Novel Selenoproteins Identified in Silico and in Vivo by Using a Conserved RNA Structural Motif*

Alain Lescure‡, Daniel Gautheret§, Philippe Carbon‡, and Alain Krol‡

From l’Unité Propre de Recherche CNRS 9002, Structure des Macromolécules Biologiques et Mécanismes de Reconnaissance, Institut de Biologie Moléculaire et Cellulaire, 15, Rue René Descartes, 67084 Strasbourg Cedex, France and S’Unité Mixte de Recherche CNRS 1889, Information Génétique et Structurale, 31, Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

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 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 data bases 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 mitochondrial thioredoxin reductase-2. The three proteins, however, possess distinct N-terminal domains. We found that another protein, SelX, displays sequence similarity to a protein involved in bacterial pilus formation. For the first time, four novel selenoproteins were discovered based on a computational screen for the RNA hairpin directing selenocysteine incorporation.

Selenium is an essential trace element whose deficiency can interfere with normal embryonic development and fertility or favor the appearance of certain cancers and viral diseases such as human immunodeficiency virus and coxsackievirus (1). The amino acid selenocysteine is the major biological form of selenium in bacteria and animals. It is found in the active site of selenoproteins and is directly involved in the catalytic reaction. In this regard, the capacity of the selenocysteine selenol group to become ionized at physiological pH, the cysteine thiol group requiring a higher pH, accounts for the higher rate of catalysis of selenoenzymes (2). Seven selenoprotein families have been characterized so far in mammals (3): the glutathione peroxidase-1, thioredoxin reductase-1 and thioredoxin reductase-2, the three proteins, however, possess distinct N-terminal domains. We found that another protein, SelX, displays sequence similarity to a protein involved in bacterial pilus formation. For the first time, four novel selenoproteins were discovered based on a computational screen for the RNA hairpin directing selenocysteine incorporation.

* This work was supported by grants from the Association pour la Recherche sur le Cancer, the Ligue Nationale contre le Cancer, the Ligue contre le Cancer du Haut-Rhin, and the Centre Volvic pour la Recherche sur le Cancer, the Ligue Nationale contre le Cancer, the Fondation Néphrologie, the Saint-Louis Hospital, the Laboratory of Endocrinology and the Laboratory of Cardiology. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank®/EBI Data Bank with accession number(s) AF166124, AF166125, AF166126, and AF166127.

† To whom correspondence should be addressed. Tel.: 33-3-88-41-70-50; Fax: 33-3-88-60-22-18; E-mail: krol@ibmc.u-strasbg.fr.

kDa selenoprotein of unknown function has been purified (4). Selenophosphate synthetase-2, the seventh selenoprotein, is remarkable in that it contains selenocysteine, but is also a key actor in the biosynthesis of this amino acid (5).

Selenocysteine is encoded by an in-frame UGA codon, implying the existence of a mechanism capable of distinguishing the UGA selenocysteine codon from a translational stop. This process requires, in eukaryotes, the presence of the selenocysteine insertion sequence (SECIS), a hairpin residing in the 3’-untranslated region of selenoprotein mRNAs that is essential for readthrough of the UGA selenocysteine codon (6). Sequence comparisons and structure-function experiments generated a consensus secondary structure model for the SECIS element in which a functional motif could be identified (7, 8).

Compelling evidence for the existence of molecular links between selenium deficiencies and biological disorders came from molecular genetics experiments. Targeted disruption of the mouse selenocysteine tRNA gene led to early embryonic lethality, implying that selenoprotein synthesis is essential to mammals (9). Studies carried out on knockout mice lacking the glutathione peroxidase underlined the protective role of selenium against free radicals (10) or coxsackievirus-induced myocarditis in Keshan disease (1). Further supporting the biological importance of this trace element, selenium labeling experiments in rats determined the existence of more selenoproteins to be identified and characterized (11). To undertake this task, we intended here to exploit the mine of information stored in EST data bases. The central question in such a project is how the relevant cDNAs can be retrieved without the knowledge of even a partial protein sequence. To circumvent the obstacle, a strategy was developed based on the absolute requirement of a SECIS element for selenoprotein translation. The finding of such a hairpin in a cDNA should therefore signal the presence of an attached coding sequence. Two assets were exploited to extract new SECIS elements from EST data bases. The first one was the detailed knowledge of the secondary structure of the SECIS element, which is conserved in all known selenoprotein mRNAs. The second one was the utilization of a program capable of detecting potential RNA secondary structures in nucleotide sequence data bases. Combined with molecular biology and in vivo experiments, this approach led to the discovery of four novel selenoproteins using a single RNA element as a structural tag.

1 The abbreviations used are: SECIS, selenocysteine insertion sequence; EST, expressed sequence tag; ORF, open reading frame; PCR, polymerase chain reaction; bp, base pair(s); HA, hemagglutinin; RACE, rapid amplification of cDNA ends; GPx, glutathione peroxidase; UTR, untranslated region; TrxR2, mammalian mitochondrial thioredoxin reductase-2; contig, group of overlapping clones.
**EXPERIMENTAL PROCEDURES**

**Computational Screen and Sequence Comparisons**—The search for new SECIS elements was conducted in GenBank™, sequence-tagged site, and EST data bases with the RNAMOT pattern search program (12, 13) with the descriptor shown in Fig. 1A. 600,300 3'- and 5'-ESTs were scanned, representing a total of ~222 × 10^9 nucleotides. 2 Positive hits were aligned with ClustalW (14). The same descriptor run against a random set of 10^9 nucleotides, 60, 5', and CDS (25%) each) yielded three hits. ORFs and ESTs were identified by BLAST searches (15) in the GenBank™ and EST data bases and aligned with ClustalW.

**Cloning of the New SECIS Elements**—The new SECIS elements were obtained by standard PCR amplification of a human B cell library or of human or other primate tRNAs, to obtain different in sequence. Into the blunt-ended pSelN plasmid was inserted either the 1121-bp 3'-fragment of R47273 to pSelN with oligonucleotide GGCCTGCAGGGATCCTGGAT(N)24 as the 5'-primers, respectively (N24 corresponds to 24 nucleotides complementary to the SECIS sequence, including the top 4 base pairs of helix I (see Fig. 1A). SECIS AA109465 was constructed by nested PCR. The PCR primers introduced a BclI site at the 5'- and a KpnI at the 3'-end of the SECIS elements in addition to a 4-bp stem below helix I (see Fig. 1B). To replace the naturally occurring SECIS element in the glutathione peroxidase reporter, the SECIS candidates were introduced in pGHA-BcK at the BclI-KpnI sites (8). This plasmid encodes a triple-HA tag fused in-frame to the N terminus of the glutathione peroxidase coding sequence (8).

**Identification and Cloning of the cDNAs Encoding the Novel Selenoproteins**—ESTs corresponding to the functional SECIS elements were identified by querying EST data bases with BLASTN at NCBI. The 1537-bp sequence (8) was entirely sequenced and named pSelZ. Similarly to pSelN, additional 5'- and 3'-end sequences of pSelN4 were ligated to the blunt-ended HinI-digested pXJ(HA)3 vector to generate pXJ(HA)3.A. Another fragment of 2066 bp, overlapping the 1544 bp 5'- to pSelN2, was obtained by screening a HeLa oligo(dT) library (a gift of P. Chambon) with a probe corresponding to the functional SECIS element with oligonucleotide ATTGC (CCGCTTACCCTC or GCGGCCGCAGGAATGGATCCTCTTTATTTGC- T(N)24 and CGGGGTACCTGGAT(N)24 as the 5'- and 3'-primers, respectively. (N24 corresponds to 24 nucleotides complementary to the SECIS sequence, including the top 4 base pairs of helix I (see Fig. 1A).)

**RESULTS**

**A Computational Screen for Novel Selenoproteins**—To scan sequences that could adopt secondary structures similar to the SECIS element, we developed a computational screen based on the pattern search program RNAMOT (12, 13). An input primary/secondary structure descriptor (Fig. 1A) for RNAMOT was inferred from sequence comparisons and the consenssus structure experimentally determined at the time of the search (7, 8). To test the validity of the descriptor, RNAMOT was run against the GenBank™ non-redundant data base (10^9 nucleotides at the time), generating 34 different SECIS elements belonging to the then known selenoprotein mRNAs. An additional hit (M35391 in Fig. 1C) was found in an intron of the human procollagen α2 chain gene. Given its localization, it is not likely to represent a bona fide SECIS element. However, it was retained because it contained all the

---

2 May 1997 release.
features of the SECIS consensus structure. Also, a search with an alternative descriptor carrying N instead of B at the top base pair of the non-Watson-Crick quartet led to the discovery of a SECIS element in the 3'-UTR of the selenophosphate synthetase-2 cDNA. This cDNA was characterized earlier, but no SECIS element could be found by the authors (5).

In a second step, the search was conducted in the GenBank™ EST data base (222 × 10^6 nucleotides). After discarding ambiguous hits containing one or more undefined nucleotides, RNAMOT found 376 sequences, including 153 mouse, 101 human, 92 Brugia malayi, and 30 other animal and plant ESTs. A sequence alignment was performed with ClustalW (14), and we plotted the derived neighbor-joining tree to obtain a clustered representation of the matches. This identified 62 sequences that could be classified into three families. One family comprised sequences corresponding to the R16491 and R23284 ESTs characterized in the GenBank™ EST data base search. After this first round of selection, 21 SECIS candidates were obtained, comprising 2 cDNAs, 17 ESTs, and 2 sequence-tagged sites.

Functional Assays of the Selected SECIS Candidates—The 21 SECIS candidates were then tested for in vivo function. The SECIS DNAs were obtained by PCR amplification of genomic DNA or cDNA libraries. Concomitant with the PCR amplification and due to the uneven stability of helix I in the different SECIS elements, an identical 4-bp stem was added below helix I in all SECIS elements (Fig. 1B) in order for the SECIS RNAs to exhibit similar stabilities. GPx being a selenoprotein, its translation requires a functional SECIS element in the 3'-UTR of its mRNA. The SECIS DNA candidates were then introduced separately into the 3'-UTR of a GPx cDNA reporter to replace the residing SECIS element. In this construct, the GPx coding sequence carries an HA tag fused in frame at the N terminus to allow detection of the translated proteins with the anti-HA antibody. Whether or not the SECIS candidates were active could be apprehehended by a rapid assay involving COS-7 transfections of the constructs, followed by Western blotting experiments. A functional SECIS candidate should lead to translation of a full-length GPx. In contrast, with an inactive SECIS element, the UGA selenocysteine codon will be recognized as a stop codon, leading to translation of a shortened 9.5-kDa polypeptide. Translation of the mRNA coding for the HA-tagged GPx, carrying its own SECIS, generated a product of ~27 kDa (Fig. 2A, lane 2). Construct GPx-mutSECIS had the G/A/A/G to A-G/G-A substitution in the non-Watson-Crick quartet.
A Computational Screen for Novel Selenoproteins

Fig. 1. In vivo functional assays of the SECIS candidates. A, capacities to direct readthrough of the glutathione peroxidase UGA reporter, transfected into COS-7 cells. The lengths of the proteins were estimated from mock-transfected COS-7 cells; lane 2, transfection of the wild-type GPx cDNA; lane 3, transfection of the GPx-mutSECIS construct harboring a debilitated SECIS element. The position of the wild-type GPx in lane 3 is a size marker for the expected translation products. Lanes 1–12 are from different gels. Controls (bars 1–3) are as described for lanes 1–3 in A. SPS2, selenophosphate synthetase-2.

Fig. 2. In vivo functional assays of the SECIS candidates. A, capacities to direct readthrough of the glutathione peroxidase UGA reporter, transfected into COS-7 cells. The lengths of the proteins were evaluated by Western blot analysis with the anti-HA antibody. Lane 1, mock-transfected COS-7 cells; lane 2, transfection of the wild-type GPx cDNA; lane 3, transfection of the GPx-mutSECIS construct harboring a debilitated SECIS element. The position of the wild-type GPx in lane 2 is a size marker for the expected translation products. Lanes 1–12 and 13–24 are from different gels. B, GPx activities arising from the transfected GPx cDNAs carrying the SECIS candidates. Average values (from three independent transfections carried out in triplicate) were obtained with respect to the transfected wild-type (wt) GPx taken as 100%. Controls (bars 1–3) are as described for lanes 1–3 in A. SPS2, selenophosphate synthetase-2.

The SECIS element that impaired its function (8), providing here also minute amounts of GPx (compare lanes 2 and 3). This construct provided the background level. Consistent with earlier observations (8), no 9.5-kDa protein appeared with GPx-mutSECIS, presumably due to the instability of such an unnatural short polypeptide in vivo. Fig. 2A shows that, among the 21 SECIS tested, only R71722, AA057045, selenophosphate synthetase-2, AA107841, R46598, and R44842 could mediate production of a full-length GPx with an efficiency comparable to that of the authentic GPx SECIS (compare with lanes 2 and 3). A seventh element, AA280511 (lane 13), also produced full-length GPx, but with a lower efficiency.

Since the active site of GPx contains an essential selenocysteine, measuring the enzymatic activity will attest that this amino acid was effectively incorporated into the protein. After transfection into COS-7 cells of the cDNA constructs carrying the SECIS candidates, GPx activities were assayed from crude cell extracts and compared with that of wild-type GPx (Fig. 2B, bar 2). As anticipated, no significant activity emanated from GPx-mutSECIS (bar 3). Wild-type or slightly higher than wild-type activities were observed with R71722 (105%), AA057045 (110%), and R46598 (~100%). AA107841, selenophosphate synthetase-2, and R44842 retained 80, 73.5, and 54% of the wild-type activity, respectively. The activity dropped to 20% with AA280511 (bar 13). A correlation between both approaches could be thus established, showing that those SECIS candidates producing full-length GPx also conferred wild-type or significant GPx activity.

Fig. 3. Tissue-specific expression patterns of the SelN, SelX, SelY, and SelZ mRNAs. Shown are the results from Northern blot hybridization of human multiple tissues. Poly(A) RNAs (CLONTECH) were hybridized sequentially with 32P-labeled probes derived from positions 739–890 in SelX, positions 1218–1899 in SelN, positions 1227–1762 in AF007144 (SelY), and positions 952–1663 in SelZf1 DNAs.

Arrows point to the estimated sizes of the mRNAs in kilobases (kb). Synthesis of a full-length, enzymatically active GPx could be obtained with seven of the selected SECIS elements, indicating that they were capable of promoting selenocysteine insertion. Possible explanations for the inactivity of the other candidates will be discussed.

Identification of the cDNAs Harboring the New Functional SECIS Element—In the previous assay, we functionally characterized the selenophosphate synthetase-2 SECIS element of the selenophosphate synthetase mRNA. Next, we sought the open reading frames lying upstream of the remaining new SECIS elements. ESTs physically linked to each SECIS element were searched in the GenBank EST data base with BLASTN. The EST sequences collected after an iterative BLASTN search were processed with the CAP program to assemble one contiguous cDNA sequence. The longest cDNAs were obtained and sequenced. The sequence of the cDNA that we found linked to SECIS AA280511 revealed that the SECIS element resides in fact on the opposite strand relative to the putative ORF. Yet constituting a potential bona fide SECIS element, we could not identify an ORF in the proper orientation. We found that SECIS elements AA107841 and R46598 corresponded to the SECIS elements of selenoprotein mRNAs characterized while our study was underway. Indeed, the sequence of the cDNA linked to SECIS AA107841 was found to be identical to that of the 15-kDa selenoprotein (4). The length of the mRNA bearing SECIS R46598, which we call SelY, was estimated to be 6 kilobases by Northern blot analysis (Fig. 3, lanes 7 and 8). This size suggested that it could correspond to the mRNA of type 2 idothyronine deiodinase, whose coding frame, deprived of the 3′-UTR, was isolated earlier (17). Our cloning and sequencing of SelY cDNA showed that it was identical to the 3′-UTR of type 2 idothyronine deiodinase (18).

Since translation of the cDNA sequences linked to the remaining three SECIS elements, R71722, AA057045, and R44842, showed no homology to known selenoproteins, the cDNAs were termed SelN, SelX, and SelZ, respectively. The sizes of the SelN, SelX, and SelZ mRNAs were estimated by Northern blot analysis to be 4.5, 1.4, and 2.2 kilobases, respectively (Fig. 3). By screening a HeLa oligo(dT) library with a probe complementary to the SelIX SECIS DNA, we identified a 1333-bp fragment presumably corresponding to the full-size...
SeIX cDNA. The sequence analysis revealed the existence of a 345-bp-long ORF with an in-frame TGA codon at position 379 (Fig. 4). As expected for a selenoprotein mRNA, its SECIS element effectively resides within the 3'-UTR. Querying EST data bases with BLAST identified a 2231-bp cDNA that was incomplete since the corresponding mRNA was 4.5 kilobases long (Fig. 3). Upstream sequences were thus obtained by screening a HeLa random-primed cDNA library and 5'-Marathon RACE, extending them by 1718 bp. Assembled together, the fragments gave rise to a 3949-bp SeIN cDNA, the sequence of which indicated that the reading frame was still open. However, the 3949-bp SeIN cDNA contained a 1414-bp ORF with a characteristic in-frame TGA codon at position 1028, which was used for subsequent analysis. Here also, the SECIS element occurred within the 3'-UTR of the SeIN cDNA (Fig. 4).

The sequencing of the EST corresponding to SECIS R44842 determined the presence of a 1505-bp cDNA that contained an ORF that obviously extended upstream of the characterized sequence. This cDNA was called SeIZ. Additional 5'-sequences were searched by 5'-Marathon RACE. Surprisingly, we obtained two different PCR fragments with different 5'-sequences. Each fragment obtained, added separately to the SeIZ cDNA, generated the 2021-bp SeIZ1 and 2041-bp SeIZ2 cDNAs. The 5'-sequences of these cDNAs differ upstream of positions 540 in SeIZ1 and 520 in SeIZ2 and are followed by the common SeIZ region (Fig. 4). Since the corresponding transcripts are approximately the same size, they could not be distinguished by Northern blot analysis with a probe complementary to the common SeIZ sequence (Fig. 3). Putative ATG initiation codons were identified by the presence of upstream sequences homologous to the Kozak consensus sequence (19) at positions 816 in SeIZ1 and 383 in SeIZ2. A TGA codon was found in the common region, potentially encoding a selenocysteine at the C-terminal penultimate position in both proteins. For SeIZ1 and SeIZ2, the SECIS element was localized 200 bp downstream of the putative TAA stop codon (Fig. 4).

Can the New SECIS Elements Mediate Readthrough of the Selenocysteine Codon in Their Own mRNA Context?—SeIX and SeIN were fused at the N terminus to an HA tag, generating constructs HASeIX and HASeIN, respectively. In SeIZ1 and SeIZ2, the putative selenocysteine codon resides at the penultimate C-terminal position in a domain common to both proteins. Therefore, only the SeIZ common region was epitope-tagged at the N terminus, giving rise to HASeIZ. After transfection of the constructs into COS-7 cells, the tag allowed immunodetection by the anti-HA antibody of the proteins contained in the cell extracts, hence evaluation of their sizes. In △SECIS constructs, the absence of the SECIS element should convert the UGA selenocysteine to a stop codon, thus producing a shortened polypeptide. Based on the cDNA sequence, HASeIX should generate either 17.2- or 15-kDa proteins, according to selenocysteine codon readthrough. Transfection of HASeIX indeed generated a major product at ~16 kDa, but also a minor one at ~10 kDa (Fig. 5A, lane 4), possibly arising from inefficient selenocysteine codon readthrough (6). Construct HASeIX△SECIS, as anticipated, produced almost exclusively the shortest form (lane 5). Obtaining the faint SECIS-independent 16-kDa band was reminiscent of what happened with GPx (lane 3) and other selenoproteins (5).

A 58-kDa product corresponding to the full-length protein produced by HASeIN was expected. Indeed, synthesis of a 60-kDa protein was observed (Fig. 5A, lane 6). Even though a shorter product of 51 kDa showed up both in the presence and absence of the SECIS element (compare lanes 6 and 7), it must be stressed that the expected full-length 60-kDa protein appeared only in the presence of the SECIS element. Since the UGA codon is located at the penultimate position in the SeIZ mRNA, we should not expect a difference in the mobilities of the full-length 48-kDa and UGA-terminated proteins. This is effectively what happened (lanes 8 and 9). We concluded from these experiments that the SECIS elements in the SeIX and SeIN mRNAs function to mediate readthrough of the selenocysteine codon, with the only ambiguity remaining for SeIZ.

SeIX, SeIN, and SeIZ Are Selenoproteins—To solve the SeIZ ambiguity, but also to assert that the new cDNAs do encode selenoproteins, in vivo labeling was performed by growing transiently transfected COS-7 cells in a medium containing Na$_{75}$SeO$_3$. The HA-tagged proteins were immunoprecipitated from the cell extracts with the anti-HA antibody and fractionated by SDS-polyacrylamide gel electrophoresis. The immunoprecipitation and the difference in size arising from the tag enabled the specific detection of the recombinant selenoproteins. For SeIX, SeIN, and SeIZ, a $^{75}$Se-labeled product was obtained only with the SECIS-containing cDNAs (Fig. 5B, compare lanes 4 and 5, 6 and 7, and 8 and 9). The positions of the bands correlated with the protein sizes predicted from the cDNA lengths and with those on the Western blot in Fig. 5A. The variable intensities of the bands may be accounted for by differential mRNA or protein stabilities or by different activities carried by different SECIS elements, as previously observed in other contexts (20). In the control experiment, the full-length GPx protein was accompanied by a lower molecular mass product of ~22 kDa, which could arise from proteolysis (lane 2). Worth noting is the lack of detection of the full-length GPx, SeIX, and SeIN proteins that were observed on the Western blots in the absence of SECIS elements (Fig. 5A), even after long exposure (data not shown). It may well be that these selenium-lacking proteins originated from weak unspecific readthrough of the selenocysteine codon under our experimental conditions.

These results conclusively demonstrate that SeIX, SeIN, and SeIZ are indeed selenoproteins. Because SeIZ exists in two isoforms, this corresponds to four novel selenoproteins: SeIX, SeIN, SeIZ1, and SeIZ2. Since the corresponding cDNAs each contain an in-frame TGA codon and a SECIS element, the selenium labeling experiments strongly argue in favor of specific selenocysteine incorporation.

Searching Functions for the New Selenoproteins—Northern blot analysis was performed to determine possible tissue-specific expression of SeIX, SeIN, and SeIZ (Fig. 3). SeIN mRNA was ubiquitously expressed, with, however, a higher accumulation in the pancreas, ovary, prostate, and spleen. The distri-
FIG. 5. Translation of and 75Se incorporation into SelN, SelX, and SelZ. A, the SECIS elements mediate UGA readthrough from their own mRNA contexts. After transfection of the constructs (with SECIS [lanes 4, 6, and 8] and lacking SECIS [lanes 5, 7, and 9]) into COS-7 cells, the HA-tagged proteins were revealed by Western blot analysis with the anti-HA antibody. Control lanes are the same as described for Fig. 2A. Migrations in lanes 1–5 and 6–9 were on 10 and 12% gels, respectively. Arrows point to the translation products mentioned under “Results”; asterisks indicate unspecific products. B, SelN, SelX, and SelZ are selenoproteins. Transfected COS-7 cells were cultured in the presence of 75Se. The HA-tagged 75Se-labeled proteins were immunoprecipitated, fractionated on a 12% gel, and revealed by autoradiography.

bution of the SelX mRNA was less homogenous than that of SelN, being preponderant in the liver and leukocytes, abundant in the pancreas, but low in the lung, placenta, and brain. SelZ mRNA showed more pronounced accumulation in the kidney, liver, testis, and prostate, but was low in the thymus.

In the course of this study, the cDNA for the selenoprotein TrxR2, a mitochondrion-specific thioredoxin reductase isoform, was cloned independently by several groups (21–23). Sequence comparisons between the SelZ1, SelZ2, and TrxR2 cDNAs, depicted schematically in Fig. 7A, indicated that they share a large common domain. The SelZ1 and TrxR2 cDNA sequences are identical from the 3'-end to residue 636 of TrxR2. In the SelZ2 cDNA, the region conserved with TrxR2 extends up to position 293 of TrxR2. The common region in the three cDNA sequences includes the 3'-part of the coding sequence with the in-frame TGA codon and the 3'-UTR, with sequence differences occurring at their 5'-ends. The three cDNAs encode three different proteins sharing a common core, but with different N-terminal domains.

Alignment of the human SelN DNA sequence with ESTs or of the SelN protein sequence with translated ESTs revealed the existence of a hypothetical ortholog in mouse and rat. The number of different ESTs was insufficient for reconstitution of complete cDNAs, but the partial assembled sequences showed conservation of the coding frames, in-frame TGA codons, and SECIS elements.

We next sought homologs to SelIX. A mouse cDNA covering the entire length of the human SelIX cDNA was reconstituted in silico by merging various overlapping mouse ESTs. The translated mouse cDNA showed 91% amino acid identity to the human SelIX protein. Furthermore, data base searches found SelIX sequence similarities to plant and prokaryotic, yeast, and Caenorhabditis elegans ORFs indexed as hypothetical proteins of unknown function. Displayed in Fig. 6, these findings show striking amino acid identities between, for example, human SelIX and Escherichia coli P39903 (24%), C. elegans P34436 (28%), and Drosophila EST AA540562 (28%). The comparison also stressed the 29% amino acid identity of the human and mouse SelIX proteins to a domain of the Neisseria gonorrhoeae, Hemophilus influenzae, Helicobacter pylori, Mycoplasma capricolum, and Streptococcus pneumoniae PILB proteins, regulators of bacterial pilus formation (24). Although the sequences are similar over their entire lengths, the alignment highlights two blocks of higher sequence conservation: PWPAF (1)→GLGHVF (2) in mammalian SelIX and GWP(A/S)F (1)→HLGHVF (2) in the SelX homologs (blocks 1 and 2 in Fig. 6; the only two positions where X and M replaced L could originate from sequence uncertainties in the corresponding ESTs). It is striking that only the mammalian SelIX proteins incorporate selenocysteine, whereas other organisms contain a cysteine instead. Sequence conservation is observed flanking the cysteine/selenocysteine (U): R(Y/H)(C/I/V/M)N in SelIX homologs and RFUIF in mammalian SelX.

**DISCUSSION**

The objective of our study was the isolation of new selenoprotein cDNAs. The existence of selenoproteins other than those previously characterized was predicted by workers based on selenium labeling experiments, but did not lead to amino acid sequence data. To circumvent the lack of protein sequence information, we assumed that a number of the desired cDNA sequences were already deposited in the EST data bases. To exploit this information, our strategy took advantage of the obligatory presence of a SECIS element in all selenoprotein mRNAs. This differs from conventional screens in two respects. The SECIS hairpin being characterized more by the high conservation of its secondary structure than by the extent of invariant sequences, alignment methods such as BLAST and FASTA were inappropriate. The originality of our approach was the use of a program capable of detecting RNA foldings such as the SECIS consensus secondary structure. Another and probably the most important aspect of our screen is that selenoprotein cDNAs contain TGA codons, obviously rendering the identification of an ORF more challenging than in other cDNAs where TGA signals the end of the ORF. Notwithstanding, the strategy paid off since the RNA structure alone was sufficient to discover four novel different selenoproteins.

Seven SECIS candidates, out of the 21 selected in silico, indeed corresponded to functional SECIS elements. This came as a surprise since the inactive candidates harbored the features defined by the SECIS consensus structure. Several possibilities can explain this paradoxical situation. The SECIS losers may lack one or more essential sequences or base pairs that could have been unintentionally omitted in the SECIS descriptor because they were not yet identified in the then known SECIS elements. Alternatively, the SECIS losers may contain sequence or base pair anti-determinants preventing them from functioning. Finally, the sequences may fold in vivo into structures slightly different from the expected one.
Three SECIS elements among the seven winners led to the discovery of the SelN, SelX, and SelZ selenoprotein mRNAs, SelZ giving rise to the SelZf1 and SelZf2 isoforms. In vivo expression of the selenoprotein mRNAs indicated that selenocysteine incorporation was actually dependent on the presence of the SECIS element. No sequence similar to SelN could be found in protein or nucleotide sequence databases. However, similarity searches were productive with SelX and SelZ. The amino acid comparisons in Fig. 6 underscored two prominent features of SelX. First, sequences similar to mammalian SelX were detected in all kingdoms. The human and mouse SelX sequences, U stands for selenocysteine, marked Sec below the sequence. For the five PILB proteins, only the domain similar to SelX is shown. Identical amino acids are shaded; invariant positions are shown by asterisks. Conserved blocks 1 and 2 mentioned under “Results” and “Discussion” are indicated. Amino acid positions are shown on the right.

![Sequence alignment of human SelX and similar proteins identified by data base searches.](image)

The amino acid sequences deduced from ORFs or ESTs were aligned with the PILB proteins. The mouse SelX contig is the translation of a cDNA contig constructed from several overlapping ESTs. Hyphens indicate gaps. In the mouse and human SelX sequences, U stands for selenocysteine, marked Sec below the sequence.

![Homologies in SelZf1, SelZf2, and TrxR2 cDNAs.](image)

A Computational Screen for Novel Selenoproteins

The C-terminal domains of SelZf1 and SelZf2 show clear formation in the bacteria *N. gonorrhoeae, H. influenzae, H. pylori, M. capricolum*, and *S. pneumoniae* (Fig. 6). PILB possesses a peptide methionine-sulfoxide reductase activity (25). Sequence comparisons established that this activity resides in a PILB subdomain different from the SelX similarity. The conserved amino acids in blocks 1 and 2 as well as the selenocysteine (Fig. 6) certainly play important roles in the function of SelX.

![Homologies in SelZf1, SelZf2, and TrxR2 cDNAs. Shown is a model for genomic organization. A, SelZf1, SelZf2, and TrxR2 cDNAs share common and specific regions. The similarly boxed common regions, the UGA selenocysteine codons, and the conserved SECIS elements are drawn. B, a BLAST search with the three cDNAs identified sequence similarities to cosmid 56c of chromosome 22q11.2 (GenBank™ accession number AC000090). The possible genomic organization, giving rise to the cDNAs in A, was obtained by joining the boxes according to the dashed line. Positions correspond to cosmid coordinates.](image)
homologies to the corresponding domain of the selenoprotein TrxR2. Interestingly, it was shown that the 293 bp at the 5′-end of the TrxR2 cDNA encode the mitochondrial targeting peptide (23), which is not found in SelZf2 (Fig. 7A). More surprisingly, the region of the cDNAs encoding the CNVNGC active site, common to the mitochondrial and cytoplasmic thioredoxin reductases and to the glutathione reductase (22), was found in the SelZf2 cDNA, but not in the SelZf1 cDNA. This suggests for SelZf1 a different function compared with SelZf2 and TrxR2. In the course of searching sequences similar to the SelZf1 and SelZf2 cDNAs, we identified genomic fragments (GenBank™ accession numbers AC000079 and AC000080) with similarity to both cDNAs. An identical genomic fragment was also shown independently by others (23) to contain se-