2004

Identification of Trace Element-Containing Proteins in Genomic Databases

Vadim N. Gladyshev
University of Nebraska-Lincoln, vgladyshev@rics.bwh.harvard.edu

Gregory V. Kryukov
University of Nebraska-Lincoln

Dmitri E. Fomenko
University of Nebraska-Lincoln, dfomenko2@unl.edu

Dolph L. Hatfield
National Institutes of Health, Bethesda, Maryland, hatfield@dc37a.nci.nih.gov

Annu. Rev. Nutr. 2004. 24:579–96
doi: 10.1146/annurev.nutr.24.012003.132241

First published online as a Review in Advance on March 12, 2004

IDENTIFICATION OF TRACE ELEMENT–CONTAINING PROTEINS IN GENOMIC DATABASES

Vadim N. Gladyshev,[1] Gregory V. Kryukov,[1] Dmitri E. Fomenko,[1] and Dolph L. Hatfield[2]

[1]Department of Biochemistry, University of Nebraska, Lincoln, Nebraska 68588-0664; email: vgladyshev1@unl.edu

[2]Section on Molecular Biology of Selenium, Laboratory of Cancer Prevention, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892; email: hatfield@dc37a.nci.nih.gov

Key Words bioinformatics, metal-containing proteins, selenium, selenoprotein, selenoproteome

■ **Abstract** Development of bioinformatics tools provided researchers with the ability to identify full sets of trace element–containing proteins in organisms for which complete genomic sequences are available. Recently, independent bioinformatics methods were used to identify all, or almost all, genes encoding selenocysteine-containing proteins in human, mouse, and *Drosophila* genomes, characterizing entire selenoproteomes in these organisms. It also should be possible to search for entire sets of other trace element–associated proteins, such as metal-containing proteins, although methods for their identification are still in development.

CONTENTS

0199-9885/04/0714-0579$14.00

INTRODUCTION

Recent advances in genomics have resulted in the generation of complete genomic sequences for a large number of organisms. The challenge now is to reliably annotate genes and regulatory elements and to determine functions of gene products. Functional annotation can be assisted by a variety of computational biology methods, such as analyses of gene occurrence, neighborhood, expression, and fusion (14, 15, 20, 48, 49). In addition to these sequence-independent methods, motif and structural analyses can yield information about protein function, but only if information is already available for other similar proteins that bear these features.

Analyses of protein sequences, on a genomic scale, for previously defined functional motifs, as well as previous direct biochemical experiments, revealed a significant number of proteins that bind micronutrients, including proteins that contain metals (Zn, Fe, Ca, Mg, Cu, Co, Mn, V, Ni, Mo, W) and other trace elements (Se, B). Trace elements have been adopted by biological systems because they provide proteins with unique coordination, catalytic, and electron transfer properties. These properties have been employed by organisms in key functions in a variety of pathways, resulting in dependence of organisms on several trace elements. In turn, trace elements can govern an organism's nutritional strategies and its evolution.

Trace elements occupy important sites in proteins, often being present as catalytic species or serving key structural functions. Therefore, the ability to identify trace element–containing proteins in genomic databases and to pinpoint the location of trace elements in proteins can greatly assist in functional annotation. Moreover, unique chemical properties of metals and other trace elements, their central locations with respect to protein function or structure, the availability of isotopes and, in many cases, the ability to manipulate their redox or coordination behavior can be exploited in a variety of ways using biophysical and spectroscopic techniques.

Analyses of full sets of trace element–containing proteins in organisms, however, have not been possible due to a lack of reliable computational tools for their identification. One exception is the recent analyses of complete or nearly complete sets of selenium-containing proteins in eukaryotes, including humans (11, 36, 40, 43, 50). In this review, we discuss these advances and describe how characterization of selenoproteomes can be exploited to improve our understanding of the biological functions of selenoproteins, the biomedical role of selenium, and the evolution of selenocysteine. We then discuss potential strategies for identification of proteins containing other trace elements.

SELENIUM-CONTAINING PROTEINS

Selenoproteins contain selenium in the form of selenocysteine (Sec). This amino acid is inserted into proteins cotranslationally in response to the UGA codon (6, 13, 30, 42, 47, 63). UGA differs from most of the other 63 codons in that it has a dual function: In addition to Sec insertion, UGA dictates the termination of protein synthesis. Other known dual uses of codons include AUG that codes for the initiation of protein synthesis and the internal insertion of methionine, and UAG, which was recently shown to code for pyrrolysine in certain archaea (29, 62). The use of UGA for Sec insertion is widespread in nature and has been described in bacteria (5, 6), archaea (57–59), and eukaryotes (13, 26, 30, 47). However, because the termination of protein synthesis is by far the most common function of UGA, available gene annotation programs interpret UGA only as stop. As a result, selenoprotein open reading frames (ORFs) have been misannotated or completely missed by gene prediction programs.

Selenoproteins typically contain only a single Sec residue, which is conserved among orthologous selenoproteins (30, 63). In functionally characterized proteins, Sec is always present at enzyme active sites. In addition, selenoproteins with known function have been observed to catalyze redox reactions. In spite of the essential function of Sec in selenoproteins, homologs (and even orthologs) in which the place of Sec is occupied by cysteine can be found in the same or other organisms (25) (Figure 1). However, Cys homologs are typically 10-fold to 1000-fold less active than selenoproteins (4, 63). The higher activity of Sec-containing proteins may be due to the unique ionization and redox properties of Sec.

The key role of Sec in catalysis by Sec-containing proteins makes selenium an essential trace element for many life forms, including humans. Disruption of Sec insertion in Sec tRNA knockout mice causes embryonic lethality (7). Low dietary intake of selenium results in decreased levels of selenoproteins, which compromises various physiological processes. Selenium plays important roles in decreasing cancer incidence (12), enhancing male reproduction (69), and supporting immune function, and has been implicated in slowing the aging process (30, 56). Selenoproteins are most likely responsible for most of the known biomedical effects of dietary selenium. Thus, information on a complete set of selenoproteins in various organisms, including humans, would provide the means for the systematic analyses of selenoprotein functions, help identify segments of the human population that may benefit most from dietary selenium supplements, and provide insights into disorders involving selenoprotein malfunction.

The essential component that distinguishes the Sec or stop function of UGA codons is the presence of an RNA stem-loop structure, designated the Sec insertion sequence (SECIS) element, in the 3′-untranslated region of eukaryotic mRNAs (3, 71) (Figure 2). In archaea, SECIS elements are also present in the 3′-untranslated regions (58, 59, 73), whereas in bacteria, these structures are located in coding regions, immediately downstream of Sec UGA codons (32, 45, 46, 60, 67). However, eukaryotic, archaeal, and bacterial SECIS elements, while conserved within individual domains of life, have no similarities to each other in terms of either conserved primary sequences or secondary structures (47).

The eukaryotic SECIS element is composed of two helices, internal and apical loops, and a quartet of non-Watson-Crick base pairs (13, 42, 47, 71) (Figure 2). The large apical loop can form an additional mini-helix that stabilizes the loop,

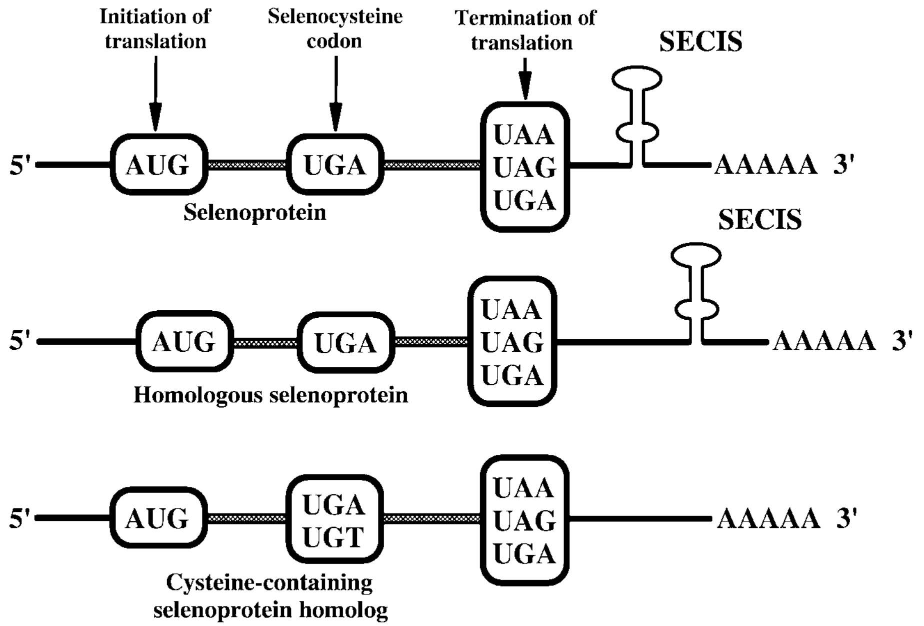

Figure 1 Organization of selenoprotein genes. Selenoprotein genes have in-frame UGA codons that code for selenocysteine (Sec) and RNA structures, Sec insertion sequence (SECIS) elements. In eukaryotes and archaea, SECIS elements are located in 3′-UTRs (*upper* and *middle* gene structures), whereas in bacteria they are present immediately downstream of UGA (not shown). Homologous selenoproteins conserve the UGA codon (compare *upper* and *middle* gene structures). Most selenoproteins also have homologs, in which cysteine (encoded by TGT or TGC) occupies the position of Sec (*lower* gene structure). Genes encoding Cys homologs do not have SECIS elements. Organization of selenoprotein genes suggested two methods for identification of these genes: by searching for SECIS elements (or pairs of orthologous SECIS elements in closely related genomes), and by searching for selenoprotein/Cys-containing protein homologous pairs.

and this has been the basis for separation of SECIS elements into type 1 (without the mini-helix) and type 2 (with the mini-helix) structures. The primary sequences of more than 95% mammalian SECIS elements contain an adenosine preceding the quartet, a TGA. . .GA motif in the quartet, and two adenosines in the apical loop or bulge (Figure 2). In addition, in two mammalian selenoprotein genes, the AA in the apical loop is replaced with CC (34, 40).

The mechanism of Sec insertion has been characterized in detail in bacteria (6), whereas some components of the Sec insertion machinery in eukaryotes and archaea are still missing (13, 42, 74). In bacteria, SECIS elements bind Sec-specific elongation factor SelB, which in turn recruits Sec tRNA (6). In archaea and

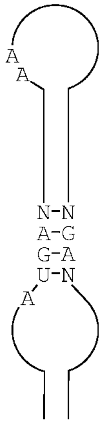

Figure 2 Eukaryotic consensus SECIS element. This element includes two helices and internal and apical loops. In addition, a quartet of non-Watson-Crick base pairs is located at the base of the upper helix. The UGA. . .GA sequence in the quartet is strictly conserved in eukaryotic SECIS elements, whereas the A that precedes the quartet and the AA in the apical loop are conserved in most, but not all, eukaryotic SECIS elements.

eukaryotes, the function of SelB is carried out by two proteins, SECIS-binding protein 2 (SBP2) and elongation factor EFSec, which form a complex, designated selenosome, in a Sec tRNA-dependent manner (13, 74). Sec tRNA is aminoacylated with serine, which is then converted to Sec by Sec synthase using selenophosphate, a selenium donor compound (28, 63). The identity of archaeal and eukaryotic Sec synthases is not known.

STRATEGIES FOR IDENTIFICATION OF SELENOPROTEIN GENES

The presence of conserved 3′-UTR RNA structures and cysteine-containing homologs provides the basis for two independent bioinformatics methods for selenoprotein identification. The first method can be used to identify selenoprotein genes by searching for SECIS elements; initial identification of these structures is followed with analysis of regions upstream of SECISes for ORFs containing conserved in-frame UGA codons (Figure 3). The second method can take advantage of the fact that most selenoprotein genes have homologs that contain cysteine in place of Sec (Figure 4). These homologs are often catalytically inefficient compared to selenoproteins but their presence allows an organism to avoid dependency on the supply of the trace element selenium (25, 26, 30, 63).

IDENTIFICATION OF SELENOPROTEINS BY SEARCHING FOR SECIS ELEMENTS

In 1999 two groups independently developed initial programs that could search for candidate SECIS elements in nucleotide sequences (36, 43). The application of these programs to mammalian EST collections identified three new selenoprotein genes, which subsequently were demonstrated to be true selenium-containing proteins in studies involving incorporation of ^{75}Se by metabolic labeling of cells. These studies marked the first examples of gene identification by searching for RNA structures on a genomic scale. The algorithm that was utilized by one of the programs, SECISearch (36), involved three steps:

1. A collection of nucleotide sequences was searched for the weak SECIS primary sequence consensus: ATGA/11–12 nucleotides/AA/18–27 nucleotides/GA (Figure 2);
2. Candidate sequences selected in the first step were analyzed for the SECIS element secondary structure consensus (helixes I and II and internal and apical loops); and
3. Free energies of candidate structures were evaluated.

This program adopted PatScan to recognize SECIS element primary sequences and secondary structures consensuses and Vienna RNA package RNAfold to evaluate separately the free energies of the structures for helix I plus internal loop and helix II plus apical loop regions of the putative SECIS element.

Subsequently, the SECISearch program was significantly improved and became sufficient to analyze smaller eukaryotic genomes. The computational screen of the entire *Drosophila* genome using this program resulted in identification of three selenoproteins: selenophosphate synthetase 2 (SPS2), G-rich, and BthD (50). Independently, the fruit fly genome was analyzed by the Geneid-based program (11), which identified the same set of selenoprotein genes, suggesting that both programs were advanced enough to analyze the entire animal genome. SECISearch was also applied to scan the nematode *C. elegans* genome (24 and unpublished data) and the green alga *Chlamydomonas* EST database (52). These experiments allowed a first view of eukaryotic selenoproteomes (see below).

However, the SECISearch and Geneid methods, even when combined, were not sufficient in identifying selenoprotein genes in larger sequences, such as mammalian genomes, because of their size and complexity. Therefore, a new approach was developed that was capable of identifying an entire set of human selenoprotein genes (40). The new program was organized as follows (Figure 3):

1. Candidate SECIS elements were identified in the human genome with SECISearch 2.0. The new version of the program, like the initial SECISearch, analyzed structural and thermodynamic features of SECIS elements; however, improvements in every step and better understanding of SECIS elements made it ~10-fold more selective (with the same specificity) than the original version of SECISearch.

2. Human/mouse and human/rat SECIS pairs were identified with a blastn-based program, which analyzed evolutionary conservation of mammalian SECIS elements. This program was the key to the successful analysis of the human genome. It was based on the observation that human, mouse, and rat SECIS elements in orthologous selenoprotein genes exhibited detectable sequence similarity. This step alone provided an increase of ~100-fold in the specificity of genomic searches.
3. Genomic sequences upstream of candidate SECIS elements were scanned with Geneid, a gene prediction program that identified ORFs with high coding potential and with in-frame TGA codons.
4. Predicted human selenoprotein genes were analyzed with mammalian selenoprotein gene signature (MSGS) criteria (38). These criteria screened selenoprotein homologs for the presence and conservation of genes, in-frame TGA codons, and SECIS elements.

Application of this algorithm to human, mouse, and rat genomes (40) identified 16 previously known selenoproteins (including those identified by the computational screen of EST databases) and 6 new selenoprotein genes (Table 1). Only two previously known mammalian selenoproteins were not found by these procedures. This was because the available human genome sequence did not contain the SPS2 gene. In addition, the thioredoxin reductase 2 [also known as thioredoxin/glutathione reductase (64, 65)] gene contained a SECIS version with a noncanonical nucleotide that preceded the quartet. Finally, the 25th selenoprotein, glutathione peroxidase 6 (GPx6), was identified in the human genome by homology searches. The computational screen did not find this gene because rodent orthologous genes encoded cysteine-containing proteins and therefore had no SECIS elements. Apparently, mouse and rat GPx6 genes replaced Sec with Cys, as evidenced by the presence of the remnant of the SECIS element in the mouse GPx6 3′-UTR (40).

Subsequent experiments involving metabolic labeling of cells with ^{75}Se confirmed Sec insertion into newly identified selenoproteins. Further statistical analyses suggested that the computational screens identified all or almost all selenoproteins common to human and rodent genomes. Thus, these studies established that the human selenoproteome consists of 25 selenoproteins, whereas mouse and rat selenoproteomes have 24 selenoproteins, and that these numbers likely represent complete sets of selenoprotein genes in these organisms (40).

SECIS-INDEPENDENT SEARCHES OF HUMAN SELENOPROTEIN GENES

An independent method for selenoprotein identification was also developed (Figure 4) (40). This method was based on the observation that most selenoproteins have homologs, in which Cys is present in place of Sec (25, 38). Thus, the search strategy of this SECIS-independent method was to identify selenoprotein

TABLE 1 Human selenoprotein genes. Sec location, protein length, and chromosomal location of selenoprotein genes are shown. All selenoproteins have a single Sec except SelP, which has 10 Sec residues.

Selenoprotein	Chromosomal location (number of exons)	Sec location in protein (length of protein)
Thyroid hormone deiodinase 1	1p32.3 (4)	126 (249)
Thyroid hormone deiodinase 2	14q31.1 (2)	133 (265)
Thyroid hormone deiodinase 3	14q32	144 (278)
Glutathione peroxidase 1	3p21.31 (2)	47 (201)
Glutathione peroxidase 2	14q23.3 (2)	40 (190)
Glutathione peroxidase 3	5q33.1 (5)	73 (226)
Glutathione peroxidase 4	19p13.3 (7)	73 (197)
Glutathione peroxidase 6	6p22.1 (5)	73 (221)
Selenoprotein H	11q12.1 (4)	44 (122)
Selenoprotein I	2p23.3 (10)	387 (397)
Selenoprotein K	3p21.31 (5)	92 (94)
Selenoprotein M	22q12.2 (5)	48 (145)
Selenoprotein N	1p36.11 (12)	428 (556)
Selenoprotein O	22q13.33 (9)	667 (669)
Selenoprotein P	5p12 (4)	59, 300, 318, 330, 345, 352, 367, 369, 376, 378 (381)
Selenoprotein R (methionine-R-sulfoxide reductase; MsrB)	16p13.3 (4)	95 (116)
Selenoprotein S	15q26.3 (6)	118 (189)
Sep15 (15 kDa selenoprotein)	1p22.3 (5)	93 (162)
Selenophosphate synthetase 2	—	60 (448)
Selenoprotein T	3q24 (6)	36 (182)
Thioredoxin reductase 1	12q23.3 (15)	498 (499)
Thioredoxin reductase 2 (TGR)	3q21.2 (16)	655 (656)
Thioredoxin reductase 3	22q11.21 (18)	522 (523)
Selenoprotein V	19q13.13 (6)	273 (346)
Selenoprotein W	19q13.32 (6)	13 (87)

genes by finding their homologs, in which Cys corresponded to the UGA codon (Figure 1).

This Sec/Cys homology method was applied to the human genome independently of the SECIS-based searches (40). First, all predicted human ORFs that contained in-frame TGA codons were identified with Geneid. The ORFs were then extended from TGA to the next terminator signal and analyzed by blastp and tblastn against all proteins that were previously predicted in completely sequenced eukaryotic genomes. This method was designed to identify pairs of sequences with homology in TGA-flanking regions, which either conserve TGA or replace TGA with TGC or TGT (cysteine codons). Second, all human proteins were analyzed against all human ESTs to identify paralogs containing TGA in place of a cysteine codon. These two Sec/Cys homology approaches recognized the majority of selenoprotein genes that were found through searches for SECIS elements, but did not identify additional selenoproteins, and thus provided further evidence that all or almost all mammalian selenoproteins were identified in the human genome.

HUMAN SELENOPROTEOME

The use of bioinformatics methods described above identified 10 new human selenoprotein genes in genomic and EST databases (34, 36, 39, 40, 43, 54). All of these were either incorrectly predicted or not detected at all in available human genome assemblies/annotations (e.g., Celera, NCBI, and Golden Path) (33, 41, 70). Analysis of organization of selenoprotein genes revealed that Sec was located in these proteins either upstream of an α-helix or very close to the C-terminus. Among the 10 new selenoproteins, GPx6 was a homolog of previously known mammalian glutathione peroxidases, whereas the C-terminal domain of SelV was homologous to SelW. Other new selenoproteins did not have known selenoprotein homologs and were validated by metabolic ^{75}Se labeling of cells that were transfected with selenoprotein constructs.

Several new selenoproteins exhibit interesting expression patterns. For example, the GPx6 mRNA was only detected in embryos and olfactory epithelium, whereas the SelV mRNA was expressed only in seminiferous tubules in testes (40). In addition, analysis of predicted secondary structures suggested that most human selenoproteins were globular proteins, with the exception of SelK, SelI, and SelS, which were predicted (and later demonstrated for SelK and SelS) to be plasma membrane proteins (40).

EUKARYOTIC SELENOPROTEOME

The number of selenoprotein genes in completely sequenced eukaryotic genomes varies from zero to more than two dozen. Whereas the human selenoproteome has 25 known proteins (Table 1), the mouse and rat genomes encode only 24 of these proteins. The difference is due to GPx6, which is a selenoprotein in humans (and pigs), but a Cys ortholog in mice and rats (40).

On the other hand, neither SECISearch-based methods nor searches for homologs of Sec insertion machinery genes identified selenoproteins in *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*. Thus, it appears that higher plants and yeast do not have the ability to insert Sec because they either lost selenoproteins during evolution or retained these proteins by replacing Sec with Cys. However, a member of the plant kingdom, green alga *Chlamydomonas reinhardtii*, has multiple selenoprotein genes (currently 10 known selenoproteins), suggesting that higher plants and yeast lost Sec independently (52, 55).

Compared to vertebrates, invertebrates have a small number of selenoproteins. For example, *D. melanogaster* genome codes for three selenoproteins (SPS2, G-rich, and BthD) (11, 50). It appears that these selenoproteins are responsible for the essential role of selenium in the fruit fly. *C. elegans* has only one known selenoprotein (thioredoxin reductase), which contains a single Sec residue (24 and unpublished data). Thus, it is possible that the entire machinery required for Sec insertion into protein is utilized for inserting a single residue into one protein in *C. elegans*. Whether Se is essential for nematodes is not known.

The largest numbers of selenoprotein genes appear to be in fish. For example, we identified more than 30 selenoprotein genes in zebrafish by homology analyses and searches of zebrafish dbEST with SECISearch (37; G.V. Kryukov & V.N. Gladyshev, unpublished data). This multiplicity of selenoprotein genes resulted from gene and genome duplications, which were followed by the retention of new selenoprotein genes. It was found that one zebrafish selenoprotein P gene (selenoprotein Pa) contained two SECIS elements and encoded a protein containing 17 Sec residues, the largest number of Sec residues found in any known protein (37). Zebrafish also had a second selenoprotein P gene (selenoprotein Pb), which contained one SECIS element and encoded a protein with a single Sec. These data revealed that the utilization of Sec is more common in fish than in mammals and perhaps in most other species. Interestingly, duplicated copies of fish selenoprotein genes, even those that are highly homologous to each other, exhibited distinct (and sometimes opposing) spatial and temporal expression patterns during embryogenesis in zebrafish (68). The reason why a greater number of selenoprotein genes occur in fish is not known, but could include the availability and relatively constant levels of selenium in seawater.

EVOLUTION OF SELENOCYSTEINE INSERTION

As described above, predicted secondary structures and Sec locations in human selenoproteins divided selenoproteins into two groups. In the first group, Sec was located upstream of an α-helix in $\alpha\beta$ domains and often was in close proximity to Cys as part of thioredoxin-like thiol-disulfide oxidoreductase motifs. In the second group, Sec was present in C-terminal extensions of other known domains. These observations suggest a mechanism for selenoprotein evolution. It appears that in organisms with a functional Sec insertion system, new selenoproteins genes evolve by replacement of a low pKa catalytic cysteine with Sec or by gene extension such that terminator or 3′-UTR TGA becomes a Sec codon.

In the former scenario, the Sec-for-Cys mutation coupled with generation of a SECIS element in the 3′-UTR enhances protein catalytic function, and the presence of a Sec-flanking α-helix may be viewed as a structural relic of accommodating a cysteine thiolate (35). With pKa ~5.5, Sec is completely ionized under physiological conditions and is a preferred residue in many redox enzymes.

In the second scenario, ORF extension may result in a new protein function. For example, animal thioredoxin reductase appeared to evolve from glutathione reductase such that the C-terminal selenotetrapeptide in the extension replaced glutathione as a substrate for the enzyme (53). Selenoprotein genes are also subject to evolutionary pressure to replace Sec with a more abundant amino acid. This may result in lineage-specific losses of selenoproteins or replacing Sec with Cys if the presence of Sec is no longer essential (e.g., mouse GPx6). The described view on selenoprotein evolution illustrates a dynamic process of steady accumulation of selenoprotein genes under selective evolutionary pressure of removing the Sec insertion machinery. This mechanism may provide tools to identify or design redox proteins, whose activity will be improved by replacing catalytic Cys with Sec.

STRATEGIES FOR IDENTIFICATION OF OTHER TRACE ELEMENT–ASSOCIATED PROTEINS

Can one use methods developed for characterization of selenoproteomes to identify metal-containing proteins? A direct application of these methods is difficult because metals are not inserted into proteins cotranslationally and RNA structures do not indicate their presence. However, understanding the properties of metal-containing proteins, and in particular, the chemistry of their binding to protein ligands, resulted in development of numerous metal-binding motifs and patterns (or motifs/patterns for cofactors that bind metals).

These bioinformatics features provide the basis for a number of commercial and public tools, which by analyzing conserved protein domains and protein functions provide predictions for sequences of interest. These tools, for example, include PROSITE, a database of protein families and domains (http://us.expasy.org/prosite/) (16, 21); InterPro, integrated resources of protein families, domains, and functional sites (www.ebi.ac.uk/interpro/) (51); BLOCKS, a blocks database (www.blocks.fhcrc.org/) (31); Pfam, protein families database (www.sanger.ac.uk/Software/Pfam/) (2); PRINTS, a protein motif fingerprint database (http://bioinf.man.ac.uk/dbbrowser/PRINTS/) (1); ProDom, a protein domain database (http://protein.toulouse.inra.fr/prodom/current/html/home.php/) (61); SMART, simple modular architecture research tool (http://smart.embl-heidelberg.de/) (44); COGs, clusters of orthologous groups (www.ncbi.nlm.nih.gov/) (66); and CDART, conserved domain architecture retrieval tool (www.ncbi.nlm.nih.gov/) (72). There is also a database that contains metalloproteins with known structures (http://metallo.scripps.edu/) (10).

Some of these tools use patterns, which are occurrences of specific clusters of amino acids in protein sequences. Patterns cover small, typically highly conserved regions of amino acid sequences and thus could be used for identification of proteins with high similarity. In addition, entire functional domains could be described based on multiple sequence alignment profiles, even if primary sequence conservation is low (8, 9, 27). Profile is a matrix of position-specific amino acid weight and gap costs, which describes similarity among sequences. For example, the PROSITE database contains a combination of Hidden Markov Model-like profiles and sequence patterns. At present, this database contains 1235 documentation entries that describe 1676 different patterns and profiles, and thus could be used for identification of metal-binding proteins with excellent sensitivity and selectivity.

PROSITE patterns and profiles were developed for proteins that contain zinc, iron, copper, magnesium, calcium, cobalt, nickel, molybdenum (together with tungsten), manganese, and vanadium (numbers of PROSITE patterns and profiles that describe metal-containing and metal-binding proteins are shown in Table 2). In some of these patterns and profiles, amino acids that are directly involved in metal coordination were identified using structural data and mutagenesis data. Cysteine and histidine appear to be the amino acids that are most often coordinate metals (23), although they do not bind all metals.

The results of application of available PROSITE motifs and patterns for metal-binding proteins to completely sequenced genomes (D.E. Fomenko & V.N. Gladyshev, unpublished) are shown in Table 3. In contrast to the studies on selenoproteins, these numbers do not reflect the actual numbers of metal-containing proteins present in genomes. This is because PROSITE patterns do not describe all metal-binding motifs, and in addition, some of them could have a low false-positive rate. Moreover, many protein databases do not contain full sets of proteins encoded in genomes and may contain some pseudogenes and duplicated entries. However, the PROSITE searches can even now reveal the extent of the use of

TABLE 2 Distribution of patterns and profiles for metal-binding proteins in the PROSITE database

Metal	PROSITE families	Patterns	Profiles
Zinc	66	77	32
Iron	48	74	5
Calcium	15	24	4
Copper	16	22	4
Magnesium	15	21	0
Nickel	6	10	0
Manganese	5	8	0
Cobalt	1	2	0
Molybdenum	2	3	0

TABLE 3 Metal-containing and metal-binding proteins identified with PROSITE patterns and profiles in completely sequenced genomes

Organism	**Total proteins**	**Zn**	**Cu**	**Mg**	**Fe**	**Ca**	**Ni**	**Co**	**Mo**
Homo sapiens	25,319	925	31	34	86	59	0	4	6
Mus musculus	35,726	673	29	38	94	45	0	4	4
Caenorhabditis elegans	21,125	457	21	26	59	24	0	1	5
Drosophila melanogaster	18,107	523	23	37	66	17	1	2	5
Arabidopsis thaliana	27,243	536	19	51	81	14	1	4	6
Saccharomyces cerevisiae	6337	163	9	23	26	2	0	2	0
Escherichia coli K12	4279	45	4	11	57	0	7	1	7
Methanococcus jannaschii	1785	7	0	3	21	0	4	1	2

specific metals in various genomes. For example, it is clear that the use of metals in proteins increased during evolution of eukaryotes (evidenced by the increase in the number of metal-containing proteins). However, the increase is not proportional for each metal. Clearly, the use of Ca and Cu, and especially Zn, increased, whereas the use of Fe, Ni, Mg, Mo, and Co was not significantly changed or even decreased during evolution from simple organisms to mammals (Table 3).

The increased use of zinc in eukaryotic proteins, however, uncovers a problem of reliable identification of these proteins. Zinc is often coordinated by cysteines, but these residues are also often used for intermolecular or intramolecular disulfide bond formation (22). For example, out of 77 available zinc-binding protein patterns, 36 patterns contain at least one cysteine, including 30 patterns containing at least one CxxC motif (two cysteines separated by two other amino acids). Likewise, iron-binding proteins are described in 74 patterns, 31 of which contain at least one cysteine, including 15 that are present in the form of the CxxC motif.

The CxxC motif is also a major redox motif in thiol/disulfide oxidoreductases, and two cysteines in this motif are directly involved in disulfide bond formation, reduction, or isomerization (17–19). At present, there are no reliable criteria for distinguishing between redox and metal-binding CxxC motifs. However, whereas the redox potential of CxxC motifs in thiol/disulfide oxidoreductases is finely tuned to perform specific redox functions, metal coordination is likely less influenced by residues that flank cysteine. In addition, the thiol/disulfide function requires only two cysteines, whereas four ligands are often used for metal coordination. These properties might be useful in developing algorithms for distinguishing metal-binding and thiol-dependent redox proteins. Cysteine is one of the least abundant, but most conserved, amino acids in proteins. Similarly, the presence of conserved histidines in proteins is also scarce. Thus, the presence of closely located conserved cysteines and histidines, separately or in combination, could also potentially be used to distinguish these proteins from thiol/disulfide oxidoreductases.

CONCLUSIONS

We are witnessing a dramatic increase in the application of bioinformatics tools to answer questions about the use of trace elements in biology. A full set of selenocysteine-containing proteins is now known for several eukaryotic genomes, providing important information about biological functions of selenium. It should be possible to extend bioinformatics methods to identify selenoproteins in most other completely sequenced genomes of eukaryotes, archaea, and bacteria, and to link specific selenoproteins with biological functions of selenium.

The use of bioinformatics methods for characterization of metalloproteomes lags behind that of selenoproteomes. This is because specific tools for identification of metal-binding proteins require further development. However, searches for metal-containing proteins are possible through motif, pattern, and profile searches. While the use of these tools is currently insufficient in detecting all metal-containing proteins, these approaches can already be extremely useful in assessing relative compositions of metalloproteomes, and in many situations, can provide a nearly complete set of proteins that bind a specific metal in an organism.

As the content of genomic and protein databases grows and the bioinformatics tools for identification of trace element–containing proteins become more refined, there is every reason to believe that most of these proteins can be identified in completely sequenced genomic databases within the next few years. These studies will be essential in fully understanding the roles that trace elements play in biology.

ACKNOWLEDGMENTS

We thank Dr. Alexey Lobanov for help with Figure 2. This work is supported by NIH grant GM061603.

The *Annual Review of Nutrition* is online at http://nutr.annualreviews.org

LITERATURE CITED

1. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, et al. 2003. PRINTS and its automatic supplement, prePRINTS. *Nucl. Acids Res.* 31:400–2
2. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, et al. 2002. The Pfam Protein Families Database. *Nucleic Acids Res.* 30: 276–80
3. Berry MJ, Banu L, Chen YY, Mandel SJ, Kieffer JD, et al. 1991. Recognition of UGA as a selenocysteine codon in type I deiodinase requires sequences in the 3′ untranslated region. *Nature* 353:273–76
4. Berry MJ, Maia AL, Kieffer JD, Harney JW, Larsen PR. 1992. Substitution of cysteine for selenocysteine in type I iodothyronine deiodinase reduces the catalytic efficiency of the protein but enhances its translation. *Endocrinology* 131:1848–52
5. Bock A, Forchhammer K, Heider J, Baron C. 1991. Selenoprotein synthesis: an expansion of the genetic code. *Trends Biochem. Sci.* 16:463–67
6. Bock A. 2000. Biosynthesis of selenoproteins—an overview. *Biofactors* 11:77–78

7. Bosl MR, Takaku K, Oshima M, Nishimura S, Taketo MM. 1997. Early embryonic lethality caused by targeted disruption of the mouse selenocysteine tRNA gene (Trsp). *Proc. Natl. Acad. Sci. USA* 94: 5531–34
8. Bucher P, Bairoch A. 1994. A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, ed. R Altman, D Brutlag, P Karp, R Lathrop, D Searls, pp. 53–61. Menlo Park, CA: AAAI Press
9. Bucher P, Karplus K, Moeri N, Hofmann K. 1996. A flexible motif search technique based on generalized profiles. *Comput. Chem.* 20:3–23
10. Castagnetto JM, Hennessy SW, Roberts VA, Getzoff ED, Tainer JA, Pique ME. 2002. MDB: the Metalloprotein Database and Browser at The Scripps Research Institute. *Nucleic Acids Res.* 30:379–82
11. Castellano S, Morozova N, Morey M, Berry MJ, Serras F, et al. 2001. In silico identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Rep.* 2:697–702
12. Combs GF, Clark LC, Turnbull BW. 2001. An analysis of cancer prevention by selenium. *Biofactors* 14:153–59
13. Driscoll DM, Copeland PR. 2003. Mechanism and regulation of selenoprotein synthesis. *Annu. Rev. Nutr.* 23:17–40
14. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. 2000. Protein function in the post-genomic era. *Nature* 405:823–26
15. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402:86–90
16. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, et al. 2002. The PROSITE database: its status in 2002. *Nucl. Acids Res.* 30:235–38
17. Fomenko DE, Gladyshev VN. 2003. Genomics perspective on disulfide bond formation. *Antioxid. Redox Signal.* 5:397–402
18. Fomenko DE, Gladyshev VN. 2003. Identity and functions of CxxC-derived motifs. *Biochemistry* 42:11214–25
19. Fomenko DE, Gladyshev VN. 2002. CxxS: Fold-independent redox motif revealed by genome-wide searches for thiol/disulfide oxidoreductase function. *Protein Sci.* 11: 2285–96
20. Galperin MY, Koonin EV. 2000, Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.* 18:609–13
21. Gattiker A, Gasteiger E, Bairoch A. 2002. ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl. Bioinform.* 1:107–8
22. Giles NM, Watts AB, Giles GI, Fry FH, Littlechild JA, Jacob C. 2003. Metal and redox modulation of cysteine protein function. *Chem. Biol.* 10:677–93
23. Gitschier J, Moffat B, Reilly D, Wood WI, Fairbrother WJ. 1998. Solution structure of the fourth metal-binding domain from the Menkes copper-transporting ATPase. *Nat. Struct. Biol.* 5:47–54
24. Gladyshev VN, Krause M, Xu XM, Korotkov KV, Kryukov GV, et al. 1999. Selenium-containing thioredoxin reductase in *C. elegans*. *Biochem. Biophys. Res. Commun.* 259:244–49
25. Gladyshev VN, Kryukov GV. 2001. Evolution of selenocysteine-containing proteins: significance of identification and functional characterization of selenoproteins. *BioFactors* 14:87–92
26. Gladyshev VN. 2001. Identity, evolution and functions of selenoproteins and selenoprotein genes. In *Selenium: Its Molecular Biology and Role in Human Health*, ed. DL Hatfield, pp. 99–113. Norwell, MA: Kluwer Acad.
27. Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* 84:4355–58
28. Guimaraes MJ, Peterson D, Vicari A,

Cocks BG, Copeland NG, et al. 1996. Identification of a novel selD homolog from eukaryotes, bacteria, and archaea: Is there an autoregulatory mechanism in selenocysteine metabolism? *Proc. Natl. Acad. Sci. USA* 93:15086–91

29. Hao B, Gong W, Ferguson TK, James CM, Krzycki JA, Chan MK. 2002. A new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science* 296:1462–66
30. Hatfield DL, Gladyshev VN. 2002. How selenium has altered our understanding of the genetic code. *Mol. Cell. Biol.* 22:3565–76
31. Henikoff JG, Greene EA, Pietrokovski S, Henikoff S. 2000. Increased coverage of protein families with the Blocks Database servers. *Nucleic Acids Res.* 28:228–30
32. Huttenhofer A, Westhof E, Bock A. 1996. Solution structure of mRNA hairpins promoting selenocysteine incorporation in *Escherichia coli* and their base-specific interaction with special elongation factor SELB. *RNA* 2:354–66
33. Kent JW, Haussler D. 2001. Assembly of the working draft of the human genome with GigAssembler. *Genome Res.* 11:1541–48
34. Korotkov KV, Novoselov SV, Hatfield DL, Gladyshev VN. 2002. Mammalian selenoprotein in which selenocysteine incorporation is supported by a new form of SECIS element. *Mol. Cell. Biol.* 22:1402–11
35. Kortemme T, Creighton TE. 1995. Ionization of cysteine residues at the termini of model alpha-helical peptides. Relevance to unusual thiol pKa values in proteins of the thioredoxin family. *J. Mol. Biol.* 253:799–812
36. Kryukov GV, Kryukov VM, Gladyshev VN. 1999. New mammalian selenoproteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *J. Biol. Chem.* 274:33888–97
37. Kryukov GV, Gladyshev VN. 2000. Selenium metabolism in zebrafish: multiplicity of selenoprotein genes and expression of a protein containing seventeen selenocysteine residues. *Genes Cells* 5:1049–60
38. Kryukov GV, Gladyshev VN. 2002. Mammalian Selenoprotein Gene Signature: identification and functional analysis of selenoprotein genes using bioinformatics methods. *Methods Enzymol.* 347:84–100
39. Kryukov GV, Kumar RA, Koc A, Sun Z, Gladyshev VN. 2002. Selenoprotein R is a zinc-containing stereospecific methionine sulfoxide reductase. *Proc. Natl. Acad. Sci. USA* 99:4245–50
40. Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, et al. 2003. Characterization of mammalian selenoproteomes. *Science* 300:1439–43
41. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
42. Lescure A, Fagegaltier D, Carbon P, Krol A. 2002. Protein factors mediating selenoprotein synthesis. *Curr. Protein. Pept. Sci.* 3:143–51
43. Lescure A, Gautheret D, Carbon P, Krol A. 1999. Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif. *J. Biol. Chem.* 274:38147–54
44. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, et al. 2002. Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* 30:242–44
45. Liu Z, Reches M, Groisman I, Engelberg-Kulka H. 1998. The nature of the minimal "selenocysteine insertion sequence" (SECIS) in *Escherichia coli*. *Nucleic Acids Res.* 26:896–902
46. Liu Z, Reches M, Engelberg-Kulka H. 1999. A sequence in the *Escherichia coli* fdhF "selenocysteine insertion sequence" (SECIS) operates in the absence of selenium. *J. Mol. Biol.* 294:1073–86
47. Low SC, Berry MJ. 1996. Knowing when not to stop: selenocysteine incorporation in eukaryotes. *Trends Biochem. Sci.* 21:203–8
48. Marcotte EM, Pellegrini M, Ng HL, Rice

DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science* 285:751–53

49. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* 402:83–86
50. Martin-Romero FJ, Kryukov GV, Lobanov AV, Carlson BA, Lee BJ, et al. 2001. Selenium metabolism in *Drosophila*: selenoproteins, selenoprotein mRNA expression, fertility, and mortality. *J. Biol. Chem.* 276:29798–804
51. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, et al. 2003. The InterPro Database: 2003 brings increased coverage and new features. *Nucleic Acids. Res.* 31:315–18
52. Novoselov SV, Rao M, Onoshko NV, Zhi H, Kryukov GV, et al. 2002. Selenoproteins and selenocysteine insertion system in the model plant cell system, *Chlamydomonas reinhardtii*. *EMBO J.* 21:3681–93
53. Novoselov SV, Gladyshev VN. 2003. Nonanimal origin of animal thioredoxin reductases: implications for selenocysteine evolution and evolution of protein function through carboxy-terminal extensions. *Protein Sci.* 12:372–78
54. Petit N, Lescure A, Rederstorff M, Krol A, Moghadaszadeh B, et al. 2003. Selenoprotein N: an endoplasmic reticulum glycoprotein with an early developmental expression pattern. *Hum. Mol. Genet.* 12:1045–53
55. Rao M, Carlson BA, Novoselov SV, Weeks DP, Gladyshev VN, Hatfield DL. 2003. *Chlamydomonas reinhardtii* selenocysteine tRNA. *RNA* 9:923–30
56. Rayman MP. 2000. The importance of selenium to human health. *Lancet* 356:233–41
57. Rother M, Wilting R, Commans S, Bock A. 2000. Identification and characterisation of the selenocysteine-specific translation factor SelB from the archaeon *Methanococcus jannaschii*. *J. Mol. Biol.* 299:351–58
58. Rother M, Resch A, Gardner WL, Whitman WB, Bock A. 2001. Heterologous expression of archaeal selenoprotein genes directed by the SECIS element located in the 3′ non-translated region. *Mol. Microbiol.* 40:900–8
59. Rother M, Resch A, Wilting R, Bock A. 2001. Selenoprotein synthesis in archaea. *Biofactors* 14:75–83
60. Sandman KE, Tardiff DF, Neely LA, Noren CJ. 2003. Revised *Escherichia coli* selenocysteine insertion requirements determined by in vivo screening of combinatorial libraries of SECIS variants. *Nucleic Acids Res.* 31:2234–41
61. Servant F, Bru C, Carrère S, Courcelle E, Gouzy J, et al. 2002. ProDom: automated clustering of homologous domains. *Brief. Bioinform.* 3:246–51
62. Srinivasan G, James CM, Krzycki JA. 2002. Pyrrolysine encoded by UAG in archaea: charging of a UAG-decoding specialized tRNA. *Science* 296:1459–62
63. Stadtman TC. 1996. Selenocysteine. *Annu. Rev. Biochem.* 65:83–100
64. Sun QA, Wu Y, Zappacosta F, Jeang KT, Lee B, et al. 1999. Redox regulation of cell signaling by selenocysteine in thioredoxin reductases. *J. Biol. Chem.* 274:24522–30
65. Sun QA, Kirnarsky L, Sherman S, Gladyshev VN. 2001. Selenoprotein oxidoreductase with specificity for thioredoxin and glutathione systems. *Proc. Natl. Acad. Sci. USA* 98:3673–78
66. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinform.* 4:41
67. Thanbichler M, Bock A. 2002. The function of SECIS RNA in translational control of gene expression in *Escherichia coli*. *EMBO J.* 21:6925–34
68. Thisse C, Degrave A, Kryukov GV, Gladyshev VN, Obrecht-Pflumio S, et al. 2003. Spatial and temporal expression patterns of selenoprotein genes during embryogenesis in zebrafish. *Gene Expr. Patterns* 3:525–32
69. Ursini F, Heim S, Kiess M, Maiorino M, Roveri A, et al. 1999. Dual function of the

selenoprotein PHGPx during sperm maturation. *Science* 285:1393–96

70. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
71. Walczak R, Westhof E, Carbon P, Krol A. 1996. A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *RNA* 2:367–79
72. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, et al. 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* 31:28–33
73. Wilting R, Schorling S, Persson BC, Böck A. 1997. Selenoprotein synthesis in archaea: identification of an mRNA element of *Methanococcus jannaschii* probably directing selenocysteine insertion. *J. Mol. Biol.* 266:637–41
74. Zavacki AM, Mansell JB, Chung M, Klimovitsky B, Harney JW, Berry MJ. 2003. Coupled tRNA(Sec)-dependent assembly of the selenocysteine decoding apparatus. *Mol. Cell* 11:773–81

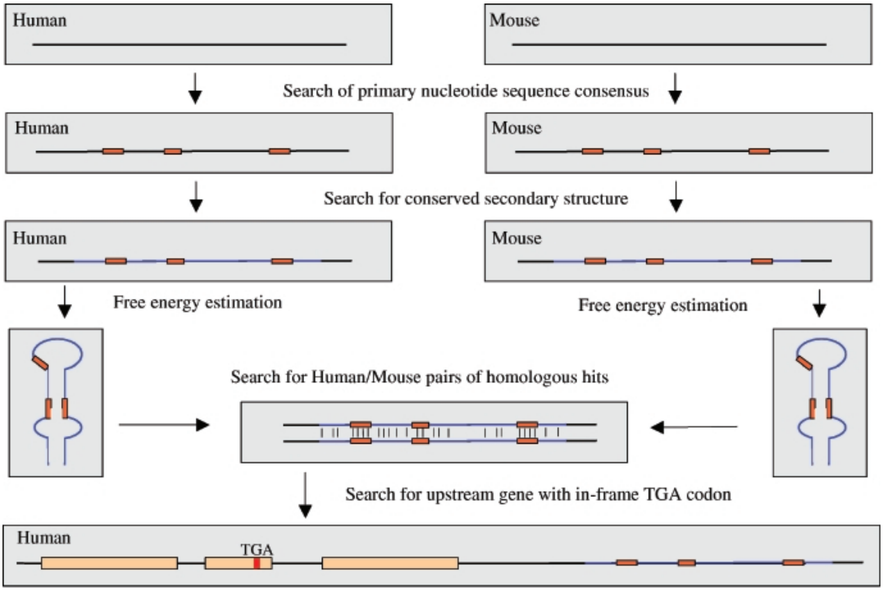
Human
Mouse
Search of primary nucleotide sequence consensus
Human
Mouse
Search for conserved secondary structure
Human
Mouse
Free energy estimation
Free energy estimation
Search for Human/Mouse pairs of homologous hits
Search for upstream gene with in-frame TGA codon
Human
TGA

Figure 3 An algorithm for a SECIS element–based search for selenoprotein genes. Selenoprotein genes are identified by searches for SECIS elements (or by searches for pairs of homologous SECIS elements in large, evolutionarily close genomes).

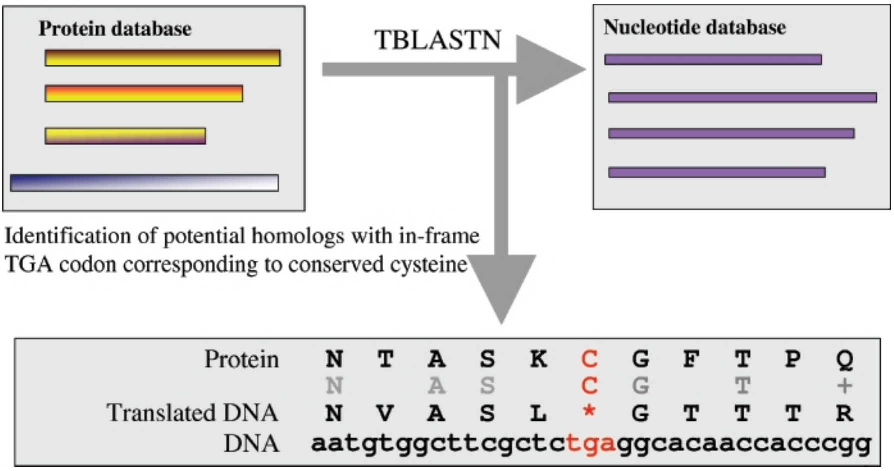
Protein database
TBLASTN
Nucleotide database
Identification of potential homologs with in-frame
TGA codon corresponding to conserved cysteine
Protein N T A S K C G F T P Q
N A S C G T +
Translated DNA N V A S L * G T T T R
DNA aatgtggcttcgctctgaggcacaaccacccgg

Figure 4 An algorithm for a Sec/Cys pair–based search for selenoprotein genes.