January 2007

High-Throughput Identification of Catalytic Redox-Active Cysteine Residues

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

Weibing Xing
University of North Carolina, Chapel Hill, NC

Blakely M. Adair
U.S. Environmental Protection Agency

David J. Thomas
U.S. Environmental Protection Agency

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

High-Throughput Identification of Catalytic Redox-Active Cysteine Residues

Dmitri E. Fomenko,[1] Weibing Xing,[2] Blakely M. Adair,[3] David J. Thomas,[3] Vadim N. Gladyshev[1]*

Cysteine (Cys) residues often play critical roles in proteins; however, identification of their specific functions has been limited to case-by-case experimental approaches. We developed a procedure for high-throughput identification of catalytic redox-active Cys in proteins by searching for sporadic selenocysteine-Cys pairs in sequence databases. This method is independent of protein family, structure, and taxon. We used it to selectively detect the majority of known proteins with redox-active Cys and to make additional predictions, one of which was verified. Rapid accumulation of sequence information from genomic and metagenomic projects should allow detection of many additional oxidoreductase families as well as identification of redox-active Cys in these proteins.

Relative to other amino acids in proteins, cysteine (Cys) residues are often conserved and functionally important. Cys residues critical to protein function are frequently used at enzyme active sites; they may bind metals such as iron, zinc, and copper; and they may be subjected to posttranslational modifications that regulate protein function or target proteins to a particular cellular location. A Cys pair may form a disulfide bond that stabilizes protein structure (*1*), and numerous additional roles of Cys have been described.

However, the presence of Cys in a protein sequence per se does not mean that this residue has a critical function, nor does it establish what that function might be. Analyses of Cys conservation may help identify catalytic Cys, but only for proteins with already known function (*2*–*4*). Cys is chemically similar to selenocysteine (Sec), a rare but widely distributed amino acid encoded by UGA and known as the 21st amino acid in the genetic code (*5*–*9*). In functionally characterized selenoprotein families, Sec is located in enzyme active sites and serves various redox functions. In addition, most selenoproteins have close homologs in which Sec is replaced by Cys.

Selenoproteins are found in all three domains of life, but a given organism contains a relatively small number of these proteins. A core of about 20 selenoproteins is widespread; however, many organisms also evolved rare or lineage-specific selenoproteins. For example, methionine-*S*-sulfoxide reductase (MsrA) occurs as a selenoprotein in green algae and a Cys-containing form in most other species (*9*), whereas the Cys-containing methionine-*R*-sulfoxide reductase (MsrB) evolved into a selenoprotein in vertebrates (*10*). Although Sec occurs only in several MsrA and MsrB sequences, it aligns with the catalytic Cys in other members of these protein families. Only the catalytic Cys aligns with Sec, whereas other Cys residues, such as the conserved Cys that serve as resolving residues or coordinate zinc, do not have Sec pairs. Thus, identification of rare Sec-containing forms of MsrA and MsrB indicates the thiol-based redox function of these proteins. Moreover, it pinpoints the location of catalytic Cys in these protein families. Similarly, the occurrence of rare Sec-containing forms of other protein families containing conserved Cys may be indicative of redox functions of these families.

We used this property of homologous Cys-Sec pairs to develop a procedure for large-scale identification of catalytic redox-active Cys in proteins. To identify Cys-Sec pairs, we analyzed, with tblastn (*11*), all Cys-containing proteins in National Center for Biotechnology Information (NCBI) databases against NCBI nucleotide sequences to identify pairs of homologous protein sequences, in which one sequence was a Cys-containing protein and the other was a candidate selenoprotein derived from a nucleotide sequence translated in all six open reading frames (ORFs) (i.e., Cys in a protein sequence aligned to TGA in a translated nucleotide sequence) (Fig. 1A). Using this procedure, we screened major nucleotide sequence databases, including completed genomes as well as environmental, whole-genome shotgun (WGS), and expressed sequence tag (EST) sequences. Independently, blastp (*11*) searches were carried out to identify Cys-containing homologs of a manually compiled set of previously identified selenoproteins. The two data sets were then combined and clustered into protein families, and a nonredundant set of Cys proteins in which each protein contained a predicted catalytic redox-active Cys was developed.

[1]Department of Biochemistry, University of Nebraska, Lincoln, NE 68588, USA. [2]Curriculum in Toxicology, University of North Carolina, Chapel Hill, NC 27599, USA. [3]Experimental Toxicology Division, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC 27709, USA.

*To whom correspondence should be addressed. E-mail: vgladyshev1@unl.edu

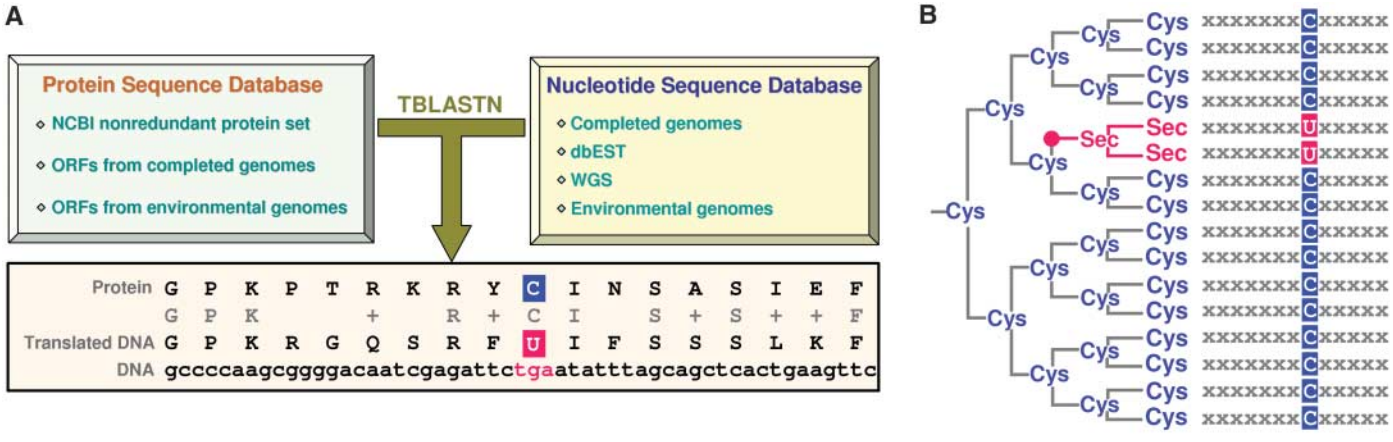

Fig. 1. High-throughout prediction of redox-active cysteine residues. (**A**) Scheme illustrating the method used for prediction of redox-active Cys residues by homology to sporadic selenoproteins [(*12*); U denotes selenocysteine]. MsrB is used as an example that illustrates the Cys-Sec pairs. (**B**) Model of evolution of proteins containing catalytic redox-active Cys. Redox-active Cys residues are conserved during evolution. In some organisms, these proteins may evolve into selenoproteins such that Sec replaces the catalytic Cys. Detection of such Cys-Sec pairs could reveal redox function for the entire protein family (or subfamily) and indicate location of the redox-active Cys.

The set of proteins containing predicted redox-active Cys had 10,412 unique sequences, which were organized into 40 protein families and superfamilies (Table 1 and table S1). Each sequence had a conserved Cys that corresponded to Sec in at least one homolog found in any of the analyzed sequence databases. The functional diversity of the identified proteins exceeded the 40 families, as some proteins with distinct functions belonged to the same family. For example, thioredoxins, protein disulfide isomerases, DsbA, DsbC, and DsbG were present in one cluster, although each of these proteins has a distinct function. The 10,412 Cys-containing sequences accounted for ~0.5% of the initial set of tested proteins and represented all completely sequenced genomes. The number of proteins found in a particular genome generally correlated with the size of the proteome, although archaea had fewer such proteins than other organisms (figs. S1 and S2).

We divided proteins represented by Cys-Sec pairs into functionally characterized proteins and proteins of unknown function. Most functionally characterized proteins detected in our search were those that used redox-active Cys in the active site for thiol-based redox catalysis, including the well-known oxidoreductases thioredoxin (fig. S3), glutaredoxin (fig. S4), peroxiredoxin (fig. S5), and glutathione peroxidase (fig. S6). Many contained a CxxC motif (two Cys separated by two residues) or motifs derived from it [e.g., CxxS in arsenate reductase (fig. S7), TxxC in peroxiredoxin, and CxxT in glutathione peroxidase] (*12*). A common property of the detected proteins was the use of a conserved nucleophilic redox-active Cys residue. During catalysis, this Cys changes redox state to a disulfide (e.g., in thioredoxin, glutaredoxin, and AhpD) or a sulfenic acid intermediate (e.g., in glutathione peroxidase, peroxiredoxin, MsrA, and MsrB). These observations suggested that the remaining Cys-containing proteins detected in the search may also be redox proteins that use redox-active Cys residues.

Conserved Cys residues are frequently used to coordinate metal ions (most often zinc, iron, and copper). However, such Cys were not present in our data set. Likewise, the data set had no proteins in which catalytic Cys residues carried out nonredox functions or were involved in posttranslational modifications or structural disulfides. Further analysis (table S1) revealed that the identified proteins represented

Table 1. Proteins identified in the searches for Cys-Sec pairs in homologous sequences. Each cell shows numbers of detected redox-active Cys–containing sequences (left) and numbers of Sec-containing sequences that support predictions of redox-active Cys residues (right). See also table S1.

Proteins	Bacterial Cys-Sec sequences	Archaeal Cys-Sec sequences	Eukaryotic Cys-Sec sequences
Functionally characterized proteins containing catalytic redox-active Cys	6222/363	257/21	1514/194
Proteins with predicted redox function and catalytic redox-active Cys	2045/256	59/9	315/360
Total	8267/619	316/30	1829/554

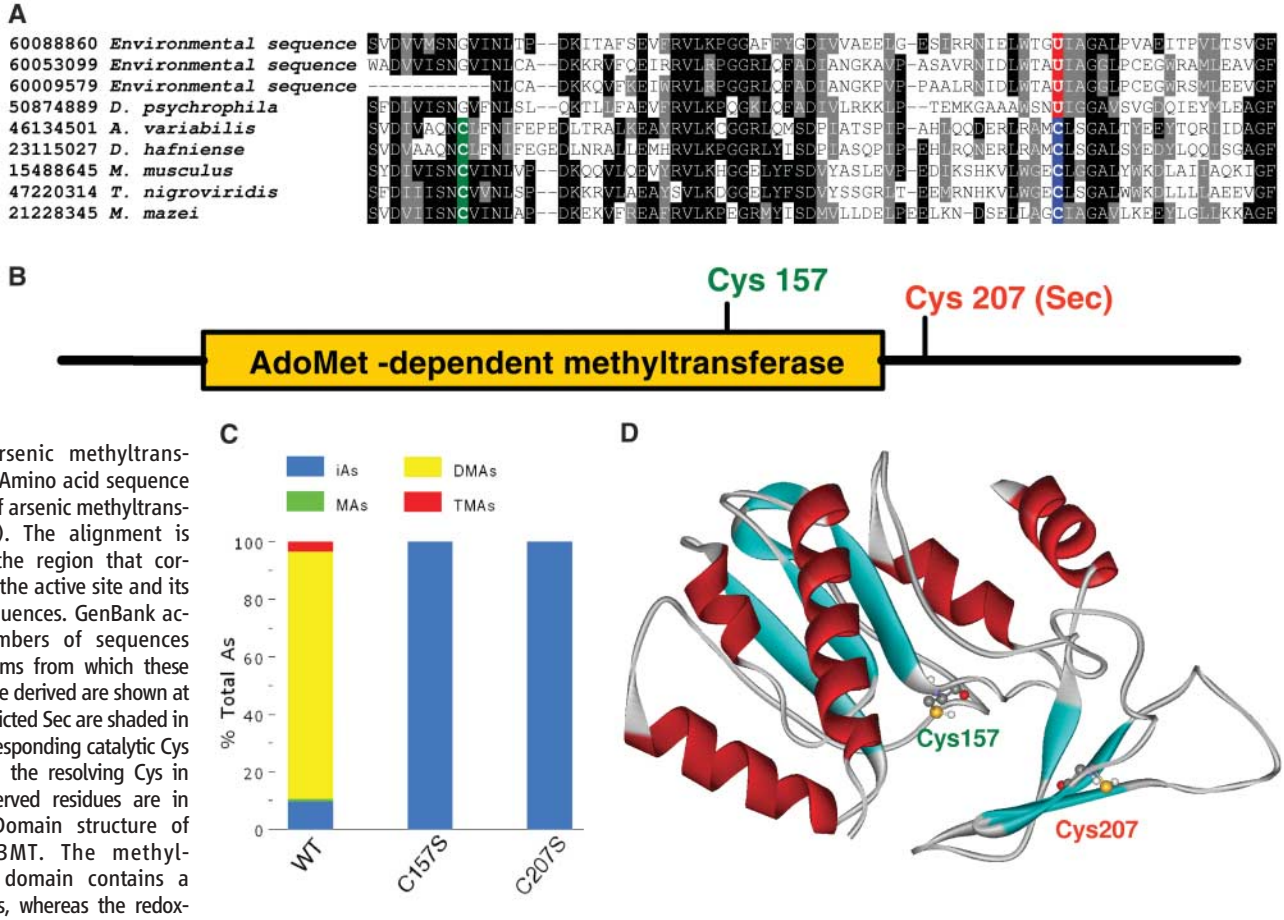

Fig. 2. Arsenic methyltransferases. (**A**) Amino acid sequence alignment of arsenic methyltransferases (*12*). The alignment is limited to the region that corresponds to the active site and its flanking sequences. GenBank accession numbers of sequences and organisms from which these sequences are derived are shown at the left. Predicted Sec are shaded in red, the corresponding catalytic Cys in blue, and the resolving Cys in green; conserved residues are in gray. (**B**) Domain structure of mouse AS3MT. The methyltransferase domain contains a resolving Cys, whereas the redox-active Cys-Sec is located downstream of this domain. (**C**) Activity of mouse AS3MT and its Cys^{157} → Ser and Cys^{207} → Ser mutants. AS3MT catalyzes AdoMet-dependent methylation of inorganic arsenicals (iAs) to its methylated forms. MAs, monomethylated arsenicals; DMAs, dimethylated arsenicals; TMAs, trimethylated arsenicals. (**D**) Structural model of mouse AS3MT. Cys^{207} and Cys^{157} are shown as ball-and-stick models.

essentially all known families of oxidoreductases that carry out thiol-based catalysis. For example, both protein families that reduce methionine sulfoxides (MsrA and MsrB) were detected, as were all known thiol peroxidase families. The data are consistent with the idea that the presence of Sec in a protein family, even if represented by only one or a few selenoprotein sequences, can indicate a redox function for the entire protein family and the location of redox-active Cys in proteins in this family. Several detected proteins had multiple conserved Cys. For example, thioredoxin, glutaredoxin, and AhpD contained conserved Cys in the form of CxxC motifs, whereas peroxiredoxin, MsrA, and MsrB had conserved Cys that were separated by variable distances. However, only those Cys that carried out catalytic functions (e.g., attacking nucleophiles) were identified by the Cys-Sec pairs, whereas Cys that served supporting functions (e.g., resolving Cys) were not detected. We further discuss several protein families predicted to use catalytic redox-active Cys.

A Cys-Sec pair represented by five Sec-containing sequences and numerous Cys-containing homologs revealed a redox-active Cys in *S*-adenosylmethionine (AdoMet)–dependent methyltransferases (Fig. 2A). However, only some members of this superfamily had Cys in this position. We constructed a phylogenetic tree and found that the methyltransferases containing the predicted redox-active Cys clustered and were present in both prokaryotes and eukaryotes. Thus, a protein family within a superfamily of AdoMet-dependent methyltransferases contained a catalytic redox-active Cys. Analysis of gene neighborhoods of bacterial methyltransferase genes revealed a functional link to arsenic detoxification.

In the detected methyltransferases, the predicted redox-active Cys was located in a C-terminal portion of the protein downstream of the common methyltransferase domain (Fig. 2B). In addition, these enzymes had a second conserved Cys residue. The common mammalian homolog of the detected methyltransferases is known as arsenic (+3 oxidation state) methyltransferase (AS3MT) (*13*, *14*). The production of methylated arsenicals in reactions catalyzed by AS3MT requires a reductant (*15*). Earlier studies with recombinant rat AS3MT found that the replacement of Cys^{156} with Ser led to a loss of catalytic activity (*16*). However, the Cys-Sec pair predicted that a different Cys donates the reducing equivalents to arsenic during the methylation reaction. We cloned the mouse AS3MT homolog and prepared $Cys^{157} \rightarrow$ Ser (corresponding to rat Cys^{156}) and $Cys^{207} \rightarrow$ Ser mutants. Both proteins were completely inactive, whereas the wild-type form efficiently converted inorganic arsenicals to monomethylated, dimethylated, and trimethylated forms (Fig. 2C). These data verified the function of the predicted redox-active Cys in arsenic methylation. We modeled the structure of mouse AS3MT (Fig. 2D). In the model, the active-site Cys was surface-exposed and protein topology supported the formation of an intramolecular disulfide during the catalytic cycle.

Predicted redox-active Cys were also found in other protein families, such as HesB-like (fig. S8), whose homologs participate in the biosynthesis of iron-sulfur proteins (*17*); DsrE (fig. S9), a small soluble protein first identified as part of the bacterial dsrABEFHCMK gene cluster (*18*); a subfamily of glutathione-*S*-transferases (*19*) (fig. S10); four families within the superfamily of rhodanese-like sulfurtransferases (*20*) (fig. S11); MoeB (fig. S12), which is thought to regenerate a thiocarboxylate group at the C terminus of MoaD in the molybdopterin synthase complex (*21*); heterodisulfide reductases (*22*) (fig. S13); and other proteins (figs. S14 to S19).

We examined the sequences that flank Cys-Sec pairs and found that a second Cys is often present in the vicinity of the redox-active Cys (fig. S20). In particular, a CxxC motif was abundant in the data set. Either the first or second Cys in this motif could serve as a redox-active residue. In addition, an increased frequency of glycine residues was found both upstream and downstream of the redox-active Cys. By contrast, negatively charged residues were extremely rare around the redox-active Cys and were absent in positions –3, +1, and +2 relative to the Cys. Analysis of secondary structures, either for the entire set of proteins or for non–thioredoxin-fold proteins (because thioredoxin-fold was the dominant fold in the data set), revealed a high frequency of β strands upstream and α helices downstream of the redox-active Cys. The Cys itself was most often present in loops (fig. S21). This distribution held for both thioredoxin-fold and non–thioredoxin-fold proteins.

It is remarkable that the searches for Cys-Sec pairs in homologous sequences could so selectively identify proteins with catalytic redox-active residues and filter out not only nonfunctional Cys but also conserved Cys that do not serve a catalytic redox function. Although this procedure selected only 0.5% of the initial sequences, it identified the majority of known families that use catalytic redox-active Cys. Our search strategy was not limited to organisms that contained selenoproteins. Moreover, it was sufficient to identify a single selenoprotein sequence (even if it was present as a hypothetical sequence from an unknown source, such as a metagenomics project) to predict a redox function for the entire family of proteins homologous to this selenoprotein. Arsenic methyltransferase provides a good illustration of this idea. Its Sec-containing sequences were detected in environmental soil sequences from a Minnesota farm (*23*), yet predictions of occurrence and location of redox-active Cys could be made for the entire protein family, which occurs in organisms from bacteria to mammals.

Our view on the evolution of Cys-Sec pairs is illustrated in Fig. 1B. In this model, most selenoproteins evolve from their redox-active Cys–containing homologs. To be selected during evolution, Sec must provide substantial advantages, as its use also results in disadvantages (slow Sec insertion, dependence on selenium, sensitivity to oxidation reactivity, and side reactions of Sec) (*10*). Thus, the Cys-to-Sec changes occur only in situations in which the catalytic properties of Sec are maximized, as is the case in redox proteins.

The major advantages of our method are its simplicity and strong predictive power. Redox-active Cys appears to be unique in that it could be so selectively identified, whereas other Cys (such as those involved in nonredox catalysis, structural disulfides, posttranslational modifications, and metal binding) could be filtered out. The set of predicted redox proteins described here should form the basis for further experimental verification. Increased availability of sequences, particularly from completed genomes and environmental projects, should further increase the predictive power of our method and lead to the identification of additional redox-active Cys.

References and Notes

1. M. Beeby *et al.*, *PLoS Biol.* **3**, e309 (2005).
2. D. E. Fomenko, V. N. Gladyshev, *Biochemistry* **42**, 11214 (2003).
3. J. S. Fetrow *et al.*, *Protein Sci.* **10**, 1005 (2001).
4. L. I. Leichert, U. Jakob, *PLoS Biol.* **2**, e333 (2004).
5. G. V. Kryukov *et al.*, *Science* **300**, 1439 (2003).
6. D. L. Hatfield, M. J. Berry, V. N. Gladyshev, *Selenium: Its Molecular Biology and Role in Human Health* (Springer, New York, 2006).
7. R. M. Tujebajeva *et al.*, *EMBO Rep.* **1**, 158 (2000).
8. P. R. Copeland, V. A. Stepanik, D. M. Driscoll, *Mol. Cell. Biol.* **21**, 1491 (2001).
9. S. V. Novoselov *et al.*, *EMBO J.* **21**, 3681 (2002).
10. H. Y. Kim, V. N. Gladyshev, *PLoS Biol.* **3**, e375 (2005).
11. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
12. Single-letter abbreviations for amino acid residues: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr.
13. B. P. Rosen, *FEBS Lett.* **529**, 86 (2002).
14. B. M. Adair *et al.*, *Environ. Chem.* **2**, 161 (2005).
15. S. B. Waters, M. Styblo, D. J. Thomas, *Chem. Res. Toxicol.* **17**, 404 (2004).
16. J. Li *et al.*, *Toxicol. Appl. Pharmacol.* **204**, 164 (2005).
17. J. R. Cupp-Vickery, J. J. Silberg, D. T. Ta, L. E. Vickery, *J. Mol. Biol.* **338**, 127 (2004).
18. C. Dahl *et al.*, *J. Bacteriol.* **187**, 1392 (2005).
19. J. D. Hayes, J. U. Flanagan, I. R. Jowsey, *Annu. Rev. Pharmacol. Toxicol.* **45**, 51 (2005).
20. D. Bordo, P. Bork, *EMBO Rep.* **3**, 741 (2002).
21. S. Leimkuhler, M. M. Wuebbens, K. V. Rajagopalan, *J. Biol. Chem.* **276**, 34695 (2001).
22. S. Madadi-Kahkesh *et al.*, *Eur. J. Biochem.* **268**, 2566 (2001).
23. S. G. Tringe *et al.*, *Science* **308**, 554 (2005).

24. Supported by NIH grants GM061603 and AG021518 (V.N.G.) and by U.S. Environmental Protection Agency–University of North Carolina Toxicology Research Program training grant T901915 (W.X.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/315/5810/387/DC1
Materials and Methods
SOM Text
Figs. S1 to S21
Table S1
References

27 July 2006; accepted 4 December 2006
10.1126/science.1133114