OXFORD

*Briefings in Bioinformatics*, 2022, **23(5)**, 1–13
https://doi.org/10.1093/bib/bbac299
**Problem Solving Protocol**

# Insights from analyses of low complexity regions with canonical methods for protein sequence comparison

Patryk Jarnot, Joanna Ziemska-Legiecka, Marcin Grynberg and Aleksandra Gruca

Corresponding author. Aleksandra Gruca. Department of Computer Networks and Systems, Silesian University of Technology, Akademicka 2A, 44-100 Gliwice, Poland. Tel.: +48322372171; E-mail: aleksandra.gruca@polsl.pl

## Abstract

Low complexity regions are fragments of protein sequences composed of only a few types of amino acids. These regions frequently occur in proteins and can play an important role in their functions. However, scientists are mainly focused on regions characterized by high diversity of amino acid composition. Similarity between regions of protein sequences frequently reflect functional similarity between them. In this article, we discuss strengths and weaknesses of the similarity analysis of low complexity regions using BLAST, HHblits and CD-HIT. These methods are considered to be the gold standard in protein similarity analysis and were designed for comparison of high complexity regions. However, we lack specialized methods that could be used to compare the similarity of low complexity regions. Therefore, we investigated the existing methods in order to understand how they can be applied to compare such regions. Our results are supported by exploratory study, discussion of amino acid composition and biological roles of selected examples. We show that existing methods need improvements to efficiently search for similar low complexity regions. We suggest features that have to be re-designed specifically for comparing low complexity regions: scoring matrix, multiple sequence alignment, e-value, local alignment and clustering based on a set of representative sequences. Results of this analysis can either be used to improve existing methods or to create new methods for the similarity analysis of low complexity regions.

**Keywords:** comparison methods; low complexity regions; protein sequence similarity

## Introduction

Protein sequences are composed of amino acid fragments of varying diversity. Fragments with low diversity in the amino acid composition are called low complexity regions (LCRs). Due to their frequent occurrence and capacity to expand through replication slippage, they can easily increase protein sequence space and contribute to novel protein functions. They are known to play a key role in protein functions and may be relevant to protein structure [1]. For example, prion-like LCRs are key regulators of protein solubility and folding [2]. Cytoplasmic human Gle1 is hyperphosphorylated in a low-complexity domain in response to stress [3]. A known LCR motif RGG/RG is generally required for RNA binding and phase separation [4]. LCRs may also form labile cross-$\beta$ polymers and hydrogel droplets [5].

Methods and algorithms for searching for similarities among protein sequences have always been important tools in biology that allowed researchers to predict protein functions from the sequence data alone. Many approaches are known from the literature that are suitable for searching for similar protein sequences. However, these methods are based on statistical models that are optimized to compare high complexity fragments. Due to that, for many years, protein regions that are characterized by low complexity of amino acids had been ignored and excluded from such type of analysis.

Recently, the research community became more interested in the so-called Dark Proteome that is mostly composed of intrinsically disordered proteins or proteins that contain intrinsically disordered regions [6–8]. Therefore, nowadays, it is crucial to revisit state-of-the art methods of protein sequence comparison in order to understand if and how they can be applied to analyse similarity of LCRs.

The community has already made some efforts to develop tools that are capable of automatically assigning functional roles of LCRs. One such example is the web server LCR hound [9] which identifies Uniprot-annotated prokaryotic LCR sequences that have the closest amino

**Patryk Jarnot** is a Research Assistant in the Department of Computer Networks and Systems, Silesian University of Technology. He develops methods for the analysis of protein domains with non-standard amino acid composition.
**Joanna Ziemska-Legiecka** is a PhD student at the Institute of Biochemistry and Biophysics PAS, where she studies low complexity regions of proteins.
**Marcin Grynberg** is an Associate Professor at the Institute of Biochemistry and Biophysics PAS, where he studies low complexity regions of proteins.
**Aleksandra Gruca** is an Associate Professor in the Department of Computer Networks and Systems, Silesian University of Technology. Her research focus on machine learning methods for protein sequence data analysis and application of predictive models for the integrated analysis of large-scale heterogeneous datasets.

**Received:** February 18, 2022. **Revised:** June 29, 2022. **Accepted:** July 1, 2022

acid or di-amino acid content in comparison to the query LCR sequence and based on this predicts a potential role of the query LCR. However, the prediction algorithm is using an amino acid content rather than sequence similarity, and therefore, we still lack of state-of-the-art methods for searching for similar LCRs in protein sequences.

Here, we present a study that compares the performance of three state-of-the art methods for searching for similarities among protein sequences: BLAST [10], HHblits [11] and CD-HIT [12]. By analysing how these methods perform in a task of searching for similar LCRs, we try to answer the following question: can these methods be applied to analyze LCRs or maybe new methods have to be invented? According to our best knowledge, scientists lack methods designed specifically for analyses of similarities among low complexity fragments of protein sequences. The aim of this study is also to raise awareness that the statistical models created for High Complexity Regions (HCRs) of proteins cannot be applied directly for a task of low complexity sequences comparison.

## Methods

The workflow of our experimental approach is shown in Figure 1. In the analysis, we need high-quality annotated data, and therefore, we used the UniProtKB/Swiss-Prot database (version: April of 2020) [13]. We identified LCRs and we divided all sequences into LCR and HCR parts. If a sequence had several LCRs, it was split into these different LCRs and the remaining HCR part of the sequence. Then, we created two datasets with sequences collected in the previous step for both LCRs and HCRs. At this point, the dataset contained amino acid sequences with simple annotations (UniProt AC and a name of a protein it belongs to). In the next step, we added information about protein families and analysed this set of sequences with BLAST, HHblits and CD-HIT tools. Then, we evaluated these methods using exploratory analysis, by looking at amino acid composition and biological role of selected results based on UniProtKB/Swiss-Prot functional annotations. To select interesting cases from BLAST and HHblits results, we filtered out results from the same families. We performed the entire process several times adjusting the parameters for each method to achieve the best results. To select similar sequences, we used e-value threshold equal to 0.0001 for BLAST and HHblits. The source code for the entire workflow is available for download (https://doi.org/10.5281/zenodo.6759535).

### LCR extraction

The definition of LCR in a protein sequence is not well specified. General agreement is that LCRs in proteins should have an excess of one or a few types of amino acid residues, but still, there is no consensus which metric is most appropriate. LCRs can be composed of homopolymers, short tandem repeats and irregular regions with low entropy. The scientific literature provides different terminology and definitions of these regions that are typically based on the sequence composition and periodicity [14]. CAST [15] and fLPS [16] are the methods capable of detecting compositionally biased regions (CBRs). While the usage of the terms LCR and CBR has been interchangeable in many contexts, use of one term or the other depends on the focus of the method used for their detection, i.e. sequence variability (LCRs) or amino acid composition (CBRs), respectively [14]. Another type of LCRs are short tandem repeats which are characterized by amino acid periodicity (i.e. repetitiveness). The example of methods that are designed specifically to discover short tandem repeats is XSTREAM [17] and T-REKS [18]. SIMPLE [19] is another method that allows to detect so-called cryptic repeats [19] which are regions of proteins containing overrepresentations of short amino acid repeats.

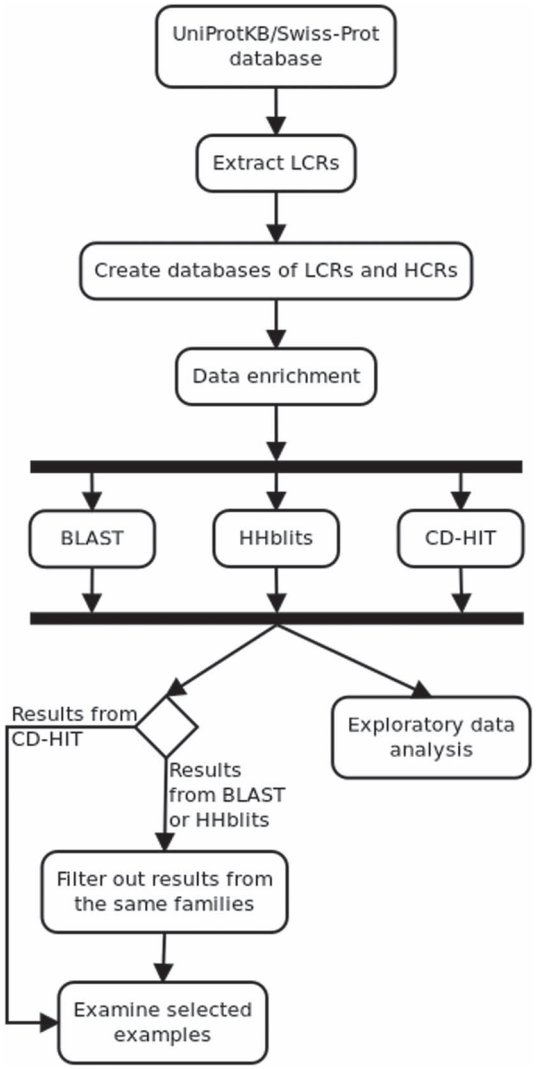


**Figure 1.** From the UniProtKB/Swiss-Prot database, we extracted LCRs and created two distinct datasets (HCRs and LCRs) enriched with family annotation that we analysed with BLAST, HHblits and CD-HIT. Then, we compared obtained similarity results and examined selected examples.

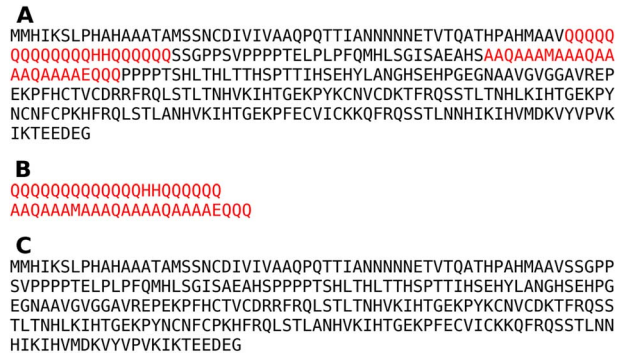


**Figure 2.** Division of the exemplary sequence (UniProtID: P39413) into a HCRs and LCRs. Panel (**A**) shows the whole sequence with identified LCRs highlighted in red. Panel (**B**) shows extracted LCRs and panel (**C**) presents all remaining non-LCR residues combined into an HCR part.

Here, we consider LCRs as sequences characterised by low entropy in amino acids composition. This assumption is based on the definition provided by authors of the SEG method which is one of the most popular tools for searching for LCRs [20].

To identify LCRs, we used SEG strict parameters (*K*1: 1.5, *K*2: 1.8, window: 15) which have been successfully employed in LCR analyses. Strict parameters of SEG ensure that identified regions are strongly compositionally biased while also allowing for a low amino acid diversity [21, 22]. The first dataset is for LCRs and the second one for HCRs. Each sequence that includes one or more LCRs is split into its corresponding LCRs, while the remaining residues are joined creating the HCR part of the sequence as presented in Figure 2. As a result, we found 26 333 LCRs in 16 418 proteins from which we created two distinct datasets. The number of HCRs is equal to the number of proteins in the database that is 562 252.

## Creation of databases for LCRs and HCRs

The main purpose of our research is to compare canonical sequence searching/clustering methods for LCRs; however, we decided to compare HCRs as a control experiment that proves correctness of our workflow.

## Data enrichment

In the first step, we enriched protein sequences with information about their protein families based on UniprotKB/Swiss-Prot annotations. We need this information to exclude from the analysis similar sequences that are derived from the same family. The rationale behind excluding these sequences from direct comparison was because we expected a high level of similarity of protein sequences within the family. However, in our analyses, we wanted to focus on non-obvious cases where HCRs from two proteins are different but LCRs are similar.

## Parameters of the methods

This section presents in details the parameters used for analysis and adjustment of their values for analysis of low complexity parts of the sequences as by default, the methods' parameters are optimised for high complexity parts (HCPs) of sequences.

### *BLAST*

BLAST is the first method we used to search for similar sequences. It uses Smith–Waterman algorithm to calculate local alignment based on a query and database sequences [10]. We converted fasta formatted datasets to BLAST-specific format.

Below, we enumerate modified parameters. We set e-value to 0.0001 and *max_target_seqs* to maximal possible value (1 073 741 798). Additionally, to improve searching for LCRs using available BLAST options, we changed *task* and *comp_based_stats* parameters. A study of selected BLAST parameter settings that can be applied for LCRs analysis can be found in [23]. *Task* option is responsible for the default parameter set adjusted for specific types of sequences. Possible options are: *blastp*, *blastp-fast* and *blastp-short*. *Blastp-short* makes the following modifications to the default options: sets scoring matrix to PAM30, sets gap opening cost to 9, sets gap extension cost to 1, sets word size to 2, clears filter options and changes e-value (which is not applicable in our case because we explicitly set it). We left the scoring matrix parameter unchanged with its default value of PAM30 since it is recommended for short sequences [24]. *Comp_based_stats* option is responsible for composition-based statistics. This option changes the scoring matrix by recalculating score values of frequently occurring amino acids in the query sequence, which is mainly caused by LCRs. It simply decreases the significance of LCRs while searching [25], and therefore, we turned it off.

### *HHblits*

HHblits is able to search for distantly related proteins that share a common ancestor. It is a part of the HH-Suite package and it uses Hidden Markov Model profiles to find similar sequences using HMM-HMM comparison [11]. Therefore, the query sequence is converted into an HMM profile. Database also stores HMM profiles which are condensed forms of multiple sequence alignments (MSAs) and represent protein sequences. In order to search for similar sequences, HHblits requires to create a database of profiles of Hidden Markov Models. We used uniclust-pipeline to create them for both LCRs and HCRs [26].

Uniclust-pipeline uses MMseqs to cluster similar sequences and to create their Hidden Markov Model profiles [26, 27]. We created two distinct datasets for both HCRs and LCRs. For HCRs, we used standard workflow, and for LCRs, we slightly modified it. To analyse LCRs with MMseqs tool, we changed two parameters. The first parameter is *mask* which is responsible for choosing a masking strategy and the second is *comp-bias-corr* which changes correction for locally biased composition of amino acids. While analysing LCRs, it is recommended to turn off both of these parameters by setting their values to 0. We also removed the *max-seqs* parameter as it is deprecated and not available in the newest version of MMseqs. The result of running uniclust-pipeline

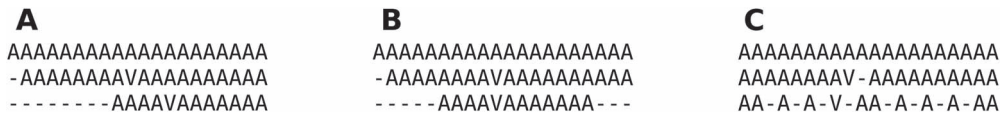


**Figure 3.** Possible ways of creating MSAs for LCRs collected from Q9V727, D3ZKD3 and Q5BGE2, respectively, that can generate different HMM profiles. These alignments were obtained using three different methods for MSA which are (**A**) MUSCLE, (**B**) Kalign and (**C**) Clustal Omega.

is several databases created with different identity thresholds: 10, 20 and 30%. From these results, we selected Uniboost30, a database with highest sequence identity that is equal to 30%. However, the threshold is still low, and therefore, the results obtained using this dataset contain more distant similarities.

We performed the analyses of both HCRs and LCRs using the same e-value as in the case of BLAST (0.0001). For LCR analysis, we modified the following parameters: *id*, *diff*, *norealign*, *sc* and *noprefilt*. These parameters control a number of results that are relevant to query sequences and their detailed analysis is provided in Supplementary material. *Id* parameter changes maximal pairwise sequence identity. We set this option to 100 (default is 90) because there are LCR families that are highly similar in their amino acid patterns. By default, HHblits selects the most diverse set of sequences, but here, we wanted to analyse the most similar ones, and therefore, we turned this setting off (*diff* parameter). Setting the *norealign* parameter disables MAC algorithm which significantly increases the number of matches [28]. *Sc* option changes the method which is used to calculate amino acid score. We examined all available parameters, and based on the empirical analyses, we chose the value of 0 which uses background frequencies. Default value for this parameter is optimized for compositional bias correction [29]. *noprefilt* parameter was introduced to speed up searches by filtering out cases that are 'obviously' not similar. However, we noticed that this parameter is filtering too many similar LCRs, so finally we disabled prefiltering.

HHblits database contains profiles of HMMs which are a condensed form of MSA. Here, we notice an issue related to the MSA creation. In Figure 3, we can see three protein alignments of three different subsequences aligned by three different tools: MUSCLE [30], Kalign [31] and Clustal Omega [32]. All of these sequences are homopolymers of alanine, but two of them have a mutation into valine. The problem is that MSA may be created in several ways. For example, in Figure 3A, the third sequence was aligned to others by inserting gaps at the beginning of the sequence, while in Figure 3B, valine is aligned properly. Figure 3C presents the actual alignment based on Clustal Omega pipeline to create profiles of HMMs for HHblits. Different ways to obtain MSA results in different profiles HMM which influences searching outcome, because a particular position will be scored differently for given amino acid.

### *CD-HIT*

The third method we analysed in this work is CD-HIT which uses a greedy incremental algorithm. In a nutshell, first all sequences are sorted by length in descendent order. Then, the longest sequence creates the first cluster. The sequence that creates a cluster is called a representative and all the following sequences are being aligned to it to determine whether they should be included into this cluster or not. If the similarity score of a particular aligned sequence is higher than a threshold, the sequence is added to the corresponding cluster. Otherwise, the sequence creates a new cluster and becomes its representative [33]. To speed up comparison time, short-word filtering and statistically based filtering were introduced [34]. We left default parameters unchanged for HCRs and we changed the minimal accepted length of sequences to its lowest value for LCRs, which is 4, since these regions are frequently short. We provide detailed analysis of this parameter in Supplementary material.

## Results analysis

We performed exploratory data analysis by investigating: (i) how the results from the selected methods overlap with each other, (ii) what is the average sequence similarity and (iii) what is the number of alignments for a given length? Finally, we present selected cases to show how BLAST, HHblits and CD-HIT analyse LCRs. We also analysed amino acid composition and biological roles of selected examples. To find more interesting cases, we filtered out results (hits) that include sequences from the same families for BLAST and HHblits. We assumed that minimal length of LCR is 6 amino acids.

To investigate biological features of the selected proteins, we performed a three layer functional analysis composed of the following stages: (i) We scanned popular DNA/protein databases (UniProt, RefSeq, STRING, Ensembl) [13, 35–37], (ii) Next we focused on the Pfam domain database [38] and (iii) the last stage was reading articles that mentioned specific proteins/protein families and their functional analyses.

# Results

## Quantitative results

To quantitatively compare the results from the three methods, we created Venn diagrams (Figure 4). In addition, we used the HCR results as a reference for the comparison. BLAST and HHblits search for similar sequences; therefore, we have created similar pairs by combining a query sequence with each hit found by a particular method. CD-HIT is a tool for clustering highly homologous sequences [33]. We created pairs of similar sequences for CD-HIT by combining all possible pairs of sequences in clusters. Table 1 shows the number

**Table 1.** In case of HCRs, HHblits found the highest number of similarities, while for LCRs, it was BLAST. CD-HIT reported low similarity between LCRs. We combined all similar pairs found by each method and calculated percentages which sum up to 100% in each of the columns

| | HCR | LCR | HCRs without same families in pairs | LCRs without same families in pairs |
|---|---|---|---|---|
| BLAST | 3,205,592 (15.83%) | 11,507,921 (71.58%) | 46,413 (0.65%) | 4,550,663 (67.56%) |
| HHblits | 15,296,119 (75.55%) | 4,331,254 (26.94%) | 7,096,205 (99.32%) | 2,105,748 (31.26%) |
| CD-HIT | 1,745,171 (8.62%) | 237,782 (1.48%) | 2,477 (0.03%) | 79,705 (1.18%) |

of similar pairs found by each method, while Figure 4 presents an overlap among them.

As shown in Table 1, the highest number of similar pairs for HCR results is found by HHblits (75.55% of all HCR pairs). If we remove pairs where both proteins belong to the same families, then HHblits results rises to 99.32% of all similar pairs found by all the methods. This was expected because HHblits is the most sensitive of the selected tools. On the other hand, in the case of LCRs, the highest number of similarities were found using BLAST which is over 71%, while HHblits results cover less than 27%. If we remove pairs that come from the same families in HCR parts for BLAST, then we remove almost 99% of similar pairs. In the case of LCRs, this number is about 60%. From all analysed methods, the lowest number of similar pairs was found by CD-HIT. In case of HCRs, it was 8.62% of all pairs, while for LCRs, it was only 1.48%. This may be due to its design and application as CD-HIT is mostly used to find sequence redundancy in datasets.

In Figure 4, we can observe that results for BLAST and HHblits are rather diverged in all cases. Only 1.6% of all HCR sequence pairs were found by both BLAST and HHblits and 2.2% in the case of LCRs. The differences between BLAST and HHblits results can be explained by the fact that BLAST finds closely related sequences, because it simply reports alignments with the lowest e-value. On the other hand, HHblits uses precalculated uniboost dataset that is optimized to boost diversity in sequences, and therefore, it finds alignments with more distant relationship. The intersection of CD-HIT and HHblits also has a small number of similar pairs. This observation is expected, because HHblits searches for distinct similarities while CD-HIT searches for close similarities or even can be used to detect redundancy in databases. Intersection of BLAST and CD-HIT shows that for HCRs, almost half of the results from CD-HIT overlaps with BLAST ( 49%). In the case of LCRs, most of the results are common with BLAST ( 92%). Intersection of all of the methods is poor for both HCRs and LCRs which may suggest that each of the selected methods covers different types of similarities. Remarkably, removing similar pairs that belong to the same families have a huge impact on HCR results, while LCR results are less affected.

For BLAST and HHblits results, we sorted alignments by e-value and divided them into 10 groups where each group contains about 10% of sequences. Therefore, the group size of BLAST is approximately 888 thousand pairs each and the group size of HHblits is approximately 222 thousand pairs. Within each group, we compared lengths of alignments to their similarity and the number of alignments. The similarity in an alignment between two sequences is denoted by percentage of similar amino acids where similar amino acids are these which have positive score in a scoring matrix.

For BLAST, the results for three best percentiles (the lowest e-value) differ from the results for other percentiles (Figure 5A). It suggests that for BLAST, e-value is length-dependent, and indeed, we can notice that length of alignments increases with decreasing e-value. Same situation, at least in the range from 0 to 100 amino acids, is visible in Figure 5B which shows average similarity of alignments for a given length. Along with increasing e-value, the average similarity in the percentiles tends to overlap. We can observe that alignments with lower e-value are longer and more similar than alignments with higher e-value, which is also reflected in higher score of alignments. Interestingly, HHblits results are different. In Figure 5C, e-value groups overlap which suggests that e-value is length-independent. In Figure 5D, we can observe that all groups are clearly divergent for alignments below 40, which is most of the results. Therefore, e-value in the HHblits results describes rather similarities between sequences than their length. It also indicates that alignments longer than 18aa are below 60% of similarity. Based on this results, we can notice that BLAST alignments are longer and more similar than HHblits alignments.

Corresponding figure for HCRs (Figure S1, see Supplementary Data available online at http://bib.oxfordjournals.org/) and comparison of lengths of alignments in different e-value percentiles for LCRs and HCRs are provided in Supplementary materials.

## Qualitative results

In this section, we discuss and analyse the most interesting cases (sequence alignments) obtained with the different methods. Here, we show data that indicate the following features: (i) some of the methods' features are more appropriate for HCRs than LCRs and (ii) similarity of LCR sequences may but does not have to indicate a similar function of these sequences.

### *BLAST*

For BLAST, we selected 5 representative alignments with their e-values that illustrate different problematic cases. At least, one of the proteins from alignments belongs

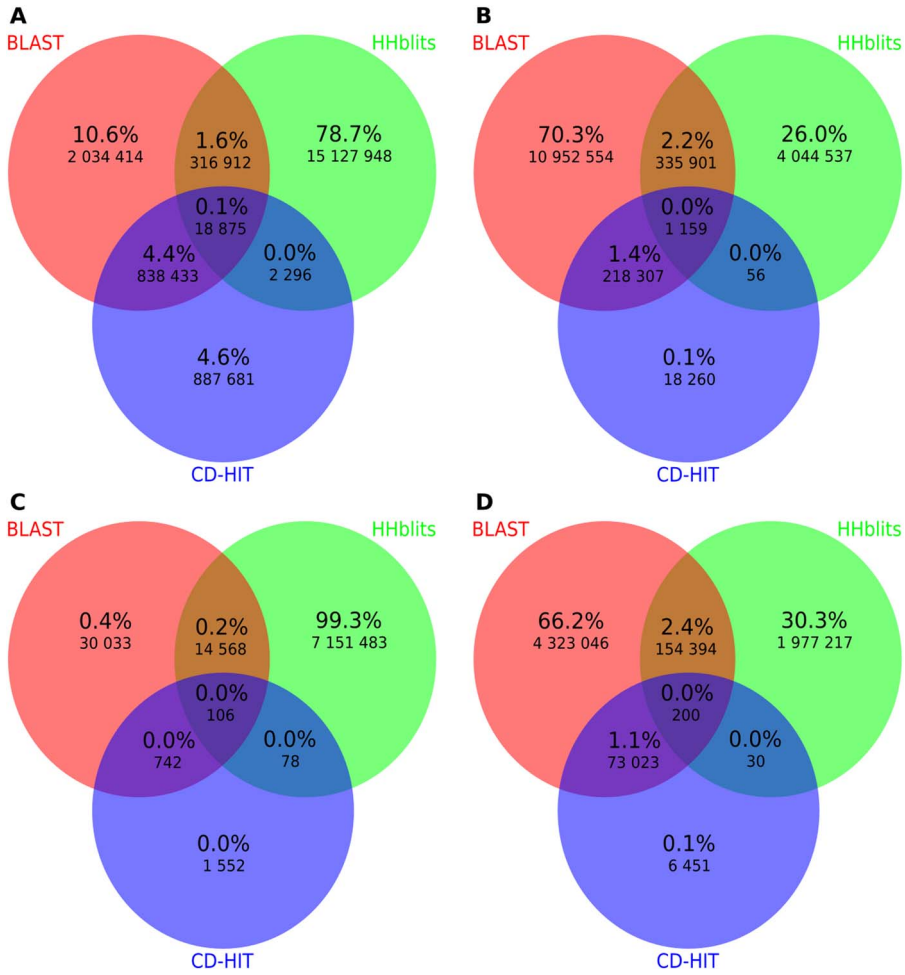


**Figure 4.** Venn diagrams present overlaps of the similar sequence pairs obtained by each of the methods. The only significant overlap is between BLAST and CD-HIT results. In the case of LCRs, CD-HIT results are subset of BLAST what is different for HCRs, where less than a half of results are common with BLAST. Other overlaps are slight, suggesting that the methods have different purposes. Panel (**A**) presents results for all HCRs, while panel (**C**) presents results for HCRs without pairs that belong to the same family. Panels (**B, D**) present the corresponding results for LCRs. Numbers in diagrams show the number of similar pairs of fragments found by specific method(s). Percentage determines how many pairs belong to the area versus all.

to the krueppel C2H2-type zinc-finger protein family. Selected alignments are presented in Table 2

The first alignment presents homopolymers of serine that consist of 25 amino acids with e-value equal to $1,10E-17$ (Table 2). The second alignment contains homopolymers of glutamine that consists of 23 amino acids. It is two amino acid shorter than the first alignment but has a lower e-value ($1,40E-23$). This situation is caused by match score assigned to serine and glutamine in the PAM matrix [39]. Matches in this matrix have different scores; therefore, two homopolymers of different amino acids may have different scores.

The third, the fourth and the fifth examples show other issues that we can observe if we use BLAST to analyse similarities between LCRs. All of these alignments are 12 aa long. The third and the fifth alignments consist of homopolymers of proline. The third one is a perfect match over the total length of LCRs. On the other hand, the fifth alignment consists of two LCRs with different lengths resulting in an alignment with a shorter LCR length. As a result, both alignments have similar lengths and scores. However, the length of homopolymer in an alignment may be crucial for its function [40]. Therefore, the fourth alignment is much better than the fifth (because both homopolymers have the same length) but have a worse e-value than the fifth alignment. N-terminal of the bottom sequence in the 4th example is 'AAPTAAPAAAATPAPTPVA' which is an imperfection of the SEG algorithm that occurs because SEG is not able to distinguish LCRs in the first part of the subsequence from the second part that is a homopolymer of proline.

We used this opportunity to analyze biological properties of the LCR pairs presented in Table 2. Remarkably, the first pair represents a polyserine stretch (Table 2). The zinc finger 865 protein is a putative transcription factor composed mainly of zinc finger motifs and an undefined low complexity N end region. Zinc fingers bind to DNA, whereas the latter fragment is usually responsible for activation or repression of transcription [41]. Serine-rich regions are usually modified and most probably function as modulators of protein binding [42]. ADP-ribosylation factor-like protein 6-interacting protein 4 (ARL6IP4) is most probably involved in splicing in nuclear speckles; however, its exact role is unknown [43, 44]. The

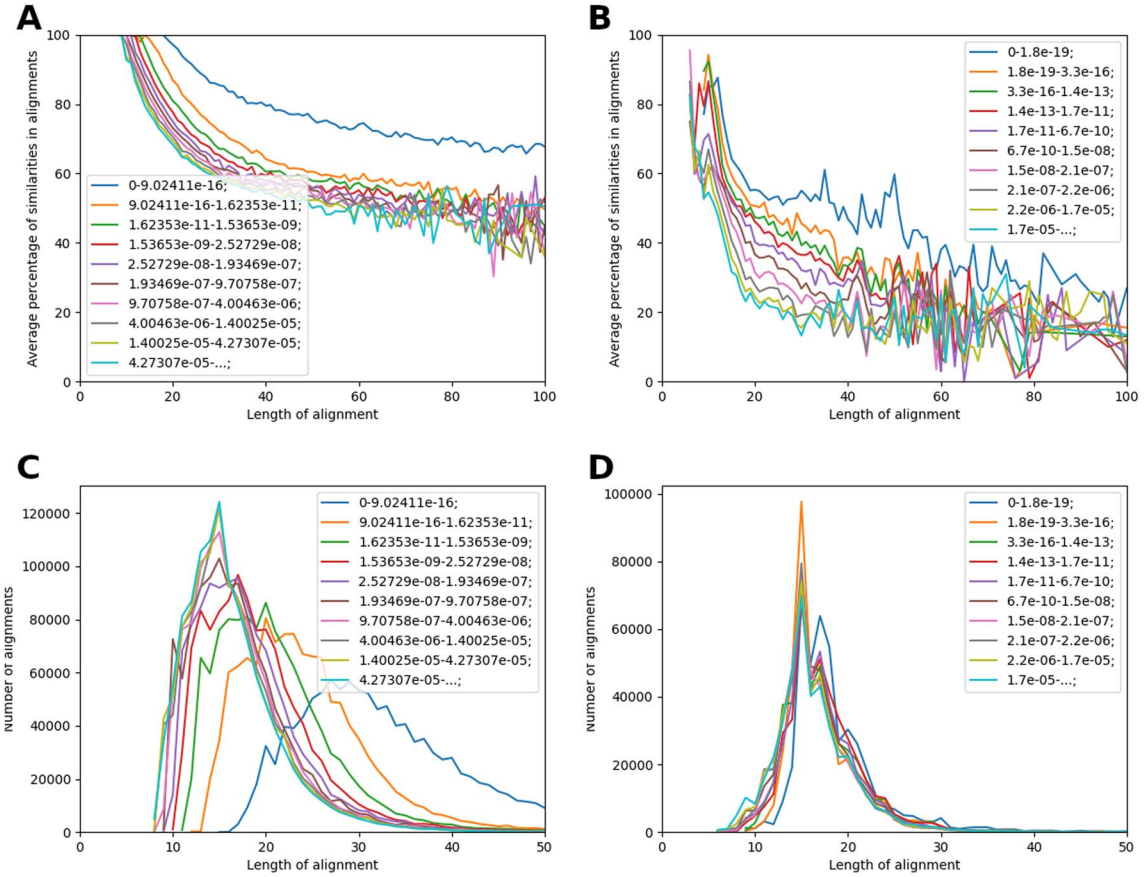


**Figure 5.** BLAST's e-value is length and similarity-dependent, while HHblits e-value is only similarity dependent. We sorted results by e-value and divided them into 10 groups of similar size. (**A, C**) show the distribution of alignments by length. (**B, D**) illustrate how alignments of a given length are averagely similar.

**Table 2.** BLAST can give ambiguous results; longer homopolymer of serine has higher e-value than shorter homopolymer of glutamine even though both homopolymers are aligned without penalty. Table shows perfect alignments collected from BLAST. It compares proteins from krueppel C2H2-type zinc-finger protein family (upper sequences) with proteins from other families (bottom sequences)

| # | Protein name | Alignments | e-value |
|---|---|---|---|
| 1 | Zinc finger protein 865 (P0CJ78)<br>midline<br>ADP-ribosylation factor-like protein 6-interacting protein 4 (Q66PJ3) | SSSSSSSSSSSSSSSSSSSSSSS (92 - 116)<br>SSSSSSSSSSSSSSSSSSSSSSS<br>SSSSSSSSSSSSSSSSSSSSSSS (257 - 281) | 1,10E-17 |
| 2 | Zinc finger protein rotund (Q9VI93)<br>midline<br>Ataxin-2 (Q99700) | QQQQQQQQQQQQQQQQQQQQQQ (739 - 761)<br>QQQQQQQQQQQQQQQQQQQQQQ<br>QQQQQQQQQQQQQQQQQQQQQQ (165 - 187) | 1,40E-23 |
| 3 | Zinc finger homeobox protein 4 (Q86UP3)<br>midline<br>Inactive histone-lysine N-methyltransferase 2E (Q3UG20) | PPPPPPPPPPPP (3112 - 3123)<br>PPPPPPPPPPPP<br>PPPPPPPPPPPP (1719 - 1730) | 3,61E-10 |
| 4 | Zinc finger homeobox protein 4 (Q86UP3)<br>midline<br>Translation initiation factor IF-2 (Q2G5E7) | PPPPPPPPPPPP (3112 - 3123)<br>PPPPPPPPPPPP<br>AAPTAAPAAAATPAPTPVAPPPPPPPPPPPP (53 - 83) | 1,84E-09 |
| 5 | Zinc finger homeobox protein 4 (Q86UP3)<br>midline<br>Glyceraldehyde-3-phosphate dehydrogenase, testis-specific (Q64467) | PPPPPPPPPPPP (3112 - 3123)<br>PPPPPPPPPPPP<br>PPPPPPPPPPPPPPPPPPP (83 - 99) | 7,75E-10 |

second pair represents human ataxin-2 and fly's rotund proteins. Ataxin-2 polyQ region seems to be responsible for dimerization/aggregationr[45–47]. Since the function of Drosophila rotund's polyQ is unknown, the parallel notion of dimerization is attractive in the context of developmental pathway in which rotund takes part [48] [49].

Moreover, a rare function can be contributed to the protein from three last examples. The polyproline stretch of ZFHX4, the zinc finger homeobox protein 4,

**Table 3.** HHblits found more distant similarities. We selected alignment results where one of the sequence is from chloroplast sensor kinase, chloroplastic protein. The midline, '—' mark indicates score above 1.5, while '+' indicates score between 0.5 and 1.5

| # | Protein name | Alignments | e-value |
|---|---|---|---|
| 1 | Chloroplast sensor kinase, chloroplastic (F4HVG8)<br>midline<br>Lamellipodin (Q70E73) | SSSSSSSSSS (39 - 49)<br>\|\|\|\|+\|\|\|\|\|\|<br>SLSSSSIKSGSSSSS (527 - 541) | 1.6e-15 |
| 2 | Chloroplast sensor kinase, chloroplastic (F4HVG8)<br>midline<br>E3 ubiquitin-protein ligase UBR4 (Q5T4S7) | SSSSSSSSSS (39 - 49)<br>\|\|\|\|\|\|\|\|\|\|\|<br>AALAASSGSSSASSSSAPVAASS (3333 - 3355) | 3.2e-13 |
| 3 | Chloroplast sensor kinase, chloroplastic (F4HVG8)<br>midline<br>Dual specificity tyrosine-phosphorylation-regulated kinase 1B (Q9Y463) | SSSSSSSSSS (39 - 49)<br>\|\|\|\|+\|\|\|\|\|\|<br>SSSTASSISSSGGSSGSSS (459 - 477) | 5.9e-15 |
| 4 | Chloroplast sensor kinase, chloroplastic (F4HVG8)<br>midline<br>Nucleolar and coiled-body phosphoprotein 1 (Q14978) | SSSSSSSSSS (39 - 49)<br>\|\|\|\|\|\|\|\|\|\|\|<br>DSSSDSDSSSSEEEEE (467 - 482) | 1.4e-14 |
| 5 | Chloroplast sensor kinase, chloroplastic (F4HVG8)<br>midline<br>Krueppel-like factor 16 (Q9BXK1) | SSSSSSSSSS (39 - 49)<br>\|\|\|\|\|+\|\|\|\|\|<br>PGGASPASSSSAASSPSSG (94 - 112) | 9.6e-13 |

```
DYRK1B   461    SSTASSISSSGGSSGSSS    478
pp150    6      SSAASGGGSSGGSSGASS    23
```

**Figure 6.** Protein low complexity region alignment of the human dual specificity tyrosine-phosphorylation-regulated kinase 1B (DYRK1B) and the small capsomere-interacting protein (pp150, pUL32, m48.2) of the cytomegalovirus obtained using FFAS03 algorithm [62].

serves as an assembly element for the human butyryl-cholinesterase into a tetramer [50–52]. We could not find any other functions ascribed to polyP of ZFHX4. The inactive histone-lysine N-methyltransferase 2E (named MLL5, KMT2E) uses the C-terminal part composed of 3 large proline-rich fragments to interact with natural cytotoxicity receptors of the NKp44 natural killer cells [53]. We could not identify any specific function(s) for the translation initiator factor IF-2. In the case of the N terminus containing the proline-rich region of testis-specific glyceraldehyde-3-phosphate dehydrogenase (GAPDHS), it was shown to bind to the tail sperm cytoskeleton [54–56] and to stabilize the structure of the enzyme itself [57].

### *HHblits*

Table 3 presents selected cases from the HHblits analysis. In each example, the query sequence (chloroplast sensor kinase, chloroplastic) is the same and it is a poly-S sequence without mutations. All examples have the same midline length, but none of them is perfectly matched to the query. However, they have lower e-value in comparison to perfectly matched alignments from BLAST of the same length.

For HHblits, alignments with a lower e-value are less similar than alignments with higher e-value found by BLAST, even if the input protein dataset is the same in both cases. Perfect alignment of serine of the same length in case of BLAST has an e-value equal to $2.95581e-05$ which is a huge difference in comparison to HHblits where the highest e-value from imperfect match is $3.2e-13$ (Table 3). Additionally, HHblits was not able to find perfect matches for a given region.

The first alignment has a better score and e-value than the second. Both of them have 100% similarity. Additionally, the second alignment has 2 mismatches and worse e-value than the first one with 3 mismatches. The third, the fourth and the fifth alignments have three mismatches and are assigned slightly different e-value, score and similarity. The third and the fifth hit sequences are 19 amino acids in length. Furthermore, the biggest difference among e-values is between the third and the fifth alignments. The fourth example is in the middle but has a different hit sequence length (16 aa).

For most of the LCRs from Table 3, we were not able to assign functions based on the literature search. However, we found an interesting similarity of the dual specificity tyrosine-phosphorylation-regulated kinase 1B (DYRK1B) to the small capsomere-interacting protein (pp150, pUL32, m48.2) of the cytomegalovirus (Figure 6). Pp150 is known to bind to capsid proteins, especially to Tri1, Tri2A and Tri2B [58, 59]. Experiments suggest that tegument protein pp150 contributes to a netlike layer that may stabilize the HCMV capsid [60]. LCR from Chloroplast sensor kinase, chloroplastic (F4HVG8) is located in the region of the transit peptide and it has to be rich in serine according to von Heijne *et al.* [61]. This serine homopolymer significantly enriches this region in the required type of amino acid which may be crucial to gain its function. We do not know the function of the LCR located in the Krueppel-like factor 16 (Q9BXK1). By similarity to the sensor kinase LCR and close vicinity to 3 zinc fingers, we speculate that it may interact with different proteins as a hinge structure.

Another example of similarity to the same serine-rich fragment of the chloroplast sensor kinase is the nucleolar and coiled-body phosphoprotein 1 (Nopp1, NOLC1) which acts as a platform to connect RNA polymerase I with enzymes responsible for ribosomal processing

and modification [63]. Experiments suggest that this serine-rich fragment may be important in binding the TFIIB transcription factor [64].

### *CD-HIT*

Our analysis of LCR results from CD-HIT is slightly different in comparison to BLAST and HHblits as results of CD-HIT are in the form of sequence clusters. Since the number of the sequences returned by CD-HIT was significantly smaller in comparison to BLAST and HHblits (Figure 4), we decided not to remove pairs in which proteins come from the same family. We noticed several interesting facts while analysing clusters created by the CD-HIT method.

In the Table 4, cluster 1 sequence P14922 is the representative of the cluster and consists of a tandem repeat of glutamine and alanine followed by a homopolymer of glutamine. This representative sequence joins two different types of LCRs to the cluster: homopolymers of glutamine and short tandem repeats of glutamine and alanine. As a result, we have sequences that are similar to representatives but it does not mean that other sequences in clusters are similar to each other. Next disadvantage of this approach is that homopolymers of glutamine (Q0CQ46) join the cluster where the representative sequence (P14922) is a combination of two types of LCRs. On the other hand, some of homopolymers of glutamine join clusters where the representative is also a homopolymer of glutamine. In such a case, we have wrong situation where homopolymers of glutamine are spreaded among different clusters.

Another disadvantage is that CD-HIT creates two different clusters from two highly similar sequences from composition and length perspectives. Table 4 contains an example of this situation. Sequence with Uniprot AC Q8TF68 which belongs to cluster 6 and sequence Q9EQJ4 which belongs to cluster 7 are highly similar but were assigned to different clusters as representative sequences. On the other hand, cluster 4 consists of the orphan sequence (Q9ZTX8). However, we can notice that this sequence is similar to the representative sequence in cluster 3. These results are correct because these sequences have different repetitive patterns. The sequence in cluster 4 is composed of the LSQQQQQQ motif, while the representative sequence in cluster 3 is composed of the LQQQQQ motif.

The fifth cluster from Table 4 (5th row) consists of serine homopolymers of different lengths. LCR from protein Q75JC9 has the longest sequence with 306 serines. The same cluster contains a region from P78424 with 11 serines. Such differences in length can be reduced by changing the cutoff option of CD-HIT, which may be adjusted by the user for a specific problem. Higher values of the cutoff option results in lower diversity of sequence lengths in clusters. By default, the parameter is turned off. Detailed analysis of this parameter is provided in Supplementary material.

## Discussion and conclusions

Our analysis confirms that selected methods are most efficient for comparing high complexity protein sequences as they rely on statistical models designed for sequences with diverse amino acid composition. This is why masking low complexity parts improves searching for homologous proteins [25]. An obvious way to include LCRs into the analysis is to disable the masking. This still does not solve the problem related to the fact that these methods were optimized for HCPs of sequences.

The methods analysed in this paper use general purpose scoring matrices. They are efficient to align typical protein sequences but fail while applying them for non-standard amino acid composition of sequences such as LCRs [65]. For example, BLOSUM, one of the most popular scoring matrix, was built from the BLOCK database which contained only about a promile of proteins with LCRs [66]. Sequence regions in proteins with non-standard amino acid composition (especially LCRs) have their own structural and amino acid preferences [1, 65]. For example, A-rich and L-rich regions promote alpha-helix formation, while H-rich and P-rich regions frequently overlap with disorder regions [67]. Whenever possible, it is recommended to use specialised scoring matrices for different types of protein domains such as intrinsically disordered regions [68]. Unfortunately, many tools for protein sequence comparison lack of parameter which enable scoring matrix selection or they allow to select predefined set of matrices only.

MSA and consequently a dataset of profiles of Hidden Markov Models for LCRs may be created in many different ways. One problematic case is shown in Figure 3 where three selected tools gave three different results. Only one method was able to align the valine correctly. The problem presented is well known as shift-errors (the erroneous positioning of a single indel whose length is preserved) and occurs in protein and genome sequences [69]. Such kind of error is especially abundant in LCR alignments. To solve this kind of error, we need data about evolution of proteins, but in many cases, it is limited due to lack of ancestral sequences [70]. MSA creation is a general issue related to HCRs as well. It is well known that automatically generated MSA need to be manually curated [71, 72].

E-value is a statistic useful while searching for similarities that have biological meaning as it provide the estimate if a particular similar sequence may occur in a database by a chance [73]. However, in the case of LCRs, especially homopolymers, the number of even identical sequences in a given database may vary a lot because in nature, some of homopolymers occur more frequently than others [74]. Therefore, to assess the significance of an alignment of LCRs, we suggest to pay more attention to score, identity and similarity metrics than e-value.

Local alignment causes loss of information about length of longer LCRs than query sequence. Such a case is presented in Table 2 for homopolymers of proline

**Table 4.** CD-HIT places dissimilar LCRs to the same clusters and similar LCRs to different clusters. Sequences where * is added to Uniprot AC are representatives of clusters

| # | Example sequences in cluster | Uniprot AC | Number of LCRs in cluster |
|---|---|---|---|
| 1 | QQQQQQQQAQQQ | Q0CQ46 | 8 (99 cluster) |
| | QAQAQAQAQAQAQAQ | Q5ABZ2 | |
| | QLQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAQAHAQ | P14922* | |
| | AQAQAQAQAQAQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQLQPLPRQ | | |
| | QLQQ | | |
| 2 | GAGGGGGGGGGGGGSGGGGGGGGAGGAGGAAAAAAGAGAVAAQAQAQAA | E9Q4N7* | 201(301 cluster) |
| | AAAAAAAAAAAGGGGGGGYGSSSSGYGV | | |
| | AAAAAAAAAAAAAAA | P35453 | |
| | GGGGGGGGGTGGGGGGG | O77215 | |
| 3 | QQQQQQQEQKQQLQQQQQQQQQLQQQQQQQQQQ | P04725 | 3 (233 cluster) |
| | LQQLQQQQQLQQQQQLQQQQQQQLQQQQQLQQQQLQQQQQQQQLQQQQQ | O14686* | |
| | QQLQQQQQQLQQQQQQQQQFQQQQQQQQ | | |
| | LQRQRQQQQLQQQQQQQLQQQQ | P52288 | |
| 4 | LQQQLSQQQQQLSQQQQQQQQLSQQQQQQLSQQQQQQLSQQQQQQLSQQ | Q9ZTX8* | 1 (743 cluster) |
| | QQQQ | | |
| 5 | SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS | Q8BTI8 | 119 (11 cluster) |
| | SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS | Q75JC9* | |
| | SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS | | |
| | SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS | | |
| | SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS | | |
| | SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS | | |
| | SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS | | |
| | SSSSSSSSSSSSS | | |
| | SSSSSSSSSSS | P78424 | |
| 6 | QQQVQAAAAAAAVAQAQAQAQAQAQAQAQAQAQAQASQASQQQQQQQQQ | Q8TF68* | 1 (625 cluster) |
| | QQQQQQQ | | |
| 7 | QQQVQAAAAAAAVAQAQAQAQAQAQAQAQAQAQAQAQAQASQASQQQ | Q9EQJ4* | 1 (723 cluster) |
| | QQQQ | | |

(alignment 5). BLAST compares entire poly-P region from *Zinc finger homeobox protein 4* protein to only a part of poly-P region from *Glyceraldehyde-3-phosphate dehydrogenase,testis-specific* protein. Therefore, local alignment based tools lose information about difference in LCR lengths. However, current knowledge says that homopolymer length may affect protein function. For instance, poly-Q region may cause neurological disease if it reach a given length [75]. Therefore, differences in lengths should be somehow penalized in LCR alignments for example by comparing them globally. Another approach to analyse the similarity of LCRs found in the literature is a metric based on Jaccard index between seqeunces [76]. In this solution, the Jaccard score may be easily penalised by difference in the number of residues between composition vectors.

Selection of representative sequences in CD-HIT may result in wrong assignment of sequences of different types of LCRs. This is especially visible in LCRs that consist of two different types of LCRs. As mentioned in Results section, if a representative sequence consists of two types of LCRs, then three types of LCRs may join this cluster: LCRs of the first type, LCRs of the second type and the combination of both LCRs. As a result, a single cluster contains different types of LCRs. Such a scenario is present in examples 1 and 2 in Table 4. The problem occurs because CD-HIT uses local alignment to join sequences to cluster that are similar to representative. A metric designed for LCR comparison is crucial for better performance of general purpose protein sequence clustering since many methods use it to create clusters [77, 78]. Another interesting approach to cluster LCRs is through tagging them by their types [79].

Even if we showed that standard methods are not appropriate for the similarity analysis of LCRs we could identify some examples that are statistically sound (see Results section for BLAST and HHblits). In this work, we collected a few examples of LCR similarity/identity that may follow the rule of transitivity. If the function of one LCR is known and we show its similarity to another LCR, then we may hypothesize that the role of the second LCR may be identical or similar, e.g. if LCR1 binds a protein, we may argue for an analogous binding mechanism of a similar LCR2.

Our work focuses on three state-of-the-art methods for protein sequence comparison that are based on the most popular approaches for protein similarity analysis such as local alignment or profiles of Hidden Markov Models. There are other tools for protein sequence comparison that can be found in the literature, but they usually rely on similar approaches. In this paper, we point out why these methods are not suitable for LCR analysis. However, since there are no tools specifically designed for LCRs, we suggest to use BLAST. It seems to be the best

choice to search for similar LCRs. HHblits may also be useful in LCR analysis, but for more distant similarities. Finally, we conclude that CD-HIT cannot be used for analysing LCRs as the local comparison to representative sequences results in badly clustered sequences that are not similar to each other. Scientists should be aware of these drawbacks while using these methods for searching for similar LCRs. On the other hand, our results may be used to improve existing methods or to design new ones especially crafted for LCR comparison.

**Key Points**

- We would like to alert the community that similarity searches reported by canonical tools may result with false positive hits, even if they use the optimal parameter set for these methods.
- We indicate which design assumptions of the selected methods are not suitable for the analysis of low complexity regions (LCRs).
- In the article, we advise on how to adjust HHblits and CD-HIT parameters to find similar LCRs more efficiently.
- This knowledge can be used to improve existing methods or to create new methods specifically designed to analyze the similarity between low complexity regions.

## Data Availability Statement

The data underlying this article are available in the UniProtKB/Swiss-Prot at www.uniprot.org, and can be accessed with major release-2020_04.

## Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

## Acknowledgements

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

## Funding

European Union through the European Social Fund (grant POWR.03.02.00-00-I029/17 to P.J.); National Science Centre (grant no. 2020/39/B/ST6/03447).

## References

1. Kumari B, Kumar R, Kumar M. Low complexity and disordered regions of proteins have different structural and amino acid preferences. *Mol Biosyst* 2015;**11**(2):585–94.
2. Franzmann TM, Alberti S. Prion-like low-complexity sequences: Key regulators of protein solubility and phase behavior. *J Biol Chem* 2019;**294**(18):7128–36.
3. Aditi, Mason AAC, Sharma M, Dawson TR, *et al.* MAPK- and glycogen synthase kinase 3-mediated phosphorylation regulates the DEAD-box protein modulator Gle1 for control of stress granule dynamics. *J Biol Chem* 2019;**294**(2):559–75.
4. Andrew Chong P, Vernon RM, Forman-Kay JD. Rgg/rg motif regions in rna binding and phase separation. *J Mol Biol* 2018;**430**(23):4650–65.
5. Kato M, Lin Y, McKnight SL. Cross-$\beta$ polymerization and hydrogel formation by low-complexity sequence proteins. *Methods* 2017;**126**:3–11.
6. Kulkarni P, Uversky VN. Intrinsically Disordered Proteins: The Dark Horse of the Dark Proteome. *Proteomics* 2018;**18**(21-22):e1800061.
7. Schafferhans A, O'Donoghue SI, Heinzinger M, *et al.* Dark Proteins Important for Cellular Function. *Proteomics* 2018;**18**(21-22):e1800227.
8. Perdigão N, Rosa A. Dark proteome database: studies on dark proteins. *High-Throughput* 2019;**8**(2):E8.
9. Ntountoumi C, Vlastaridis P, Mossialos D, *et al.* Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved. *Nucleic Acids Res* 2019;**47**:9998–10009.
10. Altschul SF, Gish W, Miller W, *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**(3):403–10.
11. Remmert M, Biegert A, Hauser A, *et al.* HHblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nat Methods* 2012;**9**(2):173–5.
12. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**(13):1658–9.
13. UniProt Consortium. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**(D1):D480–9.
14. Mier P, Paladin L, Tamana S, *et al.* Disentangling the complexity of low complexity proteins. *Brief Bioinform* 2020;**21**(2):458–72.
15. Promponas VJ, Enright AJ, Tsoka S, *et al.* CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics (Oxford, England)* October 2000;**16**(10):915–22.
16. Harrison PM. fLPS: Fast discovery of compositional biases for the protein universe. *BMC Bioinformatics* 2017;**18**:476.
17. Newman AM, Cooper JB. XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* 2007;**8**(1):382.
18. Jorda J, Kajava AV. T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* 2009;**25**(20):2632–8.
19. Albá MM, Laskowski RA, Hancock JM. Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics (Oxford, England)* 2002;**18**(5):672–8.
20. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* 1993;**17**(2):149–63.
21. Radó-Trilla N, Albà MM. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. *BMC Evol Biol* 2012;**12**(1):155.
22. Radó-Trilla N, Arató K, Pegueroles C, *et al.* Key role of amino acid repeat expansions in the functional diversification of duplicated transcription factors. *Mol Biol Evol* 2015;**32**(9):2263–72.
23. Jarnot P, Ziemska-Legiecka J, Grynberg M, *et al.* LCR-BLAST-a new modification of blast to search for similar low complexity regions in protein sequences. In: *International Conference on Man–Machine Interactions*. Springer, Cham: Springer, 2019, 169–80.
24. Pearson WR. Selecting the right similarity-scoring matrix. *Curr Protoc Bioinformatics* 2013;**43**(1):3–5.
25. Coronado JE, Attie O, Epstein SL, *et al.* Composition-modified matrices improve identification of homologs of saccharomyces

cerevisiae low-complexity glycoproteins. *Eukaryot Cell* 2006;**5**(4):628–37.
26. Mirdita M, von den Driesch L, Galiez C, *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res* 2017;**45**(D1):D170–6.
27. Steinegger M, Söding J. MMseq2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**(11):1026–8.
28. Biegert A, Söding J. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* 2008;**24**(6):807–14.
29. Söding J, Remmert M, Hauser A. HHsuite for sensitive protein sequence searching based on hmm-hmm alignment, user guide (Online) (17 January 2021, date last accessed).
30. Edgar RC. MUSCLE: multiple sequecne alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**(5): 1792–7.
31. Lassmann T, Sonnhammer ELL. Kalign - an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* 2005;**6**(1):1–9.
32. Sievers F, Wilm A, Dineen D, *et al.* (eds). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011;**7**(1):539.
33. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 2001;**17**(3):282–3.
34. Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* 2002;**18**(1):77–82.
35. O'Leary NA, Wright MW, Brister JR, *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;**44**(D1): D733–45.
36. Szklarczyk D, Gable AL, Lyon D, *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**(D1):D607–13.
37. Cunningham F, Allen JE, Allen J, *et al.* Ensembl 2022. *Nucleic Acids Res* 2022;**50**(D1):D988–95.
38. Mistry J, Chuguransky S, Williams L, *et al.* (eds). Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;**49**(D1):D412–9.
39. Dayhoff M, Schwartz R, Orcutt B. 22 a model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure* 1978;**5**:345–52.
40. Almeida B, Fernandes S, Abreu IA, *et al.* Trinucleotide repeats: a structural perspective. *Front Neurol* 2013;**4**:76.
41. Dang DT, Pevsner J, Yang VW. The biology of the mammalian krüppel-like family of transcription factors. *Int J Biochem Cell Biol* 2000;**32**(11–12):1103–21.
42. Syafruddin SE, Mohtar MA, Nazarie WFWM, *et al.* Two sides of the same coin: The roles of klf6 in physiology and pathophysiology. *Biomolecules* 2020;**10**(10):1378.
43. Sasahara K, Yamaoka T, Moritani M, *et al.* Molecular cloning and expression analysis of a putative nuclear protein, sr-25. *Biochem Biophys Res Commun* 2000;**269**(2):444–50.
44. Ouyang P. Srrp37, a novel splicing regulator located in the nuclear speckles and nucleoli, interacts with sc35 and modulates alternative pre-mrna splicing in vivo. *J Cell Biochem* 2009;**108**(1):304–14.
45. Petrakis S, Schaefer MH, Wanker EE, *et al.* Aggregation of polyq-extended proteins is promoted by interaction with their natural coiled-coil partners. *Bioessays* 2013;**35**(6):503–7.
46. Totzeck F, Andrade-Navarro MA, Mier P. The protein structure context of polyq regions. *PLoS One* 2017;**12**(1):e0170801.
47. Bondarev SA, Antonets KS, Kajava AV, *et al.* Protein co-aggregation related to amyloids: Methods of investigation, diversity, and classification. *Int J Mol Sci* 2018;**19**(8):2292.
48. St SE, Pierre MI, Galindo JP, *et al.* Control of drosophila imaginal disc development by rotund and roughened eye: differentially expressed transcripts of the same gene encoding functionally distinct zinc finger proteins. *Development* 2002;**129**(5):1273–81.
49. Li Q, Barish S, Okuwa S, *et al.* A functionally conserved gene regulatory network module governing olfactory neuron diversity. *PLoS Genet* 2016;**12**(1):e1005780.
50. Biberoglu K, Schopfer LM, Saxena A, *et al.* Polyproline tetramer organizing peptides in fetal bovine serum acetylcholinesterase. *Biochim Biophys Acta* 2013;**1834**(4):745–53.
51. Biberoglu K, Schopfer LM, Tacal O, *et al.* The proline-rich tetramerization peptides in equine serum butyryl-cholinesterase. *FEBS J* 2012;**279**(20):3844–58.
52. Peng H, Schopfer LM, Lockridge O. Origin of polyproline-rich peptides in human butyrylcholinesterase tetramers. *Chem Biol Interact* 2016;**259**:63–9.
53. Baychelier F, Sennepin A, Ermonval M, *et al.* Identification of a cellular ligand for the natural cytotoxicity receptor nkp44. *Blood* 2013;**122**(17):2935–42.
54. Westhoff D, Kamp G. Glyceraldehyde 3-phosphate dehydrogenase is bound to the fibrous sheath of mammalian spermatozoa. *J Cell Sci* 1997;**110**(15):1821–9.
55. Bunch D, Welch JE, Magyar PL, *et al.* Glyceraldehyde 3-phosphate dehydrogenase-s protein distribution during mouse spermatogenesis. *Biol Reprod* 1998;**58**(3):834–41.
56. Kuravsky ML, Aleshin VV, Frishman D, *et al.* Testis-specific glyceraldehyde-3-phosphate dehydrogenase: origin and evolution. *BMC Evol Biol* 2011;**11**(1):1–15.
57. Kuravsky M, Barinova K, Marakhovskaya A, *et al.* Sperm-specific glyceraldehyde-3-phosphate dehydrogenase is stabilized by additional proline residues and an interdomain salt bridge. *Biochim Biophys Acta* 2014;**1844**(10):1820–6.
58. Baxter MK, Gibson W. Cytomegalovirus basic phosphoprotein (pul32) binds to capsids in vitro through its amino one-third. *J Virol* 2001;**75**(15):6865–73.
59. Yu X, Jih J, Jiang J, *et al.* Atomic structure of the human cytomegalovirus capsid with its securing tegument layer of pp150. *Science* 2017;**356**(6345):eaam6892.
60. Dai X, Yu X, Gong H, *et al.* The smallest capsid protein mediates binding of the essential tegument protein pp150 to stabilize dna-containing capsids in human cytomegalovirus. *PLoS Pathog* 2013;**9**(8):e1003525.
61. von Heijne G, Steppuhn J, Herrmann RG. Domain structure of mitochondrial and chloroplast targeting peptides. *Eur J Biochem* 1989;**180**(3):535–45.
62. Jaroszewski L, Rychlewski L, Li Z, *et al.* Ffas03: a server for profile–profile sequence alignments. *Nucleic Acids Res* 2005;**33**(suppl_2):W284–8.
63. Werner A, Iwasaki S, McGourty CA, *et al.* Cell-fate determination by ubiquitin-dependent regulation of translation. *Nature* 2015;**525**(7570):523–7.
64. Miau L-H, Chang C-J, Tsai W-H, *et al.* Identification and characterization of a nucleolar phosphoprotein, nopp140, as a transcription factor. *Mol Cell Biol* 1997;**17**(1): 230–9.
65. Trivedi R, Nagarajaram HA. Substitution scoring matrices for proteins-an overview. *Protein Sci* 2020;**29**(11):2150–63.

66. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci* 1992;**89**(22): 10915–9.
67. Cascarina SM, Elder MR, Ross ED. Atypical structural tendencies among low-complexity domains in the protein data bank proteome. *PLoS Comput Biol* 2020;**16**(1):e1007487.
68. Trivedi R, Nagarajaram HA. Amino acid substitution scoring matrices specific to intrinsically disordered regions in proteins. *Sci Rep* 2019;**9**(1):1–12.
69. Landan G, Graur D. Characterization of pairwise and multiple sequence alignment errors. *Gene* 2009;**441**(1–2):141–7.
70. Bawono P, Dijkstra M, Pirovano W, *et al.* (eds). Multiple sequence alignment. In: *Bioinformatics*. Springer, 2017, 167–89.
71. Ranwez V, Chantret N. Strengths and limits of multiple sequence alignment and filtering methods. Scornavacca, Celine; Delsuc, Frédéric; Galtier, Nicolas. *Phylogenetics in the genomic era*, pp. 2.2:1–2.2:36, 2020.
72. Hubley R, Wheeler TJ, Smit AFA. Accuracy of multiple sequence alignment methods in the reconstruction of transposable element families. bioRxiv 2021.08.17.456740.
73. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci* 1990;**87**(6):2264–8.
74. Chavali S, Singh AK, Santhanam B, *et al.* Amino acid homorepeats in proteins. *Nat Rev Chem* 2020;**4**(8):420–34.
75. Laffita-Mesa JM, Paucar M, Svenningsson P. Ataxin-2 gene: a powerful modulator of neurological disorders. *Curr Opin Neurol* 2021;**34**(4):578.
76. Kastano K, Mier P, Andrade-Navarro MA. The role of low complexity regions in protein interaction modes: an illustration in huntingtin. *Int J Mol Sci* 2021;**22**(4):1727.
77. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002;**30**(7):1575–84.
78. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nat Commun* 2018;**9**(1):1–8.
79. Cascarina SM, King DC, Nishimura EO, *et al.* Lcd-composer: an intuitive, composition-centric method enabling the identification and detailed functional mapping of low-complexity domains. *NAR Genom Bioinform* 2021;**3**(2):lqab048.