Published online 18 May 2020 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* ***W77–W84***
doi: 10.1093/nar/gkaa339

PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins

Patryk Jarnot [1], Joanna Ziemska-Legiecka[2], Laszlo Dobson [3,4], Matthew Merski[5], Pablo Mier [6], Miguel A. Andrade-Navarro [6], John M. Hancock [7], Zsuzsanna Dosztányi[8], Lisanna Paladin [9], Marco Necci[9], Damiano Piovesan[9], Silvio C. E. Tosatto [9], Vasilis J. Promponas [10], Marcin Grynberg [2] and Aleksandra Gruca [1,*]

[1]Department of Computer Networks and Systems, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland, [2]Institute of Biochemistry and Biophysics PAS, Pawinskiego 5A, 02-106 Warsaw, Poland, [3]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Práter u. 50/A, 1083 Budapest, Hungary, [4]Research Centre for Natural Sciences, Magyar Tudósok Körútja 2, 1117 Budapest, Hungary, [5]Structural Biology Group, Biological and Chemical Research Centre, Department of Chemistry, University of Warsaw, Żwirki i Wigury 101, 02-089 Warsaw, Poland, [6]Faculty of Biology, Johannes Gutenberg University Mainz, Hans-Dieter-Hüsch-Weg 15, 55128 Mainz, Germany, [7]ELIXIR, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK, [8]Department of Biochemistry, ELTE Eötvös Loránd University, Budapest, Pázmány Péter stny 1/c 1117, Budapest, Hungary, [9]Department of Biomedical Sciences, University of Padova, Via Ugo Bassi 58/B, 35131 Padova, Italy and [10]Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, P.O. Box 20537, Nicosia, CY 1678, Cyprus

Received March 07, 2020; Revised April 08, 2020; Editorial Decision April 21, 2020; Accepted May 01, 2020

ABSTRACT

Low complexity regions (LCRs) in protein sequences are characterized by a less diverse amino acid composition compared to typically observed sequence diversity. Recent studies have shown that LCRs may co-occur with intrinsically disordered regions, are highly conserved in many organisms, and often play important roles in protein functions and in diseases. In previous decades, several methods have been developed to identify regions with LCRs or amino acid bias, but most of them as stand-alone applications and currently there is no web-based tool which allows users to explore LCRs in protein sequences with additional functional annotations. We aim to fill this gap by providing PlaToLoCo - PLAtform of TOols for LOw COmplexity—a meta-server that integrates and collects the output of five different state-of-the-art tools for discovering LCRs and provides functional annotations such as domain detection, transmembrane segment prediction, and calculation of amino acid frequencies. In addition, the union or intersection of the results of the search on a query sequence can be obtained. By developing the PlaToLoCo meta-server, we provide the community with a fast and easily accessible tool for the analysis of LCRs with additional information included to aid the interpretation of the results. The PlaToLoCo platform is available at: http://platoloco.aei.polsl.pl/.

INTRODUCTION

Low complexity regions (LCRs) are stretches of protein sequences that are characterized by a less diverse amino acid composition compared to the typical sequence diversity observed in proteins. LCRs can consist of single amino acid repeats (so-called homorepeats, homopolymeric regions), repetitive regions which consist of patterns of residues that are adjacent to each other (such as direpeats, tandem repeats or imperfect repeats), or compositionally biased regions (regions that lack a particular pattern but are enriched in a few amino acid types) (1).

For many years, LCRs were ignored by the scientific community, treated as a non-functional part of a proteome (or 'junk') (2). However, recent research shows that low complexity sequences may co-occur with intrinsically disordered regions and these represent a significant portion of proteins that lack 3D structural information, the so-called dark proteome (3). Recent studies have also shown that LCRs are highly conserved in many organisms (4). While the function of most LCRs is still a mystery, recent evidences suggest that LCRs often play important roles in structure stability preservation (5), adhesion (6), transduction of conformational information (7), membrane interac-

*To whom correspondence should be addressed. Tel: +48 32 2371154; Fax: +48 32 2372733; Email: aleksandra.gruca@polsl.pl

tions (8), DNA binding (9), the binding of metals by cysteine, histidine, or charge clusters (10), and in driving the formation of membraneless organelles through phase separation (11). LCRs are also directly involved in the development of various diseases, including neurodegenerative diseases and cancer (12,13).

Here we present PlaToLoCo - Platform of Tools for Low Complexity, which is a meta-server that integrates and collects the output of five different state-of-the-art tools for the discovery of LCRs. These methods (SEG, CAST, fLPS, SIMPLE and GBSC) were selected because they represent a range of approaches to detect low complexity regions. SEG (14) is the most commonly used tool for masking low complexity regions and it is often used by the similarity search tool BLAST (15). CAST (16) and fLPS (17) are methods focused on searching for compositionally biased regions. SIMPLE (18) is designed to detect highly cryptic (non-tandem) repeats and GBSC (manuscript in preparation) is a graph-based method also designed for finding repeats in protein sequences, including non-perfect repeats. PlaToLoCo not only searches for low complexity regions but also for the first time provides functional annotations for these regions such as domain detection, transmembrane segment prediction, and calculation of amino acid frequencies. In addition, the consensus or intersection of the results of the search on the query sequence can be obtained by the user.

To our best knowledge there currently exists only one tool with a similar functionality which is LCR-eXXXplorer (19). This tool offers precalculated results of CAST and SEG (with their default settings) for all UniProtKB/SwissProt database (20) entries (release 2015_01), while features currently annotated in UniProtKB for these entries are retrieved and displayed along with the detected LCRs. The main disadvantage of this approach is the limited number of LCR detection methods and the fact that it does not allow users to submit their own protein sequences.

MATERIALS AND METHODS

Different definitions of LCRs have been proposed in the literature leading to various tools for their detection (1). PlaToLoCo incorporates widely used LCR detection methods, which are representatives of the main formulations that have been proposed for this task. In this section, we present a short description of the methods implemented in the PlaToLoCo server.

SEG has been the *de facto* standard for LCR detection for at least two decades. Its popularity stems from multiple factors, including (a) a simple—yet rigorous – mathematical formulation, (b) availability of the source code and (c) its inclusion as a default masking option within the NCBI-BLAST package. Briefly, SEG uses a two-pass, sliding-window approach for calculating a measure of information content (sequence complexity or Shannon entropy) along a sequence in overlapping peptides of fixed length (trigger window length – W). SEG initially detects candidate windows satisfying a strict predefined threshold (trigger complexity – $K_2(1)$). Overlapping candidates are merged and extensions are made during the second pass, ending with 'contigs' of complexity lower than a more relaxed threshold (extension complexity – $K_2(2) > K_2(1)$). A final optimization step performs a brute-force search within each contig for the subsequence with the minimum occurrence probability. The number and sequence properties of the detected LCRs in a dataset (e.g. LCR length distribution) clearly depend on the choice of W, $K_2(1)$ and $K_2(2)$, with the authors proposing specific parameter settings for particular purposes (14).

CAST detects compositionally biased regions in a query sequence in an implicit manner by identifying regions that exhibit high similarity to homopolymers of any type of proteinogenic amino acid. Local sequence similarities against homopolymers (signifying LCR candidates) are detected using a space- and time-efficient implementation of the Smith-Waterman algorithm (21) using the BLOSUM62 scoring matrix and an infinite gap penalty. A detection step is followed by masking of the region of highest similarity; this procedure is then iterated until the highest detected similarity score values fall below a predetermined threshold.

SIMPLE identifies short motifs in a sliding window with frequencies that are higher than those in sliding windows of the same size in randomized versions of the sequence under study. Repeated motifs are given a score and those that score higher than randomized sequences are marked as significant. Additionally, an average score for all the detected windows in a sequence is generated as an indicator of how repetitive that sequence is as a whole. The method was originally published in 1986 for DNA sequences (22) and subsequently updated (23), and extended for usage with protein sequences (18).

fLPS (fast Lowest Probability Subsequences) detects compositional biases (CBs) in protein sequences. Such regions were found in functionally important proteins, e.g. in prion-like proteins (24). The underlying algorithm is based on the LPS algorithm (25), although fLPS is significantly more computationally efficient. This method has three main parts. First, a high bias probability threshold t is used to identify segments with skewed amino acid composition (QUICKSCAN). Overlapping regions are merged into contigs. Then, contigs are scanned for the LPSs using windows sizes from maximum M to minimum m (MINIMIZATION). Finally, LPSs of different residues are combined if their P-values will be higher than the combined LPS P-value (MERGE). These segments are trimmed or extended if their P-values can be further decreased.

GBSC (*Jarnot, P., Ziemska-Legiecka, J., Grynberg, M. and Gruca, A., in preparation*) identifies repetitive regions composed of one amino acid (homorepeats) or a few amino acids (STRs – short tandem repeats). This method is based on weighted paths of graphs built from consecutive 2-mers in the sequence. This algorithm also identifies imperfect homo/tandem repeats from protein sequences by scanning all the provided sequences. Because this method can detect imperfect repeats, insertions in between the repeats and mutations of amino acids within repetitive regions are less confounding to the method. The user can set the window size used to scan the sequences and the minimum number of occurrence of the repetitive pattern as parameters. The positions of tandem repeats and the information about the type of repeat are returned as outputs. The de-

tailed description of the GBSC method is provided in the Supplementary material 1 (Suppl1).

WEB SERVER DESCRIPTION

The PlaToLoCo web server available at http://platoloco.aei.polsl.pl takes a list of UniProt accession numbers or a list of protein sequences in FASTA format and the parameters of implemented methods. By default, all method parameters are set as suggested by the authors of the original papers, however, the web interface allows users to modify these values. Additionally, for SEG and fPLS, we also provide suggested, predefined parameter settings that narrow search results and are tailored to specific needs. For SEG the additional parameter settings are SEG *intermediate* ($W = 15$, $K_1 = 1.9$ and $K_2 = 2.5$) and SEG *strict* ($W = 15$, $K_1 = 1.5$ and $K_2 = 2.8$). According to previous reports, SEG *intermediate* is optimized for detecting longer and more repetitive low complexity regions in eukaryotes (26), while SEG *strict* ensures that the regions identified correspond to strongly compositionally biased sequences while also allowing for substantial sequence diversity (27). For fLPS, we provide the additional parameter set fLPS *strict* ($m = 5$, $M = 25$, and $t = 0.00001$) suggested by the method author as more suitable for detecting compositionally biased regions (17).

As the computation time may vary from a few seconds up to several hours depending on the number and length of submitted sequences and the load of the PlaToLoCo server, users are provided with jobIDs immediately after submitting the list of sequences. This unique identifier can be used to retrieve the results on demand at a later time as the results are stored on the server for seven days. When the submitted job is finished, a new panel is produced by the web server showing the summary of results, including LCR prediction results for all the submitted sequences.

Further results and statistics are available in the *sequence details* panel (Figure 1A) which is activated by clicking on the summary of a particular query sequence. The panel offers a graphical representation of the LCRs detected by the selected methods (Figure 1B). The graph consists of three parts. The amino acid sequence is shown on top (while the sequence is zoomed). Below that the LCR sequences found by the different methods are presented, followed by the Shannon entropy computed over a sliding widow of size 7, and then the selected enrichment information. The positions in the query sequences are shown at the bottom of the graph. Below there is the *amino acid frequency* section which shows the amino acid distribution of the sequence compared to the amino acid frequencies in various databases such as Uniprot/SwissProt, nextProt (28), DisProt (29) and PDB (30) (Figure 1C). Users can personalize their results in the *methods consensus* panel, by selecting their preferred LCR detection methods and selecting either the consensus result using a strict definition calculated from the intersection of the selected results, or using a more permissive definition calculated from the union of the results (Figure 1D). The selected consensus results can also be downloaded by the user in fasta format. The next section is called *Pfam and PDB details* and it provides detailed information about the Pfam domains annotated to the analyzed sequence as well as PDB structures associated with that domain (Figure 1E). Finally, the *region details* panel shows the summary of all the predicted LCRs displaying the specific amino acids enriched in the queried sequences (Figure 1F).

Implementation

In order to run the LCR identification methods we decided to unify the input and output of each method to simplify the design of the server under the wrapper module. An additional advantage of this approach is that the user can run all of the methods from the command line using unified fasta format sequences as input and receiving regions of interest as results in the same format. Other functionalities incorporated in PlaToLoCo for enriching results are: (a) Phobius (31), for simultaneously predicting secretory signal peptides and transmembrane helices, (b) PFAM (32) for displaying functional domains, and (c) amino acid frequencies in the query sequence compared to the UniProtKB/Swiss-Prot database frequencies.

PlaToLoCo uses several client-side Javascript libraries and components. The platform is based on the Angular.js library which enabled the creation of a responsive and intuitive application. It also utilizes: xml-js to parse data collected from the Pfam database, chart.js to present the amino acid frequencies chart shown in Figure 1C and feature-viewer (33) to present the sequence details shown in Figure 1B. The client communicates with the server using a RESTful API.

Server-side software was developed based on the FLASK library, which is a complete implementation of the RESTful service. Data is stored in content addressable storage (CAS). Session token and URL addresses for the results page are generated based on the query requests. We use SHA-1 to calculate the token from the requested data. This approach has several advantages. It is easy to implement and use and it allows the server to reuse previously calculated queries. The most frequent occurrence of this mechanism is the situation when the user starts an analysis and then closes the web page without copying the link/token for their job. As the user session ID is based on a token calculated from the query, it may be used to find this job if the same user submits the same query again.

The architecture design lets the user use the RESTful API directly, allowing programmable access to the PlaToLoCo server without using the convenient but restrictive web interface. Example commands in Python are also available on the webserver's API page.

CASE STUDY

PlaToLoCo is the first platform that enables the user to retrieve and examine multiple annotations of low complexity sequences at the same time. Different predictors featured in PlaToLoCo provide diversified perspectives on the problem of identification of low complexity regions. Having these different perspectives together allows a diverse coverage of cases of low complexity. This concept is illustrated here through several examples.

Although LCRs are most commonly associated with intrinsically disordered regions (IDRs), there are a handful of exceptions that form stable structures. One exception is the set of α-helical transmembrane proteins that

Figure 1. Layout of the results tab of the PlaToLoCo webserver. (**A**) summary tab, (**B**) sequence details, (**C**) amino acid frequency, (**D**) methods consensus, (**E**) Pfam and PDB details and (**F**) region details.

cross the lipid bilayer one or more times, which is unfavorable for (partially) charged functional groups, resulting in a skewed amino acid distribution. For example, steryl-sulfatase (STS_HUMAN, UniProt AC P08842) has two transmembrane regions (38,39), both detected by SEG, CAST and fLPS. Although these methods are unable to precisely detect region boundaries, it is clear that LCRs and IDRs are neither identical nor interchangable (Figure 2A).

Another example is PSPC1 (PSPC1_HUMAN, UniProt AC Q8WXF1), which multimerizes in multiple steps: the middle region of the protein forms a dimer with NONO or SFPQ via an unusual right-handed coiled-coil domain with hendecad motifs, elongated with a single α-helix. The second step has not been completely characterized, although a transition from a single α-helix to a weak coiled-coil with other PSPC1/NONO dimers has been proposed (40). Both the coiled-coils and the single α-helices have unusual residue distributions and the main forces driving the coiled-coil interactions are hydrophobic contacts supported via electrostatic interactions which results in a repeat pattern of 7 or 11 residues (41), while the single α-helices also rely on repeated charged residues that form salt bridges (42). These regions are predicted (35,36) to overlap with the LCRs and this has been confirmed by X-ray studies (43,44), and theoretical modelling (40) of the multimer (Figure 2B).

Another example of low complexity is a region of the PI3-kinase p85 subunit (P85A_HUMAN, UniProt AC P27986) that binds to an SH3 domain forming a stable structure (45). This peptide contains a significant number of prolines, likely to decrease the loss of entropy during binding. The interacting region is predicted to be an IDR (37) and is identified by several LCR tools. Moreover, the consensus of the SEG, CAST and fLPS results seems to identify the proline-rich interacting segment precisely (Figure 2C).

The inhibitory domain of the mineralocorticoid receptor is another fine example of how low complexity regions play a role in preserving structure (46). Likewise, Huntingtin (UniProt AC P42858) contains a poly-glutamine (polyQ) homorepeat that causes the disease phenotype when there are more than 36 glutamines present. Huntingtin has a

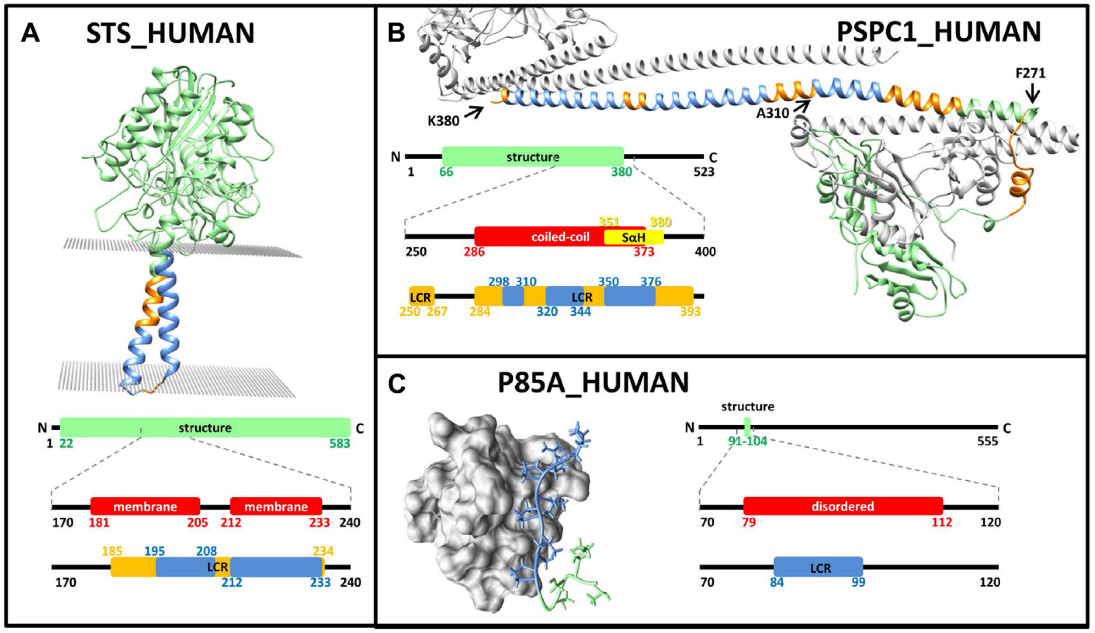

Figure 2. (**A**) The structure of steryl-sulfatase (PDB: 1p49). Membrane regions were predicted by CCTOP (red) (34). (**B**) The modelled structure of paraspeckle component multimers. Coiled-coil regions were predicted by DeepCoil (red) (35), Single-helices were predicted by CSAHdetec (yellow) (36). (**C**) Structure of SH3 domain and the p85 subunit of PI3-kinase. Disordered regions were predicted with IUPRED (red) (37). Blue: overlap of SEG, CAST and fLPS. Orange: overlap of CAST and fLPS – also highlighted on the structures.

stretch of glutamines from residue 18 to 38 and all predictors in PlaToLoCo correctly identify a reduction in sequence complexity around this region. Although, different predictors find other and different LCRs along the sequence of huntingtin.

Another protein, yeast RNA polymerase II subunit RPB1 (UniProt AC P04050) is essential to the formation of the RNA Polymerase II complex. This protein contains a sequence repeat in its C-terminus (residues 1549–1716). All methods in PlaToLoCo except SIMPLE identify this region as an LCR. GBSC in particular detects many subregions (different repeats), providing a higher level of granularity. This, together with the prediction from the Pfam database (reported in the Feature Viewer), helps in identifying the type of low complexity and provides additional information to other methods predictions.

Med1 (UniProt AC Q15648) is a component of the mediator complex and is involved in the transcriptional regulation of proteins that depend on RNA polymerase II. This protein is highly compositionally biased and contains intrinsically disordered regions from around residue 750 to its C-terminus. PlaToLoCo identifies LCRs from residue 550 onward, with CAST giving the whole region down to its C-terminus as a serine-rich region, while other predictors find more specific and sparse subregions of LC. In the feature viewer, the user gets an immediate and intuitive view of the differences between the methods' predictions.

RNA binding protein FUS (UniProt AC P35637) is a mediator of many processes, namely RNA, transport, splicing and DNA repair. It is also involved in several diseases including amyotrophic lateral sclerosis (ALS) (47,48). FUS has long low complexity IDRs and can form liquid droplets (49–51) via liquid-liquid phase separation. PlaToLoCo predictors almost unequivocally agree that this protein is almost completely composed of low complexity regions, with the exception of a central region (from residue 260 to 370). Predictors identify the N-terminal low complexity domain of FUS, a highly conserved prion-like domain composed primarily of serine, tyrosine, glycine and glutamine rich regions.

The probable serine/threonine-protein kinase fhkB is a reviewed protein in UniProt (Q1ZXH2) from *Dictyostelium discoideum*. Two regions are annotated as domains and the rest of the protein is either considered a coiled coil or compositionally biased. PlaToLoCo gives greater insight into these regions. The differences between predictors are evidenced in Figure 1B. fLPS finds low complexity regions in the middle of the sequence while the other methods focus their prediction on the N- and C-termini, more or less where the domains from UniProt are annotated. CAST identifies two continuous regions at the termini, while the other predictors give a more fragmented and detailed prediction. Since low complexity does not have a unique definition, having multiple takes on the problem allows the user to explore the phenomena at the desired level of detail.

To have an overall picture of the distribution of LCRs, we ran PlaToLoCo on different proteomes (obtained from UniProt, March 2019, with cd-hit (52) 40% redundancy control) and used the PDB (non-redundant PDB chain set, March 2019, *P*-value cutoff of 10e−7) as a control. In general, LCRs and CBRs were more commonly detected compared to repeat regions (Figure 3). However different proteomes tended to have different varieties of LCRs. For example, fLPS identified roughly the same proportion of CB segments in all the genomes, while CAST predicted an unusually high proportion of CB segments in the *Plasmodium*

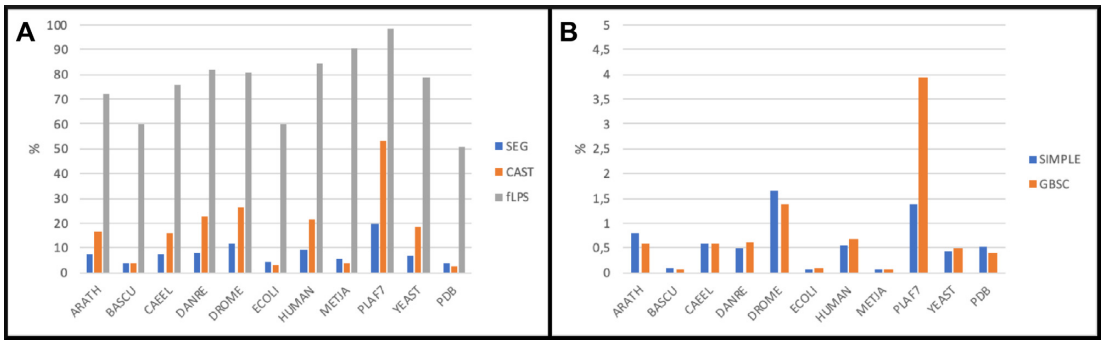

Figure 3. Panel **A**: proportion of detected low complexity and compositionally biased regions in various proteomes and in the PDB (detected LCRs to total number of residues) for SEG, CAST and fLPS. Panel **B**: proportion of repeat regions in various proteomes and in the PDB (detected repeats to total number of residues) for SIMPLE and GBSC. The exact number of residues found by each method is provided in the Supplementary material 2 (Suppl2) in Supplementary Table S2.

falciparum genome. Interestingly, this parasite also seemed to have a higher frequency of repeats that were detected by GBSC and SIMPLE as compared to the other test genomes. We also analyzed several other performance measures such as method run time, the total number of residues found by each method, and the detected overlap between methods. These statistics are provided in the Supplementary material 2 (Suppl2).

DISCUSSION

For many years low complexity regions were thought to be *junk* part of proteome. This resulted in a reduced development of tools and methods for their analysis. Most of the algorithms that have been developed to identify regions with LCRs or amino acid biases were developed in previous decades, and most of them are available only as standalone applications. Usually they are difficult or sometimes even impossible to install as they require outdated package dependencies. Furthermore, they typically provide only a list of discovered regions without any functional annotation. By developing the PlaToLoCo meta-server, for the first time, we provide the community with a fast and easily accessible tool for the analysis of LCRs, enriching them with additional information to aid the interpretation of the results.

In the future we plan to improve PlaToLoCo by providing additional functional annotations such as annotated protein regions from UniProt or predictions of protein disorder. We also plan to run PlaToLoCo against full proteomes and complete protein databases as having precomputed results for all UniProtKB proteins (or initially at least SwissProt entries) would reduce the execution time dramatically when submitting a UniProt AC as input.

Examining the overlap between the results of selected methods we notice that the output results may vary significantly when different methods are compared. This supports the conclusion that the PlaToLoCo platform provides a selection of methods that are designed to detect different types of LCRs. Since there is no single, universally accepted definition of LCRs and various methods attempt to analyse the statistical properties of the sequences or specific sequence patterns, it is important to provide the scientific community with a tool that allows the analysis of low complexity regions from different perspectives and on different granularity levels, covering a wide range of approaches to detect these regions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Grigoris D. Amoutzias, University of Thessaly, for discussing desirable features of PlaToLoCo during the early stages of this work.

FUNDING

The idea for this article was developed during the meeting of the EU COST-Action BM1405 'Non-globular proteins: from sequence to structure, function and application in molecular physiopathology'. European Union's Horizon 2020 research and innovation program [778247, 823886 to Z.D.,P.M.,M.A.A.N,L.P.,M.N.,D.P.,S.C.E.T]; European Union through the European Social Fund [POWR.03.02.00-00-I029/17 to P.J.]; M.M. acknowledges funding from the National Science Centre, Poland [2014/15/D/NZ1/00968]; Deutsche Forschungsgemeinschaft [AN735/4-1 to M.A.A.N.]; Z.D. acknowledges funding from the ELTE Thematic Excellence Programme (ED-18-1-2019-0030) supported by the Hungarian Ministry for Innovation and Technology. Funding for open access charge: Rector's research and development grant; Silesian University of Technology [02/120/RGJ20/0002] and European Social Fund [POWR.03.02.00-00-I029/17].
Conflict of interest statement. None declared.

REFERENCES

1. Mier,P., Paladin,L., Tamana,S., Petrosian,S., Hajdu-Soltész,B., Urbanek,A., Gruca,A., Plewczynski,D., Grynberg,M., Bernadó,P. *et al.* (2020) Disentangling the complexity of low complexity proteins. *Brief. Bioinform.*, **21**, 458–472.
2. Lovell,S.C. (2003) Are non-functional, unfolded proteins ('junk proteins') common in the genome? *FEBS Lett.*, **554**, 237–239.
3. Perdigão,N., Heinrich,J., Stolte,C., Sabir,K.S., Buckley,M.J., Tabor,B., Signal,B., Gloss,B.S., Hammang,C.J., Rost,B. *et al.* (2015)

Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 15898–15903.
4. Ntountoumi,C., Vlastaridis,P., Mossialos,D., Stathopoulos,C., Iliopoulos,I., Promponas,V., Oliver,S.G. and Amoutzias,G.D. (2019) Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved. *Nucleic Acids Res.*, **47**, 9998–10009.
5. Luo,H. and Nijveen,H. (2014) Understanding and identifying amino acid repeats. *Brief. Bioinformatics*, **15**, 582–591.
6. So,C.R., Fears,K.P., Leary,D.H., Scancella,J.M., Wang,Z., Liu,J.L., Orihuela,B., Rittschof,D., Spillmann,C.M. and Wahl,K.J. (2016) Sequence basis of Barnacle Cement Nanostructure is Defined by Proteins with Silk Homology. *Sci. Rep.*, **6**, 36219.
7. Brewer,S., Tolley,M., Trayer,I.P., Barr,G.C., Dorman,C.J., Hannavy,K., Higgins,C.F., Evans,J.S., Levine,B.A. and Wormald,M.R. (1990) Structure and function of X-Pro dipeptide repeats in the TonB proteins of Salmonella typhimurium and Escherichia coli. *J. Mol. Biol.*, **216**, 883–895.
8. Robison,A.D., Sun,S., Poyton,M.F., Johnson,G.A., Pellois,J.-P., Jungwirth,P., Vazdar,M. and Cremer,P.S. (2016) Polyarginine interacts more strongly and cooperatively than polylysine with phospholipid bilayers. *J. Phys. Chem. B.*, **120**, 9287–9296.
9. Kushwaha,A.K. and Grove,A. (2013) C-terminal low-complexity sequence repeats of Mycobacterium smegmatis Ku modulate DNA binding. *Biosci. Rep.*, **33**, 175–184.
10. Karlin,S. and Zhu,Z.Y. (1996) Characterizations of diverse residue clusters in protein three-dimensional structures. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 8344–8349.
11. Martin,E.W. and Mittag,T. (2018) Relationship of sequence and phase separation in protein low-complexity regions. *Biochemistry*, **57**, 2478–2487.
12. Babu,M.M. (2016) The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.*, **44**, 1185–1200.
13. Harrison,A.F. and Shorter,J. (2017) RNA-binding proteins with prion-like domains in health and disease. *Biochem. J.*, **474**, 1417–1438.
14. Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
16. Promponas,V.J., Enright,A.J., Tsoka,S., Kreil,D.P., Leroy,C., Hamodrakas,S., Sander,C. and Ouzounis,C.A. (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. *Bioinformatics*, **16**, 915–922.
17. Harrison,P.M. (2017) fLPS: Fast discovery of compositional biases for the protein universe. *BMC Bioinformatics*, **18**, 476.
18. Albà,M.M., Laskowski,R.A. and Hancock,J.M. (2002) Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics*, **18**, 672–678.
19. Kirmitzoglou,I. and Promponas,V.J. (2015) LCR-eXXXplorer: a web platform to search, visualize and share data for low complexity regions in protein sequences. *Bioinformatics*, **31**, 2208–2210.
20. The,UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
21. Smith,T. and Waterman,M. (1981) Identification of common molecular subsequences. *Mol. Biol.*, **147**, 195–197.
22. Tautz,D., Trick,M. and Dover,G.A. (1986) Cryptic simplicity in DNA is a major source of genetic variation. *Nature*, **322**, 652–656.
23. Hancock,J.M. and Armstrong,J.S. (1994) SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Bioinformatics*, **10**, 67–70.
24. Harbi,D., Kumar,M. and Harrison,P.M. (2011) LPS-annotate: complete annotation of compositionally biased regions in the protein knowledgebase. *Database*, **2011**, baq031.
25. Harrison,P.M. and Gerstein,M. (2003) A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes. *Genome Biol.*, **4**, R40.
26. Huntley,M. and Golding,G. (2002) Simple sequences are rare in the Protein Data Bank. *Proteins*, **48**, 134–140.
27. Radó-Trilla,N. and Albà,M. (2012) Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. *BMC Evol. Biol.*, **12**, 155.
28. Zahn-Zabal,M., Michel,P.-A., Gateau,A., Nikitin,F., Schaeffer,M., Audot,E., Gaudet,P., Duek,P.D., Teixeira,D., Rech de Laval,V. *et al.* (2020) The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Res.*, **48**, D328–D334.
29. Hatos,A., Hajdu-Soltész,B., Monzon,A.M., Palopoli,N., Álvarez,L., Aykac-Fas,B., Bassot,C., Benítez,G.I., Bevilacqua,M., Chasapi,A. *et al.* (2020) DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.*, **48**, D269–D276.
30. Burley,S.K., Berman,H.M., Bhikadiya,C., Bi,C., Chen,L., Di Costanzo,L., Christie,C., Dalenberg,K., Duarte,J.M., Dutta,S. *et al.* (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.*, **47**, D464–D474.
31. Käll,L., Krogh,A. and Sonnhammer,E. L.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
32. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
33. Paladin,L., Schaeffer,M., Gaudet,P., Zahn-Zabal,M., Michel,P.-A., Piovesan,D., Tosatto,S. C.E. and Bairoch,A. (2020) The Feature-Viewer: a visualization tool for positional annotations on a sequence. *Bioinformatics*, doi:10.1093/bioinformatics/btaa055.
34. Dobson,L., Reményi,I. and Tusnády,G.E. (2015) CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res.*, **43**, W408–W412.
35. Ludwiczak,J., Winski,A., Szczepaniak,K., Alva,V. and Dunin-Horkawicz,S. (2019) DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics*, **35**, 2790–2795.
36. Dudola,D., Tóth,G., Nyitray,L. and Gáspári,Z. (2017) Consensus prediction of charged single alpha-helices with CSAHserver. *Methods Mol. Biol.*, **1484**, 25–34.
37. Mészáros,B., Erdos,G. and Dosztányi,Z. (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.*, **46**, W329–W337.
38. Dobson,L., Reményi,I. and Tusnády,G.E. (2015) The human transmembrane proteome. *Biol. Direct*, **10**, 31.
39. Hernandez-Guzman,F.G., Higashiyama,T., Pangborn,W., Osawa,Y. and Ghosh,D. (2003) Structure of human estrone sulfatase suggests functional roles of membrane association. *J. Biol. Chem.*, **278**, 22989–22997.
40. Dobson,L., Nyitray,L. and Gáspári,Z. (2015) A conserved charged single α-helix with a putative steric role in paraspeckle formation. *RNA*, **21**, 2023–2029.
41. Burkhard,P., Stetefeld,J. and Strelkov,S.V. (2001) Coiled coils: a highly versatile protein folding motif. *Trends Cell Biol.*, **11**, 82–88.
42. Süveges,D., Gáspári,Z., Toth,G. and Nyitray,L. (2009) Charged single α-helix: a versatile protein structural motif. *Proteins: Struct. Funct. Bioinf.*, **74**, 905–916.
43. Lee,M., Sadowska,A., Bekere,I., Ho,D., Gully,B.S., Lu,Y., Iyer,K.S., Trewhella,J., Fox,A.H. and Bond,C.S. (2015) The structure of human SFPQ reveals a coiled-coil mediated polymer essential for functional aggregation in gene regulation. *Nucleic Acids Res.*, **43**, 3826–3840.
44. Passon,D.M., Lee,M., Rackham,O., Stanley,W.A., Sadowska,A., Filipovska,A., Fox,A.H. and Bond,C.S. (2012) Structure of the heterodimer of human NONO and paraspeckle protein component 1 and analysis of its role in subnuclear body formation. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 4846–4850.
45. Renzoni,D.A., Pugh,D.J., Siligardi,G., Das,P., Morton,C.J., Rossi,C., Waterfield,M.D., Campbell,I.D. and Ladbury,J.E. (1996) Structural and thermodynamic characterization of the interaction of the SH3 domain from Fyn with the proline-rich binding site on the p85 subunit of PI3-kinase. *Biochemistry*, **35**, 15646–15653.
46. Vlassi,M., Brauns,K. and Andrade-Navarro,M.A. (2013) Short tandem repeats in the inhibitory domain of the mineralocorticoid receptor: prediction of a β-solenoid structure. *BMC Struct. Biol.*, **13**, 17.

47. Vance,C., Rogelj,B., Hortobágyi,T., De Vos,K.J., Nishimura,A.L., Sreedharan,J., Hu,X., Smith,B., Ruddy,D., Wright,P. *et al.* (2009) Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science*, **323**, 1208–1211.
48. Kwiatkowski,T.J. Jr, Bosco,D.A., Leclerc,A.L., Tamrazian,E., Vanderburg,C.R., Russ,C., Davis,A., Gilchrist,J., Kasarskis,E.J., Munsat,T. *et al.* (2009) Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science*, **323**, 1205–1208.
49. Burke,K.A., Janke,A.M., Rhine,C.L. and Fawzi,N.L. (2015) Residue-by-residue view of in vitro FUS granules that bind the C-terminal domain of RNA polymerase II. *Mol. Cell*, **60**, 231–241.
50. Molliex,A., Temirov,J., Lee,J., Coughlin,M., Kanagaraj,A.P., Kim,H.J., Mittag,T. and Taylor,J.P. (2015) Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell*, **163**, 123–133.
51. Patel,A., Lee,H.O., Jawerth,L., Maharana,S., Jahnel,M., Hein,M.Y., Stoynov,S., Mahamid,J., Saha,S., Franzmann,T.M. *et al.* (2015) A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell*, **162**, 1066–1077.
52. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.