— **Minireview** —

From RNA Structure to the Identification of New Genes: The Example of Selenoproteins

Alain Lescure,[a] Daniel Gautheret,[b] Delphine Fagegaltier,[a] Philippe Carbon,[a] and Alain Krol *, [a]

[a]UPR 9002 du CNRS, Structure des Macromolécules Biologiques et Mécanismes de Reconnaissance, Institut de Biologie Moléculaire et Cellulaire, 67084 Strasbourg Cedex, France and [b]UMR 1889 du CNRS, Information Génétique et Structurale, 31, Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

(Received June 30, 2000)

Selenocysteine is incorporated into selenoproteins by an in-frame UGA codon whose readthrough requires the selenocysteine insertion sequence (SECIS), a conserved hairpin in the 3′ untranslated region (3′UTR) of eukaryotic selenoprotein mRNAs. To identify new selenoproteins, we developed a strategy that obviates the need for prior amino acid sequence information. A computational screen was used to scan nucleotide sequence databases for sequences presenting a potential SECIS secondary structure. The computer-selected hairpins were then assayed *in vivo* for their functional capacities and the cDNAs corresponding to the SECIS winners were identified. Four of them encoded novel selenoproteins as confirmed by *in vivo* experiments. Among these, SelZf1 and SelZf2 share a common domain with the mitochondrial thioredoxin reductase TrxR2. The three proteins, however, possess distinct N-terminal domains. We found that another protein, SelX, displays sequence similarity to a protein involved in bacterial pili formation. For the first time, four novel selenoproteins were discovered based on a computational screen for the RNA hairpin directing selenocysteine incorporation.

Key words —— selenocysteine, selenoprotein, RNA motif

INTRODUCTION

Selenocysteine is the major biological form of the essential trace element selenium. In this particular amino acid, selenium replaces sulfur. In the selenocysteine biosynthesis pathway (reviewed in Ref. 1), the first step consists in the charging of serine onto a specialized tRNA, the tRNA^{Sec}. Selenocysteine synthase subsequently ensures the seryl- to selenocysteyl- conversion on the tRNA^{Sec}. Selenocysteine is then cotranslationally incorporated into selenoproteins, and for those selenoenzymes for which a function has been ascribed, it was found in the active center.[2] To date, seven selenoprotein families have been described. What characterizes them is their implication in oxidation–reduction reactions encountered in various pathways. In effect, one can cite the glutathione peroxidase family for scavenging free radicals, the thioredoxin reductase-thioredoxin couple acting in essential functions such as synthesis of deoxynucleotides and maintaining the redox status of the cell, and lastly the iodothyronine deiodinases involved in developmental processes.[3] Because it contains selenocysteine, especially appealing is selenophosphate synthetase SPS2, the enzyme implicated in selenocysteine biosynthesis by converting selenite to phosphoselenoate.[4] Aside from these proteins, three others of unknown functions have been characterized, SelW, SelP and the 15 kDa selenoprotein,[5] (reviewed in Ref. 3).

Sequencing of the first identified selenoprotein cDNAs revealed that selenocysteine is encoded by an in-frame UGA codon. Its cotranslational incorporation into proteins appeals to an original mechanism in order to discriminate UGA selenocysteine from UGA stop codons. In prokaryotes, this is readily achieved through a stem–loop structure localized

*To whom correspondence should be addressed: UPR 9002 du CNRS, Structure des Macromolécules Biologiques et Mécanismes de Reconnaissance, Institut de Biologie Moléculaire et Cellulaire, 67084 Strasbourg Cedex, France. Tel.: +33-3-88-41-70-50; Fax: +33-3-88-60-22-18; E-mail: A.Krol@ibmc.u-strasbg.fr.

immediately downstream of the UGA selenocysteine codon. This RNA motif is recognized by SelB, a specialized elongation factor which also binds the selenocysteyl-tRNA^{Sec}. Thus, the simultaneous binding of SelB to the mRNA and charged tRNA^{Sec} enables designation of the UGA selenocysteine codon and the direct presentation of the charged tRNA^{Sec} to the A site of the ribosome.

Much less is known in eukaryotes about the selenocysteine incorporation mechanism. Remarkably, readthrough of the UGA selenocysteine codon has been shown to require the presence of SECIS (SElenoCysteine Insertion Sequence), an RNA stem-loop structure residing in the 3′ untranslated region (3′UTR) of selenoprotein mRNAs.[6] SECIS is essential for recognition of UGA as a selenocysteine codon rather than a codon for termination of translation; its localization within the 3′UTR differ among selenoprotein mRNAs. Rather surprisingly, it could even be found lying 5 kb downstream from the selenocysteine codon in the type 2 iodothyronine deiodinase.[7] Recently, the purification and characterization of SBP2, a protein which binds SECIS specifically, has been reported.[8]

From the above description, it is obvious that the challenge rests on the discovery of the structural and functional relationships between the SECIS element and the selenocysteine codon. In particular, it is especially important to establish how the SECIS element can fulfill its role while residing far downstream from the codon to be designated. Does this proceed by RNA-RNA interactions, RNA-protein interactions, or both? To address the issue, it was mandatory to determine the structure of the SECIS element at the outset. Deduced from such investigations, it appeared that the detailed knowledge of structure-function constraints in the SECIS element can constitute an invaluable asset for unveiling new selenoprotein mRNAs in databases (see below).

Structure-Function Studies of the SECIS Element

In an earlier work, we proposed an experimentally-derived secondary structure model for the SECIS element, generated by enzymatic and chemical structure probing of the SECIS RNA, as well as by molecular modeling.[9] Further, sequence comparisons with more than 30 different SECIS elements allowed us to refine the model, leading to the Form 1 consensus secondary structure shown in Fig. 1.

The stem-loop structure proposed by the consensus model consists in two helices I and II interrupted by an internal loop, an apical loop surmounting helix II. The SECIS element is characterized by a low degree of sequence conservation, exhibiting invariant residues only in the internal and apical loops. The most striking sequence/structure conservation occurs within helix II, where resides a quartet of non-Watson-Crick base pairs with the invariant G.A/A.G tandem occupying the central position. Clearly, based on these observations, the SECIS element should be seen as an RNA stem-loop in which the global secondary structure is conserved, rather than the nucleotide sequence. Phylogenetic comparisons of additional SECIS sequences suggested that for those SECIS that possess large apical loops, base pairing can occur within the terminal loop.[10] By structure probing,[11] we could experimentally confirm the model, yielding the forms 1 and 2 of the SECIS element (Fig. 1). Interestingly, it appeared possible to classify the SECIS elements into form 1 or 2, according to the type of selenoprotein mRNA they belong to.

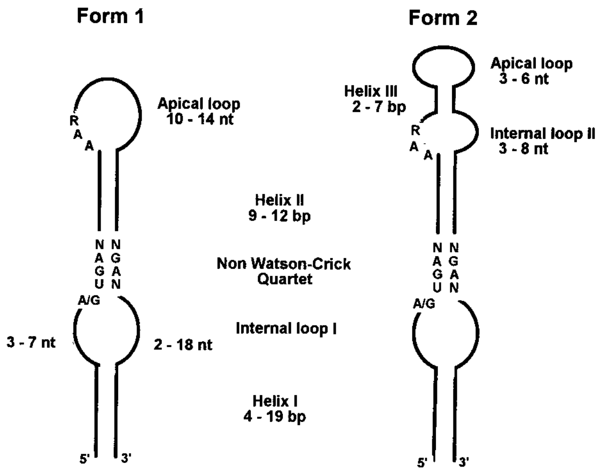

Fig. 1. Consensus Secondary Structure Models for the SECIS Element

Obtained by enzymatic and structure probings, sequence comparisons and molecular modeling.[9–11] The characteristic features of the two forms of the SECIS element are indicated. N, any nucleotide; R, purine. This model served as a basis for elaborating the descriptor used by RNAMOT.

The functional relevance of the structural features identified was subsequently tested, providing simultaneously a means for validating the merits of the model.[12] Single point mutations were introduced into the SECIS element by site-directed mutagenesis at defined structural motifs or sequences. The resulting SECIS mutants were individually introduced into the 3′UTR of the cDNA of the glutathione peroxidase (GPx) in place of the residing wild-type SECIS element. After transfection into COS-7 cells of the GPx cDNAs carrying the SECIS variants, crude cell extracts were prepared. Since GPx contains a selenocysteine in the active center, the effects of the SECIS mutations can be readily reported by monitoring the residual GPx activity. These experiments verified the predictions of the SECIS structural model and especially pinpointed the crucial role of the non-Watson-Crick quartet for the SECIS function.[12)]

Identification of New Selenoprotein mRNAs Using the SECIS Element as a Molecular Tag

Experiments with rats, fed a ^{75}Se-adequate diet, suggested that there should exist more selenium-containing proteins than those described in the Introduction, which would not yet have been identified.[13)] As a matter of fact, it is very likely that identification of these, as yet undiscovered, selenoproteins would be very informative to better understand the pleiotropic role of selenium. Information concerning these proteins is certainly available in the hundreds of thousands of DNA sequences stockpiled in nucleic acid databases, as a result of the various sequencing projects that are underway. Since we were not in possession of the least, even partial, peptide sequence, the classical computer search programs could not be used to fetch them. The key issue, then, consisted in elaborating a ploy for deciphering the hidden information. Since every selenoprotein mRNA contains a SECIS element, the discovery of a novel SECIS element should establish the physical link with an as yet unidentified selenoprotein mRNA. Given the constraints in maintenance of the conservation of the SECIS structure, it appeared to us that this element could be fit for acting as a molecular tag to be sought in databases. Since the information in the databases lies in one dimension only - the sequence itself - we needed an intermediary that would directly convert the sequence itself to SECIS RNA secondary structures. To perform this step, we took advantage of RNAMOT, an algorithm capable of detecting RNA secondary structures by reading out sequences from a nucleic acid database.

A descriptor for the SECIS element was deduced from the structural studies described above (Fig. 1) and used by RNAMOT. This computer program searched, in databases, nucleotide sequences that were able to adopt a secondary structure similar to that proposed by the descriptor. The search conducted on different nucleic acid databases led to the identification of a large number of candidates that were further screened. The abilities of the selected sequences to act as bona fide SECIS elements were tested in vivo with the glutathione peroxidase reporter system described above which asked whether the SECIS hits could functionally replace the residing SECIS element of the GPx mRNA. From these experiments, it turned out that six of the candidates could lead to selenocysteine incorporation into GPx.[14)] The new SECIS elements were used as a molecular handle to uncover the corresponding open reading frames (ORF). Hereby, the corresponding cDNAs have been obtained and sequenced. Among them, three appeared to encode selenoproteins identified earlier, for which only the coding region had been identified before, or new proteins characterized while this study was underway: the selenophosphate synthetase-2, the 15 kDa selenoprotein and the type 2 iodothyronine deiodinase. The sequences of the other three corresponded to new selenoproteins, which we called SelX, SelN and SelZ. Surprisingly, the SECIS of SelZ appeared to match two cDNA sequences, SelZf1 and SelZf2, with distinct 5′ extremities. Inspection of the cloned sequences revealed that, as expected for selenoprotein cDNAs, they all show the presence of an in-frame UGA codon and the SECIS element localized within the 3′UTR (Fig. 2). *In vivo* labeling, by growing transiently transfected COS-7 cells in a medium containing radioactive selenium, confirmed incorporation of selenium into the new proteins. Moreover, we showed that this incorporation was dependent on the presence of the SECIS element.[14)]

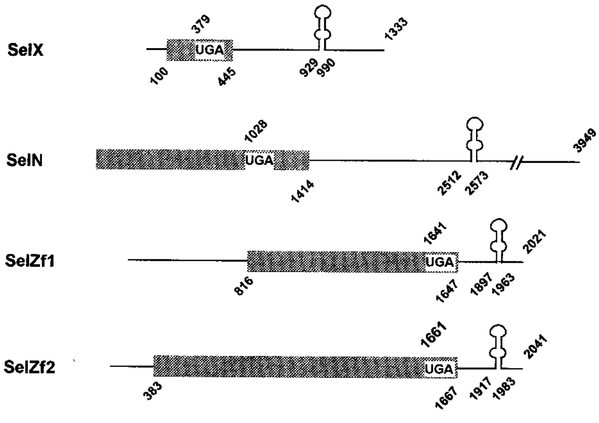

Fig. 2. Diagrammatic Representations of the New Selenoprotein SelX, SelN, SelZf1 and SelZf2 cDNAs

The coding and untranslated regions are represented by gray boxes and single lines, respectively. UGA selenocysteine codons are boxed; SECIS elements are depicted by stem-loop structures.

New Selenoproteins toward New Functions

Tissue specific expression of the new selenoproteins was analyzed by Northern blot, and potential functions were deduced from amino acid similarities to known proteins. SelN appeared to be ubiquitously expressed in all tissues considered but showed no homology to any known protein. In contrast, SelX is similar to a protein of unknown function detected in all organisms from bacteria to mammals.[14] Prominently, however, selenocysteine occurs only in mammals whereas other organisms contain a cysteine. Amino acid sequence comparison also showed that SelX presents similarities to a domain of the bacterial transcriptional regulator PILB involved in pili formation. Search for amino acid similarities predicted that SelZf1 and SelZf2 share a common domain with the mitochondrial thioredoxin reductase, but possess distinct N-terminal domains. Sequence comparisons suggested that the SelZf1, SelZf2 and the mitochondrial thioredoxin reductase mRNAs arise from the same gene by alternative splicing, resulting in the addition of different 5′ segments to a common core. This mechanism likely generates three different selenoproteins with specialized functions or localizations.

CONCLUSION

Sequence comparisons and structure probing revealed that SECIS elements probably correspond to an extreme situation of RNA motifs where most of the functional determinants are structure-based, and only a few of them consist in sequence conservation. This observation stresses the central role played by the RNA shape in molecular recognition. However, the programmes available so far to identify new genes are based on sequence similarities. In our case, the information gathered on the structure of the SECIS element combined with an original algorithm, was used to screen the hundreds of thousands of nucleotide sequences stockpiled in the databases. This led us to the identification of four novel selenoproteins. These results demonstrate once again the value of mRNA 3′-UTR as a repository of functional RNA motifs instrumental in post-transcriptional control. Undoubtedly, this strategy could be extended to the discovery of other genes whose expression is regulated by common structural motifs localized in the untranslated regions of mRNAs.

REFERENCES

1) Atkins J.F., Böck A., Matsufuji S., Gesteland R.F., "The RNA World, Second Edition" Gesteland R.F., Cech T.R., Atkins J.F. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NewYork, pp. 637–673, 1999.

2) Stadtman T.C., *Annu. Rev. Biochem.*, **65**, 83–100 (1996).

3) Burk R.F., Hill K.E., *Bioessays*, **21**, 231–237 (1999).

4) Guimaraes M.J., Peterson D., Vicari A., Cocks B.G., Copeland N.G., Gilbert D.J., Jenkins N.A., Ferrick D.A., Kastelein R.A., Bazan J.F., Zlotnik A., *Proc. Natl. Acad. Sci. U.S.A*, **93**, 15086–15091 (1996).

5) Gladyshev V.N., Jeang K.T., Wootton J.C., Hatfield D.L., *J. Biol. Chem.,* **273**, 8910–8915 (1998).

6) Berry M.J., Banu L., Chen Y.Y., Mandel S.J., Kieffer J.D., Harney J.W., Larsen P.R., *Nature,* **353**, 273–276 (1991).

7) Buettner C., Harney J.W., Larsen P.R., *J. Biol. Chem.,* **273**, 33374–33378 (1998).

8) Copeland P.R., Fletcher J.E., Bradley A.C., Hatfield D.L., Driscoll D.M., *EMBO J.,* **19**, 306–314 (2000).

9) Walczak R., Westhof E., Carbon P., Krol A., *RNA*, **2**, 367–379 (1996).

10) Grundner-Culemann E., Martin G.V.III, Harney J.W., Berry M.J., *RNA*, **5**, 625–635 (1999).

11) Fagegaltier D., Lescure A., Walczak R., Carbon P., Krol A., *Nucleic Acids Res.* **28**, 2679–2689 (2000).

12) Walczak R., Carbon P., Krol A., *RNA*, **4**, 74–84 (1998).

13) Behne D., Kyriakopoulos A., Weiss N.C., Kalckloesch M., Westphal C., Gessner H., *Biol. Trace Elem. Res.,* **55**, 99–110 (1996).

14) Lescure A., Gautheret D., Carbon P., Krol A., *J. Biol. Chem.*, **274**, 38147–38154 (1999).