Bioinformatics, 31(13), 2015, 2208–2210
doi: 10.1093/bioinformatics/btv115
Advance Access Publication Date: 20 February 2015
Applications Note

OXFORD

Sequence analysis

LCR-eXXXplorer: a web platform to search, visualize and share data for low complexity regions in protein sequences

Ioannis Kirmitzoglou[†] and Vasilis J. Promponas*

Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, CY 1678, Nicosia, Cyprus

*To whom correspondence should be addressed.
[†]Present address: Department of Life Sciences, Imperial College London, London, UK
Associate Editor: John Hancock

Received on July 21, 2014; revised on January 31, 2015; accepted on February 17, 2015

Abstract

Motivation: Local compositionally biased and low complexity regions (LCRs) in amino acid sequences have initially attracted the interest of researchers due to their implication in generating artifacts in sequence database searches. There is accumulating evidence of the biological significance of LCRs both in physiological and in pathological situations. Nonetheless, LCR-related algorithms and tools have not gained wide appreciation across the research community, partly due to the fact that only a handful of user-friendly software is currently freely available.
Results: We developed LCR-eXXXplorer, an extensible online platform attempting to fill this gap. LCR-eXXXplorer offers tools for displaying LCRs from the UniProt/SwissProt knowledgebase, in combination with other relevant protein features, predicted or experimentally verified. Moreover, users may perform powerful queries against a custom designed sequence/LCR-centric database. We anticipate that LCR-eXXXplorer will be a useful starting point in research efforts for the elucidation of the structure, function and evolution of proteins with LCRs.
Availability and implementation: LCR-eXXXplorer is freely available at the URL http://repeat.biol.ucy.ac.cy/lcr-exxxplorer.
Contact: vprobon@ucy.ac.cy
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

During the past 30 years, the main focus of research related to regions of local compositional extremes (low complexity regions; LCRs) was their identification for the purpose of sequence masking (Altschul *et al.*, 1994; Wootton and Federhen, 1993; Ye *et al.*, 2006) for eliminating spurious hits in database searches (Promponas *et al.*, 2000; Tsoka *et al.*, 1999). Several studies have been published showcasing the abundance and importance of such regions at the molecular/structural (e.g. Radivojac *et al.*, 2006; Tamana *et al.*, 2012), functional (e.g. Andrade *et al.*, 2001; Haerty and Golding, 2010), organismic (e.g. Miskinyte *et al.*, 2013; Pizzi and Frontali, 2001) and habitat level (e.g. Nandi *et al.*, 2003). Despite the apparent biological importance of LCRs there's a distinct lack of tools or services capable of helping biologists to study them in depth. Most of the methods capable of detecting LCRs were developed for the sole purpose of masking them and are meant to be used from the command line as part of a sequence analysis or search pipeline. While some tools, such as SEG (Wootton and Federhen, 1993), CAST (Promponas *et al.*, 2000) or BIAS (Kuznetsov and Hwang, 2006) do offer more advanced reports as an option, their results are mostly meant to be parsed by a computer software and not a biologist.

In this work, we present LCR-eXXXplorer, an online service to search, visualize and share LCRs in protein sequences. We highlight

its unique features that may facilitate research efforts towards understanding the biological roles of proteins with LCRs.

2 Functionality

2.1 General description

LCR-eXXXplorer is built upon a customized instance of GBrowse (Stein *et al.*, 2002) modified to properly work with protein sequences. It currently contains 545 000 sequences (retrieved from UniProt/SwissProt) annotated with over 16 million LCR-related annotations. Along with information about sequence complexity, LCR-eXXXplorer displays external annotations from UniProt, as well as predicted disordered and binding regions by utilizing IUPRED (Dosztányi *et al.*, 2005) and ANCHOR (Dosztányi *et al.*, 2009; Mészáros *et al.*, 2009) respectively. Data are stored in a MySQL database, using a database schema based on the SeqFeature schema internally used by GBrowse (see Supplementary Methods and Supplementary Fig. S1).

2.2 Key functionality

A basic keyword-based search functionality (allowing wildcards) is available for retrieving protein sequences with matching UniProtKB Accession(s)/Entry Name(s) or gene name(s). Moreover, the 'Advanced Search' option (specifically implemented for this process as a custom-made GBrowse plug-in) facilitates more fine-tuned queries. Using the basic search mode, users are able to retrieve up to 500 entries using simple keyword search (e.g. with a single UniProt identifier or accession number). An 'Advanced Search' may be initiated by querying a suitable combination of UniProt fields (e.g. gene or protein name, source organism) or LCR properties (e.g. type of LCR, percent of masked residues)—yet, only the AND Boolean operator is currently supported for combining search criteria. Under this mode, batch search functionality is also available using a list of UniProt accession numbers: this feature enables users to take advantage of the powerful UniProt search engine and come up with a list of entries specifying complex search criteria. Results can be displayed in the browser (with a limit of 15 000 entries) or downloaded in a plain text tab-delimited formatted file providing statistics on the LCR content for further processing (with a limit of 50 000 entries). Different options of masking protein sequences are provided for each individual sequence from the graphical GBrowse 'protein details' view and sequences are available in FASTA format.

The Downloads section offers LCR-eXXXplorer the option of downloading the complete set of sequences in FASTA formatted files masked for LCRs, the complete set of annotations in GFF3 format or a CSV formatted table with LCR statistics for each sequence in the database.

Users may also search for data in LCR-eXXXplorer using BLASTP (Ye *et al.*, 2006) powered by the user-friendly SequenceServer (Priyam *et al.*, manuscript in preparation). Three underlying databases (unmasked, SEG or CAST masking with default parameters) are provided, with the masked databases being a unique feature of this service; this configuration is shown to improve database search results (Kirmitzoglou, 2014; Kirmitzoglou *et al.*, in preparation). Furthermore, users may initiate BLASTP searches against the sequence databases hosted at the NCBI web servers (http://www.ncbi.nlm.nih.gov/) using as input query the currently displayed sequence; several options of applying masking using any combination of amino acid residue types and detection algorithm are available.

The main strength of LCR-eXXXplorer—setting it apart from similar services—is its visualization capabilities. Displaying LCRs in a protein sequence is more informative when information regarding other functional or structural features is also shown (Supplementary Fig. S2). By taking advantage of the underlying GBrowse capability to display features stored on a remote web accessible server, LCR-eXXXplorer incorporates selected annotations from UniProt into the main browser interface. UniProt annotations displayed in LCR-eXXXplorer are of two major types: (i) general annotations associated with the protein sequence (e.g. protein name, gene ontology terms, PDB accession IDs) and (ii) position-specific annotations, which may include domains, sites, secondary structure etc. These annotations are fetched from UniProt/SwissProt on-the-fly for the protein sequence of interest. This is facilitated by a custom-designed cgi-bin script and the retrieved features are further post-processed to a format suitable for the LCR-eXXXplorer.

Using the same underlying mechanism, LCR-eXXXplorer can display tracks generated by another instance of GBrowse, a Distributed Annotation System (DAS) server or valid GFF3 files generated by the user. The only requirement is that the remote tracks must use the same coordinates system, which in the case of LCR-eXXXplorer is the protein sequence itself. Thus, users may practically display results from any LCR-detection tool (or any other protein sequence analysis tool) alongside the data provided by LCR-eXXXplorer.

2.3 Comparison to similar services

Two services for providing access to protein sequence LCR-related data are currently available online. The one most closely related to LCR-eXXXplorer is LPS-annotate (Harbi *et al.*, 2011), which identifies LCRs based on the LPS algorithm (Harrison and Gerstein, 2003), compared to SEG. These LCR annotations are accompanied with disordered region predictions by DISOPRED (Buchan *et al.*, 2010). Even though LPS-annotate is an invaluable resource for researchers interested in compositionally biased proteins, its main drawback is the lack of any effective visualization options. Moreover, the underlying database (according to data available at the LPS-annotate website) has not been updated since 2009. Recently, the HRaP server (Lobanov *et al.*, 2014) was developed, specializing in the study of homopolymeric repeats, which comprise a highly specialized case of LCRs, thus it is not further discussed herein. A detailed presentation of web-based services providing information related to LCRs is presented in Kirmitzoglou (2014).

3 Future Developments

The current version of the LCR-eXXXplorer web server offers several tools for facilitating research on proteins with LCRs, including BLAST search and interactive visualization by exploiting inherent GBrowse features. Given the genuine interest of our research group in LCR-containing proteins, we plan to expand this service in the near future.

More specifically, we are in the process of automating the LCR-eXXXplorer update procedure to regularly synchronize with UniProt updates. Moreover, the customizations performed on different GBrowse modules require some additional work (and appropriate documentation) for enabling full programmatic access to our service through the REST interface already available for GBrowse.

An important improvement destined for the next version of LCR-eXXXplorer is enabling full support of Boolean queries against fields in the underlying database. The modular (both in terms of data and software) architecture of LCR-eXXXplorer enables easy incorporation of novel datasets (e.g. complete genome sequences) and LCR detection tools in future versions.

Funding

The "Cyprus Research Promotion Foundation" is the local agency that funded the projects (with money provided from the "Republic of Cyprus" and the "EU European Regional Development Fund (ERDF)" (The CyTera project, NEA ΥΠΟΔΟΜΗ/ΣΤΡΑΤΗ/0308/31 and ΠΕΝΕΚ/ΕΝΙΣΧ/0308/77).

Conflict of interest: none declared.

References

Altschul,S.F. *et al.* (1994) Issues in searching molecular sequence databases. *Nat. Genet.*, **6**, 119–129.

Andrade,M.A. *et al.* (2001) Protein repeats: structures, functions, and evolution. *J. Struct. Biol.*, **134**, 117–131.

Buchan,D.W.A. *et al.* (2010) Protein annotation and modelling servers at University College London. *Nucleic Acids Res.*, **38**, W563–W568.

Dosztányi,Z. *et al.* (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, **25**, 2745–2746.

Dosztányi,Z. *et al.* (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.

Haerty,W. and Golding,G.B. (2010) Low-complexity sequences and single amino acid repeats: not just "junk" peptide sequences. *Genome Natl. Res. Council Can. Genome Conseil national de recherches Canada*, **53**, 753–762.

Harbi,D. *et al.* (2011) LPS-annotate: complete annotation of compositionally biased regions in the protein knowledgebase. *Database*, baq031.

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.

Kirmitzoglou,I. (2014) Development of algorithms and software for unravelling the biological role of low complexity regions in protein sequences. PhD Thesis, University of Cyprus, Nicosia, Cyprus.

Kuznetsov,I.B. and Hwang,S. (2006) A novel sensitive method for the detection of user-defined compositional bias in biological sequences. *Bioinformatics*, **22**, 1055–1063.

Lobanov,M.Y. *et al.* (2014) HRaP: database of occurrence of HomoRepeats and patterns in proteomes. *Nucleic Acids Res.*, **42**, D273–D278.

Mészáros,B. *et al.* (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*, **5**, e1000376.

Miskinyte,M. *et al.* (2013) The genetic basis of *Escherichia coli* pathoadaptation to macrophages. *PLoS Pathogen*, **9**, e1003802.

Nandi,T. *et al.* (2003) The low complexity proteins from enteric pathogenic bacteria: taxonomic parallels embedded in diversity. In: *Silico Biol.*, **3**, 277–285.

Pizzi,E. and Frontali,C. (2001) Low-complexity regions in *Plasmodium falciparum* proteins. *Genome Res.*, **11**, 218–229.

Priyam, A. *et al.* SequenceServer: BLAST searching made easy. *in preparation*.

Promponas,V.J. *et al.* (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. *Bioinformatics*, **16**, 915–922.

Radivojac,P. *et al.* (2006) Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. *Proteins Struct. Funct. Bioinf.*, **63**, 398–410.

Stein,L.D. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

Tamana,S. *et al.* (2012) Sequence features of compositionally biased regions in three dimensional protein structures. In: IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), 2012, pp. 270–275.

Tsoka,S. *et al.* (1999) Reproducibility in genome sequence annotation: the *Plasmodium falciparum* chromosome 2 case. *FEBS Lett.*, **451**, 354–355.

Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.

Ye,J. *et al.* (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**, W6–W9.