BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3394–3395
doi:10.1093/bioinformatics/btu551

Sequence analysis

Advance Access publication August 14, 2014

SCD*Finder*, a Web-based tool for the identification of putative novel ATM and ATR targets

Lukas Cara[1], Medina Baitemirova[2], Franklin Duong[2], Maia Larios-Sanz[2,*] and Albert Ribes-Zamora[2,*]

[1]Department of Mathematics, Computer Science and Cooperative Engineering, University of St. Thomas and [2]Bioinformatics Program, Department of Biology, University of St. Thomas, 3800 Montrose Blvd, Houston, TX, 77006 USA

Associate Editor: John Hancock

ABSTRACT

Motivation: The S/TQ cluster domain (SCD) constitutes a new type of protein domain that is not defined by sequence similarity but by the presence of multiple S/TQ motifs within a variable stretch of amino acids. SCDs are recognized targets for DNA damage response (DDR) kinases like ATM and ATR. Characterizing DDR targets is of significant interest. The aim of this work was to develop a web-based tool to allow for easy identification and visualization of SCDs within specific proteins or in whole proteome sets, a feature not supported by current domain and motif search tools.

Results: We have developed an algorithm that (i) generates a list of all proteins in an organism containing at least one user-defined SCD within their sequence, or (ii) identifies and renders a visual representation of all user-defined SCDs present in a single sequence or batch of sequences.

Availability and implementation: The application was developed using Pearl and Python, and is available at the following URL: http://ustbioinfo.webfactional.com/scd/.

Contact: ribesza@stthom.edu or lariosm@stthom.edu

Received on June 17, 2014; revised on August 8, 2014; accepted on August 11, 2014

1 INTRODUCTION

ATM and ATR are the chief controlling kinases of a cell's response to DNA damage (Ciccia and Elledge, 2010; Shiloh and Zhiv, 2013). The complete set of ATM and ATR (ATM/ATR) targets is currently unknown, but several high-throughput studies suggest that the final tally could be well over a thousand proteins (Matsuoka *et al.*, 2007). Furthermore, there is growing evidence for ATM functions even in the absence of DNA damage, although direct ATM/ATR targets in those pathways are currently mostly uncharacterized (Shiloh and Zhiv, 2013).

Both ATM and ATR preferentially phosphorylate substrates on serine or threonine residues that are followed by a glutamine (S/TQ). Their specificity for this motif is such that mutations on the glutamine can be as deleterious for function as mutating the serine or the threonine (Kim *et al.*, 1999). In addition, several well-known ATM/ATR targets contain clusters of more than one S/TQ motif within a small region of the protein. This observation led to the description of the SCD, a new type of domain that is not defined by a rigidly conserved sequence similarity but rather by the presence of at least three S/TQ motifs within a stretch of 100 amino acids (Traven and Heierhorst, 2005). This domain is present in 43 of 81 of the better-characterized ATM targets in mammalian cells, and it can also be found in more than half of 686 human proteins identified in a high-throughput analysis as containing phosphorylated S/TQ motifs (Matsuoka *et al.*, 2007). Studies using a more stringent SCD definition (three S/TQ in 50 amino acids) found this domain present four times more abundantly in the yeast proteome than expected by random generation and concentrated in proteins belonging to pathways known to be under ATM/ATR control (Cheung *et al.*, 2012).

Searching for proteins that may contain an SCD is not a simple task, given the domain's arbitrary length and structural diversity that has been observed to date. Current tools for motif or domain identification (like SMART, which is designed to search for signaling domain sequences) typically rely on conserved sequence similarity and are thus not best suited for searching for SCD domains. Even ELM, which identifies functional motifs in eukaryotic proteins and has a Phosphoblast tool that identifies phosphorylated serines, threonines and tyrosines, does not identify SCDs. Similarly, most motif-finding tools do not support the search for multiple motifs in regions of limited length. While it is possible to use the ScanProsite tool with a user-defined pattern, the SCD is not part of the default ProSite collection of motifs. A user can submit a customized search pattern, but whole genome analysis is not possible without integrating to an external bioinformatics pipeline to perform the necessary filtering of the initial results, which requires programming abilities by the user. To facilitate this process, we have created SCD*Finder*, a web-based tool that allows a user to easily search for SCD-containing proteins in different proteomes using a user-determined SCD definition. This program can also be used to identify and visualize user-defined SCD domains within a given sequence or in a batch of submitted sequences.

2 METHODS

SCD*Finder* is implemented mainly in the Python programming language. It uses Prosite's *ps_scan* Pearl script (available from: ftp://ftp.expasy.org/databases/prosite/ps_scan/) to search for proteins containing a S/TQ

*To whom correspondence should be addressed.

cluster domain (SCD; Sigrist *et al.*, 2002). SCD*Finder* then performs three levels of filtering on the results returned. The first level of filtering is necessary to ensure that the SCD found by the *ps_scan* script matches the desired SCD length definition. For example, if we define an SCD as consisting of three S/TQs in 100 residues, then the query SCD*Finder* passes to *ps_scan* is [ST]Q-X[100]-[ST]Q-X[100]-[ST]Q, where [100] is the number of residues between the S/TQ, even though this will return results that have SCDs longer than our desired 100 residues (e.g. TQ-[74AA]-SQ-[90AA]-SQ). Results that do not match the definition will subsequently be removed in this first filtering stage. The second filtering is optional and counts a protein with multiple SCDs as a single hit. The final level of filtering is also optional and accomplishes two things: it restricts the results to a single hit for those proteins that have isoforms and filters out proteins that have the same descriptions but different RefSeq IDs. Finally, using jQuery and Python, SCD*Finder* creates visualizations highlighting the location of the SCDs in the analyzed proteins.

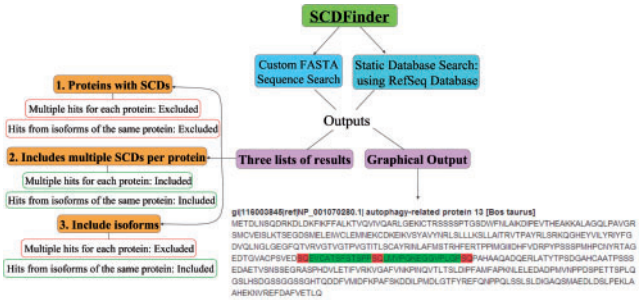

Fig. 1. Flowchart depicting the step-wise process followed in SCD*Finder*. Also shown is an example of the output, showing a retrieved protein sequence highlighted (S/TQ motif in red, SCD in green) for easy identification of the SCD

3 RESULTS

SCDFinder is an online, open-access tool that can be accessed at http://ustbioinfo.webfactional.com/scd/. This SCD search tool is easy to use and offers two distinct interfaces that allow user-defined queries: a static database search and a custom FASTA sequence search. Both allow the user to customize the SCD definition, or use the default definition of three S/TQs in 100 residues.

3.1 Static database search

In the static database search, users can identify all SCD-containing proteins in a given proteome of interest. The user chooses the desired organism from a dropdown menu (the current list contains 17 completed eukaryotic proteomes retrieved from NCBI in January 2013). The user also stipulates the SCD definition by manipulating the number of S/TQ motifs allowed in a specified stretch of amino acids (the default is three S/TQs in 100 amino acids). Multiple definitions can be assigned to each search, and the hits are returned as a hyperlinked list of proteins separated by SCD definition. The summary box, showing a tally of all hits, links to result lists that include the NCBI RefSeq accession number and function for each protein retrieved. Clicking on each hit generates a flat file of the FASTA protein sequence with each S/TQ motif highlighted in red and the SCD highlighted in green (see Fig. 1). Each result list also has an advanced results tab, which shows itemized lists of proteins containing multiple SCDs found per protein type, as well as a list, which also includes SCD-containing isoforms. All run files are downloadable in a CSV format.

3.2 Custom FASTA sequence search

The custom FASTA sequence search allows the user to directly paste sequences of interest into a search window, or to upload a file containing multiple sequences for analysis. Users can also define the SCD based on their preferred cutoff for number of S/TQs and length of SCD. This search option also has an advanced results tab that links to graphical representations of multiple SCD occurrences per protein (highlighted as described above) and a link showing SCDs found in isoforms of each protein submitted. Like with the static database search, all run files for this search option are downloadable in a CSV format.

4 DISCUSSION

SCD*Finder* is a new Web-based tool for the detection and visualization of user-defined SCDs within single proteins, a set of proteins, or entire proteomes. It can be used to identify putative novel ATM/ATR substrates or to reveal new pathways under the possible control of these kinases. This tool may be particularly useful to investigate the function of ATM/ATR in the absence of DNA damage by identifying potential direct targets that may have been missed in studies that fail to reproduce the conditions under which these kinases are activated. Other valuable applications may include its use in comparative proteomics studies or to determine the optimal SCD definition for different organisms.

ACKNOWLEDGEMENT

We would like to thank F.A. San Lucas for technical assistance.

Funding: This project was generously supported by funds from the Smith Cullen Chair of Biology and from the Committee on Student Research at the University of St. Thomas.

Conflict of interest: none declared.

REFERENCES

Cheung,H.C. *et al.* (2012) An S/TQ cluster domain census unveils new putative targets under Tel1/Mec1 control. *BMC Genomics*, **13**, 664.

Ciccia,A. and Elledge,S.J. (2010) The DNA Damage Response: making it safe to play with knifes. *Mol. Cell*, **40**, 179–204.

Kim,S.T. *et al.* (1999) Substrate specificities and identification of putative substrates of ATM kinase family members. *J. Biol. Chem*, **274**, 37538–37543.

Matsuoka,S. *et al.* (2007) ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science*, **316**, 1160–1166.

Shiloh,Y. and Ziv,Y. (2013) The ATM protein kinase: regulating the cellular response to genotixic stress and more. *Nat. Rev. Mol. Cell. Biol*, **14**, 197–210.

Sigrist,C.J.A. *et al.* (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform*, **3**, 265–274.

Traven,A. and Heierhorst,J. (2005) SQ/TQ cluster domains: concentrated ATM/ATR kinase phosphorylation site regions in DNA-damage-response proteins. *Bioessays*, **27**, 397–407.