with MMseqs, the two sequences are composed of PPQGY, differing in the number of prolines that are degenerated at the C-terminal. These sequences contain similar repeats and can be well aligned.

The sequences presented in Panel **(F)** were interpreted differently by GBSC compared to CD-HIT and MMseqs. In the first sequence, GBSC identified only DR repeats, while in the second sequence, it recognized a combination of DR and ER repeats. As a result, GBSC assigned both sequences to

(A) STRs clustered by CD-HIT only
P1: AP-PA_PA-AP
S1: APAPAPAPSAPAP
S2: AAAAAAAAAAAA
P2: AA-AA

(B) STRs clustered by MMseqs only
P1: DH-HD_HD-DH
S1: HDHDHDHDHDHDHDHDHDHDHDHDH
S2: DYDYDYDYDYDYD
P2: DY-YD_YD-DY

(C) STRs clustered by GBSC only
P1: NS-SN_SN-NS~NN-NN
S1: NSNNNSNNNSNSNSNSNSNNSNNNNINNNNSNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNN
S2: SNSNSNSNSNSNSNSNSNSNSNSNSNSNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
P2: NS-SN_SN-NS~NN-NN

P1: GN-NG_NG-GN
S1: GNGNSGNGNGNGASNGNG
S2: GNGYSNGNGYTNGNGHTNGNGNYNGNGHVNGNG
P2: GN-NG_NG-GN

(D) STRs clustered by GBSC and CD-HIT
P1: RS-SR_SR-RS
S1: RSRSRSRSRSRSRSRSRYSRSKSRSRTRSRSRS
S2: RSRSRSWSRSRSRSRRYSRSRSR
P2: RS-SR_SR-RS

(E) STRs clustered by GBSC and MMseqs
P1: GY-YP_PP-PQ_PQ-QG_QG-GY_YP-PP
S1: PPQGYAPPPQGYPPQGYPPQGYPPQGYPPQGYPPPPQGAPPQG
S2: YPPQGYPPQGYPPPPPQGYPPQGYPPQGYPPQGYPPPPQGPPP
QGPPPQAAPPQ
P2: GY-YP_PP-PQ_PQ-QG_QG-GY_YP-PP

(F) STRs clustered by CD-HIT and MMseqs
P1: DR-RD_RD-DR
S1: DRDRDRDRDKERDRDRDRDRERDR
S2: DRDRERDRDRERDRDRDRERERTRERERER
P2: DR-RD_RD-DR_RE-ER

(G) STRs clustered by all methods
P1: DE-EY_ED-DE_LE-ED
S1: LEDEYRTLEDEYETLEDEYGILEDEY
S2: LEDEYRTLEDEYETLEDEYGILEDEY
P2: DE-EY_ED-DE_LE-ED

Fig. 5. Selected examples that characterise all areas of the Venn diagram for strict parameters of GBSC. Each example is described by **(P1)** GBSC model assigned to the first sequence, **(S1)** sequence of the first region, **(S2)** sequence of the second region, and **(P2)** GBSC model assigned to the second sequence.

different clusters. In contrast, CD-HIT and MMseqs classified both sequences as similar. Glutamic acid, which is abundant in the second sequence but rare in the first, was partially aligned and partially treated as mutations in the second sequence.

Panel **(G)** contains a pair of identical sequences. It is important to highlight that most pairs found at the intersection of all methods are either identical or contain only a small number of mutations.

Biological interpretation of selected cluster

In order to identify an interesting case study, we analysed the STRs in clusters obtained from GBSC analysis. After a thorough analysis, one cluster caught our attention. In all of the 216 parameter combinations (see Supplementary Material S4), this RG/GR cluster appeared in the results and had GO terms assigned. Detailed descriptions of proteins with assigned GO terms and motif RG/GR are provided in Supplementary Material S5. We analyzed the group of 51 STRs from fungi, plant and animal kingdoms characterised by the RG/GR motif. These are known RNA binding regions, named GAR domains, RGG boxes or RG domains [Burd and Dreyfuss, 1994]. The list of protein names obtained using the following parameters; w=5, g=3, l=10, x=4, is available in Supplementary Material S5. Among those are well known RNA binders, like U3 small nucleolar RNA-associated protein 25, interleukin enhancer-binding factor 3 or 40S ribosomal protein S2. There are also DNA binders, such as the transcription regulatory protein SNF2 and methyl-CpG-binding domain protein 2. All STRs from this cluster have very well established RNA/DNA binding capabilities. Supplementary Material S5 shows UniProt accession numbers and their STRs annotated as RNA/DNA binders either by sequence homology or by known functions according to the QuickGO database[Huntley et al., 2009]. There are 18 protein STRs in this set that are known to bind DNA and 40 RNA binders. Out of 51 proteins, only 7 have no GO annotation. These are: A0A1D9BZF0 (germ cell nuclear acidic protein), A7TQN2 (SQS1), M9PBE2 (E3 ubiquitin-protein ligase Hakai), P12978 (Epstein-Barr nuclear antigen 2), Q3KSV2 (Epstein-Barr nuclear antigen 2), Q29EQ3 (PHD finger protein rhinoceros) and Q7YZH1 (PHD finger protein rhinoceros). This group of proteins has a common feature; all of them are known to bind proteins that bind RNA or DNA, but there is no evidence that they bind DNA/RNA in the RG/GR region[Ribeiro and Crossan, 2023, Pandit et al., 2009, Bawankar et al., 2021, Glaser et al., 2017, Hsieh et al., 1999, Kelleher, 2021, Steele et al., 2009, Andersen et al., 2017]. Our results suggest that, apart from ability to bind proteins, all those 7 proteins possess GR repeats known to bind nucleic acids, making them high-confidence candidates for further experimental validation.

Discussion

In this work, we introduced and analyzed a novel STR clustering method that utilizes the graph representation of STRs created during the identification process. This method focuses on clustering STRs based on their repetitive pattern, ignoring information about insertions and mutations between these repeats which makes it is suitable for degenerate STRs. We compared this method with other approaches for both STR identification and protein sequence clustering. Subsequently, we conducted exploratory data analysis to investigate the results and examined selected cases in detail. Additionally, we performed Gene Ontology analysis to assess the functional composition of the clusters. Finally, we analyzed a specific cluster and discussed its biological relevance.

GBSC is able to cluster STRs composed of adjacent types, as it is aware about locations of these motifs. Adjacent STRs can be critical for the functional and structural properties of proteins. For instance, a poly-P fragment inhibits β-Sheet structure in neighbouring poly-Q fragment [Darnell et al., 2007]. However, currently available tools for protein

sequence clustering struggle to manage these adjacent STRs effectively [Jarnot et al., 2022]. Methods like MMseqs and CD-HIT, which focus on whole protein sequences, skip the information about STR locations, making it impossible for them to cluster adjacent STRs properly. In contrast, GBSC not only creates separate clusters for adjacent STRs but also appends their subtypes to their corresponding clusters.

GBSC clusters sequences based on repeating patterns while skipping information about mutations between them. Repeats in proteins can exhibit various mutations and insertions. A well-known example of this is collagen, which requires glycine at consistent intervals and permits nearly random residues in between [Naomi et al., 2021]. However, statistical-based methods consider these insertions when calculating the final score, even allowing for changes in repetitive patterns [Jarnot, 2024]. As a result, fragments lacking relevant STRs may be aligned with a higher alignment score. By skipping these insertions when determining clusters for STRs, GBSC proves to be more suitable for such cases. A similar approach is employed by XSTREAM and T-REKS, where these repetitions serve as seeds for longer repeats [Jorda and Kajava, 2009, Newman and Cooper, 2007]. However, these methods are designed only for the identification of STRs, not for clustering.

The sequence models produced by GBSC are interpretable, as GBSC tags identified and clustered sequences using a graph-based model of pattern that is interpretable and describes the sequences. Some tools for identifying regions with biased compositions lack descriptive information about the identified patterns which can be crucial for selecting the appropriate algorithm for a given task, even if a non-interpretable model has better performance [Linardatos et al., 2020]. For example, T-REKS is a method suitable for exact and highly degenerate tandem repeats. However, it does not provide a consensus on what constitutes repeatability, making it challenging to understand which seeds were used during identification, and therefore it is unable to cluster TRs based on identified regions. Therefore, our method may be the optimal choice when interpretability of similarity between STRs is a critical factor.

Finally, we have developed the C measure, an evaluation function to assess the functional composition of the GBSC clusters. We have also analyzed the influence of the GBSC parameters for the C measure values identifying the optimal parameters for the Swiss-Prot database. The function uses annotation of Gene Ontology to protein sequences, a common approach in functional analysis of the gene or protein groups. However, in the case of STRs, such approach can introduce errors, as those regions may not be associated with the Gene Ontology functions as they are usually assigned to the whole protein sequences. In addition to the quantitative analysis provided in the Supplementary Material S5 we also presented a qualitative analysis of the cluster with an RG/GR motif that has been identified using the C measure. Our case analysis clearly shows previously missed functions of a set of proteins. We were able to hypothesise on it thanks to their presence in an RG/GR cluster with all other members being known to bind RNA or DNA.

In conclusion, it is well known that whole-protein sequence based clustering is not suitable for an analysis of low-complexity regions. More than half of all proteins, however, have low-complexity regions, with such regions even more prominent in 'modern' proteins involved in development and the neuronal functions. To address this, we have developed Graph-Based Sequence Clustering (GBSC) a new method to identify and cluster short tandem repeats (STRs), an important class of low-complexity regions. GBSC was developed with the purpose of analysing in large protein datasets to functionally characterize STRs. The graph-based method supports the clustering of imperfect repeats as well as the clustering and interpretation of multiple adjacent STRs as single functional patterns. Notably, the constructed clusters comprise fragments with similar protein-level annotations. We have demonstrated the utility of this novel approach for inference of the functions of the STRs themselves on a cluster of over 50 RG-repeat proteins, allowing us to newly link RNA/DNA-binding function with multiple STRs with high confidence.

Competing interests

No competing interest is declared.

Acknowledgments

We thank David P. Kreil for valuable comments on this manuscript.

Funding

This work was supported by National Science Centre [2020/39/B/ST6/03447 to P.J., J.Z., M.G., and A.G.] and COST-Action [CA21160 to J.Z., M.G., V.P., and A.G.]

References

A. Abnousi, S. L. Broschat, and A. Kalyanaraman. Alignment-free clustering of large data sets of unannotated protein conserved regions using minhashing. *BMC Bioinformatics*, 19(1):83, 3 2018.

M Mar Albà, Roman A Laskowski, and John M Hancock. Detecting cryptically simple protein sequences using the simple algorithm. *Bioinformatics*, 18(5):672–678, 2002.

Peter Refsing Andersen, Laszlo Tirian, Milica Vunjak, and Julius Brennecke. A heterochromatin-dependent transcription machinery drives pirna expression. *Nature*, 549 (7670):54–59, 09 2017. ISSN 1476-4687.

P. Bawankar, T. Lence, C. Paolantoni, I. U. Haussmann, M. Kazlauskiene, D. Jacob, J. B. Heidelberger, F. M. Richter, M. P. Nallasivan, V. Morin, N. Kreim, P. Beli, M. Helm, M. Jinek, M. Soller, and J. Y. Roignant. Hakai is required for stabilization of core components of the m6a mrna methylation machinery. *Nature Communications*, 12 (1):3778, Jun 2021.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Christopher G. Burd and Gideon Dreyfuss. Conserved structures and diversity of functions of rna-binding proteins. *Science*, 265(5172):615–621, 1994.

Sean M Cascarina and Eric D Ross. The lcd-composer webserver: high-specificity identification and functional analysis of low-complexity domains in proteins. *Bioinformatics*, 38(24):5446–5448, 2022.

UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1): D523–D531, 2023.

Gregory Darnell, Joseph PRO Orgel, Reinhard Pahl, and Stephen C Meredith. Flanking polyproline sequences inhibit β-sheet structure in polyglutamine segments by inducing ppii-like helix structure. *Journal of molecular biology*, 374

(3):688–704, 2007.

Robert C. Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 10 2010.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2021.

Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575–1584, 2002.

Roman Feldbauer, Lukas Gosch, Lukas Lüftinger, Patrick Hyden, Arthur Flexer, and Thomas Rattei. Deepnog: fast and accurate protein orthologous group assignment. *Bioinformatics*, 36(22-23):5304–5312, 2020.

Laura V. Glaser, Simone Rieger, Sybille Thumann, Sophie Beer, Cornelia Kuklik-Roos, Dietmar E. Martin, Kerstin C. Maier, Marie L. Harth-Hertle, Björn Grüning, Rolf Backofen, Stefan Krebs, Helmut Blum, Ralf Zimmer, Florian Erhard, and Bettina Kempkes. EBF1 binds to EBNA2 and promotes the assembly of ebna2 chromatin complexes in b cells. *PLoS Pathog*, 13:e1006664, 2017.

Paul M Harrison. flps 2.0: rapid annotation of compositionally-biased regions in biological sequences. *PeerJ*, 9:e12363, 2021.

L. He, Y. Li, R. L. He, and S. S. Yau. A novel alignment-free vector method to cluster protein sequences. *Journal of Theoretical Biology*, 427:41–52, 8 2017.

Michael D Hebert, Piotr W Szymczyk, Karl B Shpargel, and A Gregory Matera. Coilin forms the bridge between cajal bodies and smn, the spinal muscular atrophy protein. *Genes & development*, 15(20):2720–2729, 2001.

James J.-D. Hsieh, Sifang Zhou, Lin Chen, David B. Young, and S. Diane Hayward. CIR, a corepressor linking the DNA binding factor CBF1 to the histone deacetylase complex. *Proceedings of the National Academy of Sciences*, 96(1):23–28, 1999.

Rachael P Huntley, David Binns, Emily Dimmer, Daniel Barrell, Claire O'Donovan, and Rolf Apweiler. Quickgo: a user tutorial for the web-based gene ontology browser. *Database*, 2009:bap010, 2009.

Patryk Jarnot. Challenges in adjusting scoring matrices when comparing functional motifs with non-standard compositions. *Scientific Reports*, 14(1):31777, 2024.

Patryk Jarnot, Joanna Ziemska-Legiecka, Marcin Grynberg, and Aleksandra Gruca. Insights from analyses of low complexity regions with canonical methods for protein sequence comparison. *Briefings in Bioinformatics*, 23(5):bbac299, 2022.

Julien Jorda and Andrey V Kajava. T-reks: identification of tandem repeats in sequences with a k-means based algorithm. *Bioinformatics*, 25(20):2632–2638, 2009.

Mohamed Kamel, Kristina Kastano, Pablo Mier, and Miguel A Andrade-Navarro. Rep2: A web server to detect common tandem repeats in protein sequences. *Journal of Molecular Biology*, 433(11):166895, 2021.

Erin S Kelleher. Protein–protein interactions shape genomic autoimmunity in the adaptively evolving rhino-deadlock-cutoff complex. *Genome Biology and Evolution*, 13(7):evab132, 06 2021. ISSN 1759-6653.

Gal Kesten-Pomeranz, Yaniv Nikankin, Anja Reusch, Tomer Tsaban, Ora Schueler-Furman, and Yonatan Belinkov. Induction meets biology: Mechanisms of repeat detection in protein language models, 2026.

Vojtěch Kubáň, Pavel Srb, Hana Štégnerová, Petr Padrta, Milan Zachrdla, Zuzana Jaseňáková, Hana Šanderová, Dragana Vítovská, Libor Krasny, Tomáš Koval', et al. Quantitative conformational analysis of functionally important electrostatic interactions in the intrinsically disordered region of delta subunit of bacterial rna polymerase. *Journal of the American Chemical Society*, 141(42):16817–16828, 2019.

Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

Xuehui Li and Tamer Kahveci. A novel algorithm for identifying low-complexity regions in a protein sequence. *Bioinformatics*, 22(24):2980–2987, 10 2006. ISSN 1367-4803.

Yuchao Liang, Siqi Yang, Lei Zheng, Hao Wang, Jian Zhou, Shenghui Huang, Lei Yang, and Yongchun Zuo. Research progress of reduced amino acid alphabets in protein analysis and prediction. *Computational and Structural Biotechnology Journal*, 20:3503–3510, 2022.

Wei-Hsiang Lin and Edo Kussell. Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. *Nucleic acids research*, 40(6):2399–2413, 2012.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

Camille Moeckel, Manvita Mareboina, Maxwell A. Konnaris, Candace S.Y. Chan, Ioannis Mouratidis, Austin Montgomery, Nikol Chantzi, Georgios A. Pavlopoulos, and Ilias Georgakopoulos-Soares. A survey of k-mer methods and applications in bioinformatics. *Computational and Structural Biotechnology Journal*, 23:2289–2303, 2024. ISSN 2001-0370.

Ruth Naomi, Pauzi Muhd Ridzuan, and Hasnah Bahari. Current insights into collagen type i. *Polymers*, 13(16):2642, 2021.

Aaron M Newman and James B Cooper. Xstream: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC bioinformatics*, 8(1):1–19, 2007.

Shatakshi Pandit, Sudakshina Paul, Li Zhang, Min Chen, Nicole Durbin, Susan M W Harrison, and Brian C Rymond. Spp382p interacts with multiple yeast splicing factors, including possible regulators of prp43 dexd/h-box protein function. *Genetics*, 183(1):195–206, 09 2009. ISSN 1943-2631.

Joana Pereira, Lorenzo Pantolini, Janani Durairaj, and Torsten Schwede. Large-scale protein clustering in the age of deep learning. *Current Opinion in Structural Biology*, 94:103078, 2025. ISSN 0959-440X. doi: 10.1016/j.sbi.2025.103078.

Vasilis J Promponas, Anton J Enright, Sophia Tsoka, David P Kreil, Christophe Leroy, Stavros Hamodrakas, Chris Sander, and Christos A Ouzounis. Cast: an iterative algorithm for the complexity analysis of sequence tracts. *Bioinformatics*, 16(10):915–922, 2000.

Jonathan Ribeiro and Gerry P Crossan. Gcna is a histone binding protein required for spermatogonial stem cell maintenance. *Nucleic Acids Research*, 51(10):4791–4813, 03

2023. ISSN 0305-1048.

Christian JA Sigrist, Edouard De Castro, Lorenzo Cerutti, Béatrice A Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. New and continuing developments at prosite. *Nucleic acids research*, 41(D1): D344–D347, 2012.

Latishya Steele, Madina J. Sukhanova, Jinhua Xu, Gabriel M. Gordon, Yongsheng Huang, Long Yu, and Wei Du. Retinoblastoma family protein promotes normal r8-photoreceptor differentiation in the absence of rhinoceros by inhibiting de2f1 activity. *Developmental Biology*, 335(1): 228–236, 2009. ISSN 0012-1606.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31 (6):926–932, 2015.

Alina C Teuscher, Elisabeth Jongsma, Martin N Davis, Cyril Statzer, Jan M Gebauer, Alexandra Naba, and Collin Y Ewald. The in-silico characterization of the caenorhabditis elegans matrisome and proposal of a novel collagen classification. *Matrix Biology Plus*, 1:100001, 2019.

Ze-Gang Wei, Xu Chen, Xiao-Dan Zhang, Hao Zhang, Xing-Guo Fan, Hong-Yan Gao, Fei Liu, and Yu Qian. Comparison of methods for biological sequence clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(5):2874–2888, 2023.

John C Wootton and Scott Federhen. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & chemistry*, 17(2):149–163, 1993.