SECISaln, a web-based tool for the creation of structure-based alignments of eukaryotic SECIS elements

Charles E. Chapple¹, *, Roderic Guigó¹,² and Alain Kro³

¹Institut Municipal d’Investigació Mèdica, ²Centre de Regulació Genòmica, Universitat Pompeu Fabra and Parc de Recerca Biomedica de Barcelona, Carrer del Doctor Aiguader 88, 08003, Barcelona, Catalonia, Spain and ³Unité Architecture et Réactivité de l’ARN, Université Louis Pasteur de Strasbourg, CNRS, 15 rue René Descartes, F-67084 Strasbourg, France

Received on October 7, 2008; revised on December 18, 2008; accepted on January 7, 2009

Advance Access publication January 29, 2009
Associate Editor: Ivo Hafacker

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

Summary: Selenoproteins contain the 21st amino acid selenocysteine which is encoded by an inframe UGA codon, usually read as a stop. In eukaryotes, its co-translational recoding requires the presence of an RNA stem-loop structure, the SECIS element in the 3 untranslated region of (UTR) selenoprotein mRNAs. Despite little sequence conservation, SECIS elements share the same overall secondary structure. Until recently, the lack of a significantly high number of selenoprotein mRNA sequences hampered the identification of other potential sequence conservation. In this work, the web-based tool SECISaln provides for the first time an extensive structure-based sequence alignment of SECIS elements resulting from the well-defined secondary structure of the SECIS RNA and the increased size of the eukaryotic selenoproteome. We have used SECISaln to improve our knowledge of SECIS secondary structure and to discover novel, conserved nucleotide positions and we believe it will be a useful tool for the selenoprotein and RNA scientific communities.

Availability: SECISaln is freely available as a web-based tool at http://genome.crg.es/software/secisaln/.

Contact: charles.chapple@crg.es

Supplementary information: Supplementary data are available at Bioinformatics online.

Selenoproteins are a diverse family of proteins characterized by the presence of the 21st amino acid, selenocysteine (Sec or U). Selenocysteine is co-translationally inserted into the growing polypeptide chain in response to UGA, otherwise read as a stop codon. The correct recoding of UGA to Sec requires the presence of a stem-loop structure, the SECIS element in the 3 untranslated region (UTR) of selenoprotein gene transcripts. Accordingly, the presence of a suitable SECIS element has been used in many studies as a tool for the computational prediction of novel selenoproteins (Castellano et al., 2001; Kryukov et al., 1999; Lescure et al., 1999) and a specialized tool for SECIS prediction, SECISearch (Kryukov et al., 2003), has already been described and has been widely used.

There are two types of eukaryotic SECISes, type I and type II differing at the apex by the presence of the additional helix 3 in type II (Fagegaltier et al., 2000; Grundner-Culemann et al., 1999; Walczak et al., 1996, see Fig. 1). Although the SECIS structure is conserved, there is little sequence conservation beyond the consecutive non-Watson-Crick base pairs UGAN/KGAW constituting the quartet, an unpaired A 5 to UGAN and a run of As in the apical loop/internal loop 2 (Fagegaltier et al., 2000; Walczak et al., 1996). Of these only the UGA/GA of the quartet is invariant1 (e.g. Buettner et al., 1996; Lobanov et al., 2007). Here, we describe SECISaln, a web-based tool that creates structure-based alignments of an extensive dataset of eukaryotic SECIS sequences. Its implementation led us to uncover novel, conserved sequence elements.

SECISaln will predict a SECIS element in the query sequence, split it into its constituent parts and align these against a precompiled database of eukaryotic SECIS elements. The user can choose whether the database sequences are sorted by protein family or by species, thereby offering the possibility of comparing the submitted sequence to other, known SECISes. In addition, SECISaln returns a graphical image of the predicted structure of the user-submitted sequence as well as a multiple structural alignment of all SECIS elements of that type already present in the database. SECISaln uses SECISearch for the SECIS prediction step, described in detail in Kryukov et al. (2003) and is not intended as a replacement for SECISearch. Our patterns and free-energy cutoffs are not stringent and will result in a high false positive rate if used to identify novel SECIS elements. Ideally, SECISaln should be used on sequences which are known to contain a SECIS element, and its main application is the detailed characterization of structural features in the identified SECIS elements, through the multiple structural comparison to other known SECIS elements.

In addition to being the first structural alignment tool for SECIS elements, SECISaln also provides the largest available, manually curated collection of eukaryotic SECISes. Our SECIS collection was built by searching for homologs of all known eukaryotic selenoproteins in NCBI’s RefSeq mRNA and TIGRs EGO databases. We ran TBLASTN searches using the human (when available, other species when not) selenoproteins as queries. We then extracted SECISes.

¹With one exception, the SelT genes of Toxoplasma gondii and Neospora canine have a non-canonical GGA/GA sequence instead (Novoselov et al., 2007).

*To whom correspondence should be addressed.
Fig. 1. Eukaryotic SECIS element consensus sequence. Novel conserved residues are shown in magenta. Where a specific nucleotide is shown, it was observed in that position in 50% or more of the aligned sequences. Where a class of nucleotides is shown, that class was observed in that position in 70% or more of the aligned sequences. Y = U or C, K = G or U, N = any nucleotide, W = A or U, R = A or G, M = A or C. Quartet: four consecutive non-Watson–Crick base pairs. Base pairs forming the quartet were called abcd/a'b'c'd' for the sake of clarity in the text. Position ‘z’ is the first nucleotide after the run of Ms, positions 2H3/2H3 are the second base pair of Helix 3 and 1ap the first nucleotide of the apical loop. The range of possible lengths for helix 1 is hard to determine because it depends on the local 2D structure of the mRNA 3'UTR.

ACKNOWLEDGEMENTS

The authors would like to thank David Martin for his help with CGI scripting and Marco Mariotti for beta testing the software.

Funding: Spanish Ministry of Education and Science (to R.G.); BioSapiens European Network of Excellence (to R.G.); National Institute for Bioinformatics (www.inab.org) a platform of ‘Genoma España’ (to R.G.); ToxNuc-E program (to A.K.); ACI BCMS of the Spanish Ministry of Education and Science (to C.E.C.).

Conflict of Interest: none declared.

REFERENCES


Grundner-Culemann,E. et al. (1999) Two distinct secis structures capable of directing selenocysteine incorporation in eukaryotes. RNA, 5, 625–635.


