Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Applications Note

Sequence analysis

The LCD-Composer Webserver: High-Specificity Identification and Functional Analysis of Low-Complexity Domains in Proteins

Sean M. Cascarina[1,*] and Eric D. Ross[1,*]

[1]Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, CO 80523, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Low-complexity domains (LCDs) in proteins are regions enriched in a small subset of amino acids. LCDs exist in all domains of life, often have unusual biophysical behavior, and function in both normal and pathological processes. We recently developed an algorithm to identify LCDs based predominantly on amino acid composition thresholds. Here, we have integrated this algorithm with a webserver and augmented it with additional analysis options. Specifically, users can: 1) search for LCDs in whole proteomes by setting minimum composition thresholds for individual or grouped amino acids, 2) submit a known LCD sequence to search for similar LCDs, 3) search for and plot LCDs within a single protein, 4) statistically test for enrichment of LCDs within a user-provided protein set, and 5) specifically identify proteins with multiple types of LCDs.

Availability: The LCD-Composer server can be accessed at http://lcd-composer.bmb.colostate.edu. The corresponding command-line scripts can be accessed at https://github.com/RossLabCSU/LCD-Composer/tree/master/WebserverScripts.

Contact: Sean.Cascarina@colostate.edu or Eric.Ross@colostate.edu

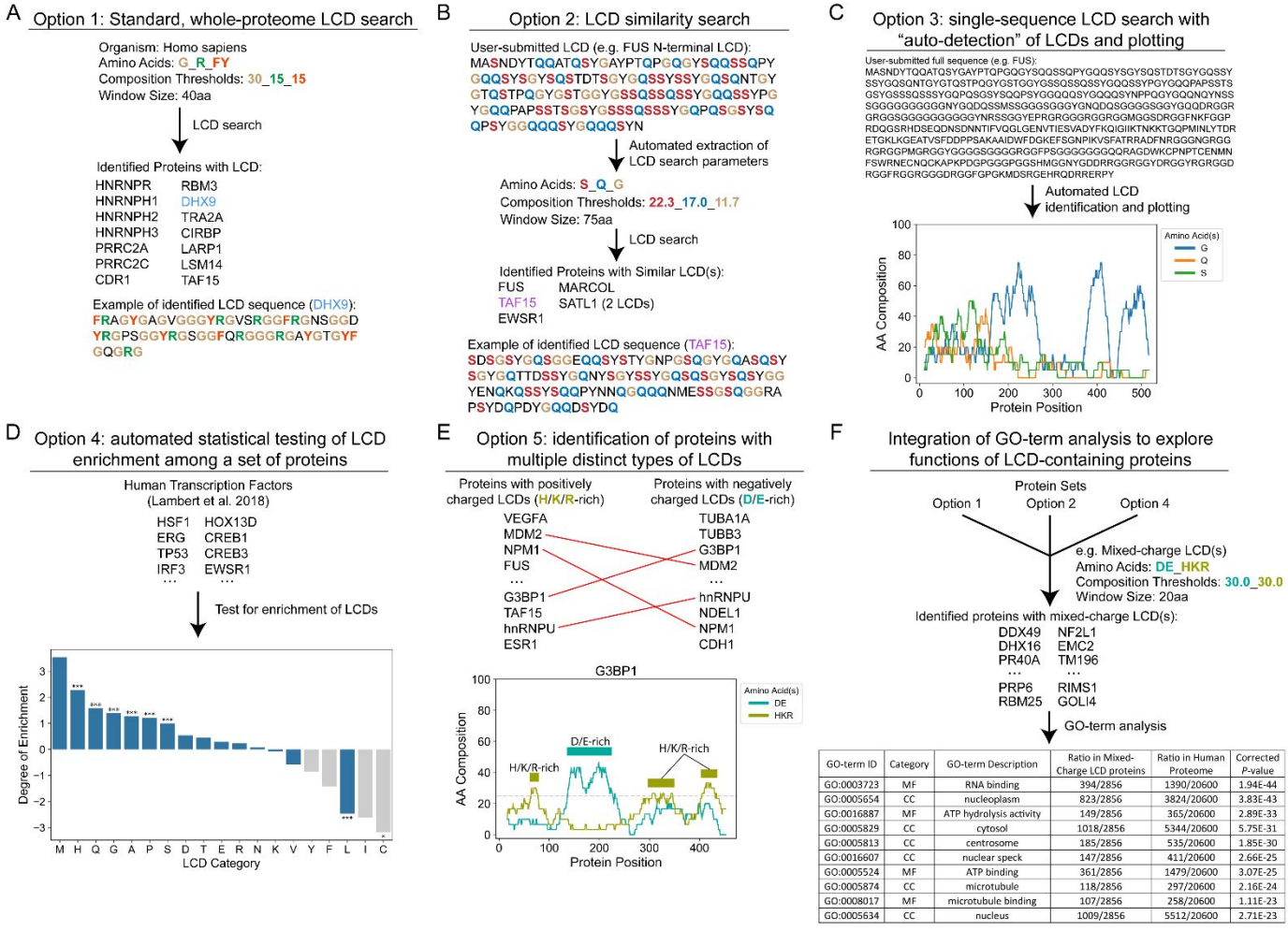
1 Introduction

Low-complexity domains (LCDs) in proteins are regions in which a small subset of amino acids comprise an unusually large percentage of that region. LCDs that are enriched in different amino acids tend to be associated with distinct molecular functions and biophysical behavior, which may contribute to their functional specialization. Consequently, local enrichment of one or more specific amino acids provides a direct and intuitive way to identify/classify LCDs (Cascarina and Ross, 2018; Cascarina, King, *et al.*, 2021). Additionally, LCDs have been associated with a variety of human diseases including cancer (Lu *et al.*, 2021), neuromuscular disorders (Harrison and Shorter, 2017), and pathogenic infections (Davies *et al.*, 2017; Cascarina and Ross, 2022).

2 The LCD-Composer Webserver

Our low-complexity domain composition scanner (LCD-Composer) offers user control over multiple search parameters, including sliding window size, amino acid(s) to use as defining features to identify LCDs, minimum composition thresholds associated with the defining amino acid(s), and minimum spacing of the defining amino acid(s). These parameters are described on the server "Help" page and in Cascarina, King, *et al.* (2021). LCD searches can be based on simple criteria or on multiple criteria simultaneously (e.g. LCDs with combined serine/threonine composition ≥50% *and* combined phenylalanine/tryptophan/tyrosine composition ≥20%). LCD searches can be performed on UniProt reference proteomes or individual, user-defined proteins. The corresponding command-line scripts can be run on any FASTA-formatted proteome. The LCD-Composer server provides several options for performing LCD searches (Fig 1), summarized in the sections below.

S.M. Cascarina and E.D. Ross



GO-term ID	Category	GO-term Description	Ratio in Mixed-Charge LCD proteins	Ratio in Human Proteome	Corrected *P*-value
GO:0003723	MF	RNA binding	394/2856	1390/20600	1.94E-44
GO:0005654	CC	nucleoplasm	823/2856	3824/20600	3.83E-43
GO:0016887	MF	ATP hydrolysis activity	149/2856	365/20600	2.89E-33
GO:0005829	CC	cytosol	1018/2856	5344/20600	5.75E-31
GO:0005813	CC	centrosome	185/2856	535/20600	1.85E-30
GO:0016607	CC	nuclear speck	147/2856	411/20600	2.66E-25
GO:0005524	MF	ATP binding	361/2856	1479/20600	3.07E-25
GO:0005874	CC	microtubule	118/2856	297/20600	2.16E-24
GO:0008017	MF	microtubule binding	107/2856	258/20600	1.11E-23
GO:0005634	CC	nucleus	1009/2856	5512/20600	2.71E-23

Fig 1. Examples of LCD-Composer analyses using each option.

3.1 Option 1

Users can customize search parameters and perform an LCD search on a selected proteome, with the option of limiting the search to a single representative isoform for each protein or including all known isoforms for the corresponding organism (Fig 1A). This option is equivalent to running the LCD-Composer command-line script.

3.2 Option 2

Users can submit a "query" LCD sequence of their choosing: LCD-Composer then searches for compositionally similar LCDs by extracting search parameters from the query sequence (Fig 1B). Users can choose the number of compositional features (with a maximum of 4) to use as search parameters. Identified LCDs are automatically ranked according to compositional similarity to the user-submitted query sequence (defined as the normalized Manhattan distance between amino acid compositions of each identified LCD compared to the query LCD). LCD searches can be performed across organisms (e.g. a query LCD sequence from yeast can be used to search for similar LCDs in humans).

3.3 Option 3

Users can submit a single protein sequence and search for LCDs using customizable search parameters. This option is particularly useful when defining LCDs based on multiple simultaneous criteria, which can be challenging without a quantitative definition of domain boundaries. Additionally, this option offers automated, publication-quality plotting of amino acid composition as a function of protein position using our CompositionPlotter algorithm [(Cascarina, Kaplan, *et al.*, 2021); Fig 1C].

3.4 Option 4

Users can submit a list of proteins and test for statistical enrichment/depletion of certain types of LCDs within that protein set. Enrichment tests can be specific to a user-defined LCD type (e.g. testing if Q/N-rich LCDs are enriched among transcription factors) or can be performed "naïvely", where enrichment tests are automatically performed for the 20 canonical LCD classes [Fig 1D; (Lambert *et al.*, 2018)].

3.5 Option 5

Proteins containing multiple distinct types of LCDs ("co-occurring" LCDs) can be specifically associated with certain functions (Cascarina, King, *et al.*, 2021). For example, users may be interested in proteins that contain both a positively charged LCD and a negatively charged LCD

(Fig 1E). After searching for multiple types of LCDs, users can submit the results of each search for comparison: only proteins that contain every type of user-defined LCD are returned to the user.

3.6 Automated GO-term analysis for LCD-containing proteins

Molecular functions of LCD-containing proteins can be remarkably LCD-type-specific (Cascarina, King, *et al.*, 2021), suggesting the existence of functional niches for each type of LCD. Therefore, the LCD-Composer webserver also offers automated gene ontology (GO)-term analysis that can be performed in conjunction with LCD searches (Fig 1F), using the user-selected UniProt proteome to define the background set of proteins for enrichment analyses.

3.7 The LCD-Composer server complements existing LCD servers

Existing servers designed to identify LCDs or evaluate compositional biases in protein sequences include ProBias (Kuznetsov, 2008), LCR-eXXXplorer (Kirmitzoglou and Promponas, 2015), LCR-hound (Ntountoumi *et al.*, 2019), PlaToLoCo (Jarnot *et al.*, 2020), SAPS [(Brendel *et al.*, 1992); currently available at https://www.ebi.ac.uk/Tools/seqstats/saps], and Composition Profiler (Vacic *et al.*, 2007). In general, these servers excel at LCD identification and/or whole-protein composition analyses in the absence of pre-defined LCD features of interest. LCD-Composer complements these servers by enabling high-specificity LCD identification at whole-proteome scale with pre-defined compositional features of interest or with search parameters calculated from a user-submitted query LCD sequence.

3.8 High-throughput analyses using command-line scripts

We also release command-line versions of each option available on the LCD-Composer server, which enable high-throughput analyses of multiple proteomes or LCD types, as well as analyses of alternative proteomes not supported by the server.

Acknowledgements

We thank Michael P. Hughes for critical advice on webserver architecture, and Kacy R. Paul for server testing during initial development.

Funding

This work was supported by the National Institute of General Medical Sciences [grant number R35GM130352 to E.D.R.]; and the National Science Foundation [grant number MCB-1817622 to E.D.R.].

Conflict of Interest: none declared.

References

Brendel,V. *et al.* (1992) Methods and algorithms for statistical analysis of protein sequences. *Proc. Natl. Acad. Sci.*, **89**, 2002.

Cascarina,S.M., Kaplan,J.P., *et al.* (2021) Generalizable compositional features influencing the proteostatic fates of polar low-complexity domains. *Int. J. Mol. Sci.*, **22**, 8944.

Cascarina,S.M., King,D.C., *et al.* (2021) LCD-Composer: an intuitive, composition-centric method enabling the identification and detailed functional mapping of low-complexity domains. *NAR Genom. Bioinform.*, **3**, lqab048.

Cascarina,S.M. and Ross,E.D. (2022) Phase Separation by the SARS-CoV-2 Nucleocapsid Protein: Consensus and Open Questions. *J. Biol. Chem.*, 101677.

Cascarina,S.M. and Ross,E.D. (2018) Proteome-scale relationships between local amino acid composition and protein fates and functions. *PLOS Comput. Biol.*, **14**, e1006256.

Davies,H.M. *et al.* (2017) Repetitive sequences in malaria parasite proteins. *FEMS Microbiol. Rev.*, **41**, 923–940.

Harrison,A.F. and Shorter,J. (2017) RNA-binding proteins with prion-like domains in health and disease. *Biochem. J.*, **474**, 1417–1438.

Jarnot,P. *et al.* (2020) PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins. *Nucleic Acids Res.*, **48**, W77–W84.

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.

Kuznetsov,I.B. (2008) ProBias: a web-server for the identification of user-specified types of compositionally biased segments in protein sequences. *Bioinformatics*, **24**, 1534–1535.

Lambert,S.A. *et al.* (2018) The Human Transcription Factors. *Cell*, **172**, 650–665.

Lu,J. *et al.* (2021) Emerging Roles of Liquid–Liquid Phase Separation in Cancer: From Protein Aggregation to Immune-Associated Signaling. *Front. Cell Dev. Biol.*, **9**, 631486.

Ntountoumi,C. *et al.* (2019) Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved. *Nucleic Acids Res.*, **47**, 9998–10009.

Vacic,V. *et al.* (2007) Composition Profiler: A tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics*, **8**, 211.