Sequence analysis

The LCD-Composer webserver: high-specificity identification and functional analysis of low-complexity domains in proteins

Sean M. Cascarina ☑ * and Eric D. Ross*

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

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on May 6, 2022; revised on October 12, 2022; editorial decision on October 19, 2022; accepted on October 21, 2022

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 (i) search for LCDs in whole proteomes by setting minimum composition thresholds for individual or grouped amino acids, (ii) submit a known LCD sequence to search for similar LCDs, (iii) search for and plot LCDs within a single protein, (iv) statistically test for enrichment of LCDs within a user-provided protein set and (v) specifically identify proteins with multiple types of LCDs.

Availability and implementation: 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 et al., 2021b). 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 (Cascarina and Ross, 2022; Davies et al., 2017).

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 et al. (2021b). 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 proteomes. 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.

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 four) 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 et al., 2021a; 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 ‘naively’, 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 et al., 2021b]. 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 et al., 2021b], 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-exXplorer (Kirmizigolou 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 a 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 [R35GM130352 to E.D.R.]; and the National Science Foundation [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. USA, 89, 2002–2006.
Cascarina, S.M. et al. (2021a) Generalizable compositional features influencing the proteostatic fates of polar low-complexity domains. Int. J. Mol. Sci., 22, 8944.
Cascarina, S.M. et al. (2021b) LCD-Composer: an intuitive, composition-centric method enabling the identification and detailed functional mapping of low-complexity domains. NAR Genom. Bioinformatics, 3, lka048.
Cascarina, S.M. and Ross, E.D. (2022) Phase separation by the SARS-CoV-2 nucleocapsid protein: consensus and open questions. J. Biol. Chem., 298, 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.