Relationship Between the Occurrence of Cysteine in Proteins and the Complexity of Organisms

Attila Miseta and Peter Csutora

Department of Clinical Chemistry, Faculty of Medicine, Pécs University, Pécs, Hungary

The occurrence and relative positions of cysteine residues were investigated in proteins of various species. Considering random mathematical occurrence for an amino acid coded by two codons (3.28%), cysteine is underrepresented in all organisms investigated. Representation of cysteine appears to correlate positively with the complexity of the organism, ranging between 2.26% in mammals and 0.5% in some members of the *Archeabacteria* order. This observation, together with the results obtained from comparison of cysteine content of various ribosomal proteins, indicates that evolution takes advantage of increased use of cysteine residues. In all organisms studied except plants, two cysteines are frequently found two amino acid residues apart (C-$(X)_2$-C motif). Such a motif is known to be present in a variety of metal-binding proteins and oxidoreductases. Remarkably, more than 21% of all of cysteines were found within the C-$(X)_2$-C motifs in *Archea.* This observation may indicate that cysteine appeared in ancient metal-binding proteins first and was introduced into other proteins later.

Introduction

Cysteine is unique among coded amino acids because it contains a reactive sulph-hydryl group. Therefore, two cysteine residues may form a cystine (disulfide link) between various parts of the same protein or between two separate polypeptide chains. The formation of a disulfide link is coupled with the folding of a protein and is assisted by the enzymes thiol-disulfide oxidoreductase and protein disulfide isomerase (Loferer and Hennecke 1994; Noiva 1994; Raina and Missiakas 1997). It is known that cytosolic proteins contain relatively few disulfide links, and the glutaredoxin and thioredoxin systems are involved in keeping the -SH groups in reduced form (Derman et al. 1993; Stewart, Aslund, and Beckwith 1998; Zhong et al. 1998).

Simple structural constraints may also interfere with the interaction of cysteines. Because of the rigid planar nature of the peptide bond, two neighboring cysteines are unable to interact. The same is true if two cysteines are spaced by only one amino acid. Based on the physical dimensions and the alignments of amino acid residues in alpha helix and beta turn configurations, two cysteines are closest to each other when spaced by two other amino acids (C-$(X)_2$-C motif; X may stand for any coded amino acid). Indeed, it is known that C-$(X)_2$-C motifs are well conserved parts of various iron-sulfur proteins and oxidoreductases (Shuber et al. 1986; Ammendola et al. 1992). However, we recognized that the occurrence of C-$(X)_2$-C domains was not limited to these groups of proteins. In addition, the number of C-$(X)_2$-C domains appeared amazingly high in a large randomly selected population of known *Saccharomyces cerevisae* proteins.

These observations prompted us to analyze if cysteines occurred randomly or organized preferentially into C-$(X)_2$-C motifs in known proteins of various species. Before carrying out such an analysis, the occurrence of cysteines and other coded amino acids in proteins of various species had to be analyzed. Cysteine is coded by the codon triplets UGU and UGC, respectively. All known organisms contain cysteines in their proteins.

Since all known living organisms contain the same 20 coded amino acids, there is no direct way to tell if one or another amino acid was introduced sooner or later during evolution. Alternatively, life on earth might have started with primitive organisms containing the known 20 coded amino acids. Evidence supporting the gradual expansion of the genetic code is primarily based on the anomalous codon allocations of various species (Osawa et al. 1992; Bauman and Oro 1993).

In addition, Doring and Marliere (1998) showed that mutant tRNAs that incorporate cysteine at positions corresponding to the isoleucine or methionine codons may be maintained in *Escherichia coli.* The toxicity of cysteine miscoding was low, proving that cysteine is an acceptable substitution at most protein positions. Thus, the incorporation of cysteine may overcome the steric and polar restrictions that limit the evolution of the genetic code.

Therefore, it is possible that cysteine was introduced in a relatively late stage of evolution through a takeover of some codons of another amino acid (serine, glycine). If the synthesis of an amino acid side chain progresses by the same or a similar biosynthetic pathway(s), the resulting amino acids are physicochemically similar (e.g., Ile, Val, Leu, Phe, Tyr, Asp, Glu) (Miseta 1989). Importantly, physicochemically related amino acids occupy similar codons (Jungck 1978; Weber and Lacey 1978). Therefore, the genetic code is designed to minimize the chance that a single nucleotide base change will result in replacement with a physicochemically different amino acid.

The result of nontargeted introduction of a cysteine must have been that many cysteines were not at ''optimal places.'' Subsequent evolution must have corrected for these ''errors,'' leaving cysteines in positions where their physicochemical properties could be used to the advantage of the organism. Simultaneously, mutations

Key words: cysteine, amino acids, evolution, ribosomal proteins, secondary structure.

Address for correspondence and reprints: Attila Miseta, Department of Clinical Chemistry, Faculty of Medicine, Pécs University, 7624 Pécs, Ifjúság u.13., Hungary. E-mail: miseta@clinics.pote.hu.

Mol. Biol. Evol. 17(8):1232–1239. 2000

resulting in cysteine-coding triplets facilitated the development of new, more complex, proteins.

Based on this hypothesis, one might expect to find simple organisms with relatively few cysteines and complex organisms with significantly more cysteines. Likewise, one might anticipate finding cysteines within a limited number of protein domains in simple organisms.

In the present report, we attempt to highlight the relationship between the complexity of various organisms and the occurrence of cysteine in their proteins.

Materials and Methods

Protein Sequences, Sequence Data Files, and the Analysis of Amino Acid Composition

The sequence analysis was carried out on sequence data files created from sequences stored in the SwissProt database (CDPROT31). The sequence data files were created with the aid of a selection rule. The selection rule specified the scientific name of species. When indicated, the subcellular locations of proteins were also specified. Then, the saved data files containing the catalogs for protein sequences were edited. This editing was restricted to the removal of a few corrupted sequence files for which the subsequent data analysis programs could not run.

The following data files were created: *Homo sapiens* (human), 2,681 protein sequences; *Mus musculus* (mouse), 1,727; *Bos taurus* (bovine), 719; *Drosophila melanogaster* (fruit fly), 624; *Caernorhabditis elegans,* 690; *Zea mays* (corn), 287; *Orysa sativa* (rice), 219; *Lycopersicon esculentum* (tomato), 117; *Saccharomyces cerevisiae* (baker's yeast), 3,127; *E. coli,* 3,151; *Rhodobacter sphaeroides,* 58; *Pseudomonas aeruginosa,* 214; *Haloarcula marismortui (*formerly *Halobacterium marismortui*), 62; and *Thermus aquaticus,* 88. In some cases, proteins from related species were used for the analysis. The data files for various *Cyanobacteria* contained 556 sequences, those for *Archea* contained 553 sequences, and those for *Sulpho-archea* contained 137 sequences. Data files for ribosomal proteins included 66 sequences for humans, 134 sequences for yeast, 59 sequences for *E. coli,* and 51 sequences for *H. marismortui.*

Since data were generated and intended for comparison among data files containing different numbers of protein sequences, it was important to calculate the standard deviations of any results for smaller and larger data files. Therefore, we generated five data files each containing 50 *S. cerevisiae* sequences with a random generator. The occurrence of the 20 coded amino acids was analyzed in the five data files, and mean values and standard deviations were calculated. We found that in this case the average SD value for the 20 amino acids was 5.57%. For example, we found 291.2 ± 36.8 cysteine residues (mean ± SD). When each of the five *S. cerevisiae* databases contained 100 or 200 randomly selected protein sequences instead of 50, the average SD values decreased to 3.50% and 2.21%, respectively.

Doubtless, smaller data files do not represent the average protein composition of a species as well as larger ones. We also note that these data files were not created by a random generator, but rather by the specific interests of researchers. For this reason, we intended to be as careful as possible in interpreting data in small data files. No data file containing fewer than 50 sequences was used. Also, the creation of combined data files of related species with higher numbers of sequences was aimed at decreasing the standard error.

The amino acid composition of proteins within a data file was analyzed, with the combined total of the 20 coded amino acids being 100%. The cysteine (and any other amino acid) content was expressed as a percentage of total amino acids.

Correction of Amino Acid Composition Data for the GC Contents Versus the AT Contents of Coding DNA

It is known that the GC contents versus the AT contents of coding sequences vary within a wide range in different species. Cysteine may be coded by UGU or UGC. Consequently, only first– and second–codon position GC contents affect the occurrence of cysteine. The total coding GC contents and the averaged first plus second codon letter GC contents (values in italics) of the species listed in the report are as follows: humans—52.75%, *49.24%*; bovines—52.86%, *48.37%*; mice—52.76%, *49.25%*; *D. melanogaster*—54.53%, *48.95%*; *C. elegans*—42.37%, *43.94%*; maize—55.69%, *50.67%*; rice—56.16%, *51.03%*; tomatoes—42.74%, *45.26%*; *S. cerevisae*—39.72%, *40.60%*; Cyanobacteria—32.36%, *41.8%*; *E. coli*—51.37%, *49.60%*; *R. sphaeroides*—67.86%, *57.69%*; *P. aureginosa*—64.06%, *55.09%*; *H. marismortui*—63.69%, *55.00%*; *T. aquaticus*—66.57%, *55.45%*; *Sulpho-archea*—43.12%, *45.71%.*

The increase in the first–codon position GC content is expected to have a negative effect on the cysteine content of the species, because the first codon letter is U. The increase of the second codon letter GC content should have a positive effect on the cysteine content, because the second codon letter is G. In addition, (A + U) + (G + C) = 1 (where 1 is the total nucleotide content of the species). Since (G + C) = 1 − (A + U) the relationship between the probability of coding for cysteine (Pcys) and the (G + C) content of the coding DNA may be described as Pcys = (G + C)*(1 − (G + C)). Because the probability curve is quite flat in the 40%–60% GC content range, corrected values never differ from raw data by more than 3.5%. Whether or not data were corrected for the actual GC content of the species is indicated in the text and/or figure legends.

Analysis of Amino Acid Spacing and Protein Homology

The above-mentioned data files were used to analyze the occurrence of various amino acid sequences. The occurrence of C-$(X)_n$-C sequence was analyzed. C is the single-letter label for cysteine, X is any other amino acid, and *n* is the number of repeats, 0–22 in our investigations. Since the probe sequence of C-$(X)_n$-C may be overlaid on any sequence of given length once more than the C-$(X)_{n+1}$-C probe sequence, appropriate

correction may be carried out. However, we found that the resulting correction did not significantly alter our results, and it was omitted for this reason.

For the C-$(X)_2$-C sequence, we also analyzed the occurrence of all individual sequences in the data files of humans, fruit flies, *E. coli,* and *Archea.* Amino acid sequence alignments were done according to the method developed by Myers and Miller (1988). We used a structure genetic matrix based on the work of Feng, Johnson, and Doolittle (1984) with an open gap cost of 10 and a unit gap cost of 5.

Results

The Occurrence of Cysteine in Proteins of Various Species

First, we investigated the occurrence of cysteine in known proteins of various species. In order to do this, the occurrences of the 20 coded amino acids were counted in representative samples (data files) of humans, bovines, mice, fruit flies, *C. elegans,* maize, rice, tomatoes, yeast, *Cyanobacteria, E. coli, R. sphaeroides, P. aeruginosa, H. marismortui, T. aquaticus,* and *Sulpho-archea.* The total of all 20 coded amino acids was considered 100%.

Cysteine is a relatively rare amino acid within the proteins of investigated organisms (fig. 1). Analyses of human, bovine, and mouse proteins revealed almost identical (2.26%) occurrences of cysteine. Proteins of the fruit fly and *C. elegans* contained somewhat less cysteine (1.90% and 1.97%, respectively). Proteins of maize, rice, and tomatoes contain 1.62%–1.69% cysteine. Among the eukaryotes studied, yeast contained the least cysteine (1.21%).

Among the prokaryotes studied, *Cyanobacteria, E. coli, P. aeruginosa,* and *R. sphaeroides* contained 1.03%–1.13% cysteine, which is somewhat less than that seen in *S. cerevisiae.*

Representatives of the newly established *Archea* order were also studied (Woese and Gupta 1981; Woese, Kandler, and Wheelis 1990). We found that the cysteine contents of the extreme halophil *H. marismortui* and the thermophil *T. aquaticus* were the lowest among the species investigated (0.49% and 0.41%, respectively). Similarly, low representation of cysteine is apparent in *Halobacterium salinarium* (data not shown) and members of the *Sulpho-archea* group. All data were corrected for the actual GC contents of the species (used as background probability) as described in *Materials and Methods.*

We note that the actual and GC-corrected cysteine levels were nearly identical for most species. (The largest difference between these values was in *S. cerevisiae,* for which the actual cysteine content was 1.25%, whereas the GC corrected was 1.21%.)

While the proteins of the *Archea* order may contain four- to fivefold less cysteine than mammalian proteins, it appears that cysteine is underrepresented in all organisms investigated. Considering that there are 61 sense codons in these organisms, and two codons are allocated to cysteine, $(100/61) \times 2 = 3.28\%$. The GC-corrected probability range is 3.23%–3.28% for the species studied.

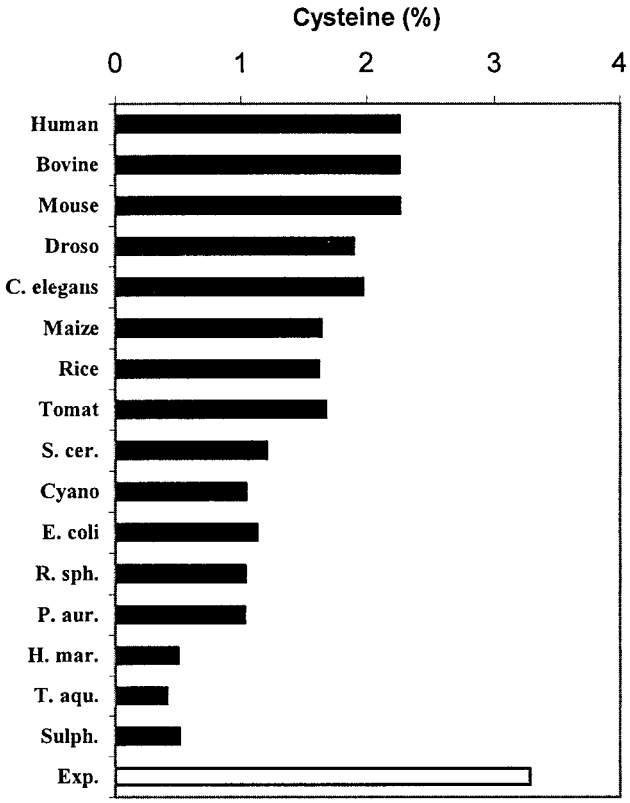

FIG. 1.—The occurrence of cysteine in proteins of various species. The cysteine content is expressed as a percentage (Y-axis). The total of all 20 coded amino acids was considered 100%. All data were corrected for the GC contents versus the AT contents of the species as described in *Materials and Methods.* Since cysteine has 2 codons out of 61 sense codons, the mathematical probability for the occurrence of cysteines is 3.28% (last column). The GC corrected data range is 3.23%–3.28% for various species.

We also calculated the number of proteins with and without at least one cysteine residue in some species. Figure 2 shows that about 92% of human proteins contain one or more cysteine residues. Similar numbers were obtained for fruit fly and yeast proteins. A modest increase in the number of cysteine-free proteins is apparent for *E. coli,* but more than 50% of the known *H. marismortui* proteins do not contain a single cysteine residue. Similar percentages may be obtained for other related species listed in figure 1.

We note that the average lengths of proteins are not equal in different data files (species) (see *Materials and Methods* for details).

The Occurrence of Cysteine in the Ribosomal Proteins of Various Species

It is known that one or more disulfide links are frequently found in excreted or plasma membrane proteins. In contrast, cytosolic proteins often lack disulfide links. It is known that cytosolic cysteines are maintained in the thiol form by the glutaredoxin and thioredoxin systems, and the scarcity of disulfide bridges in cyto-

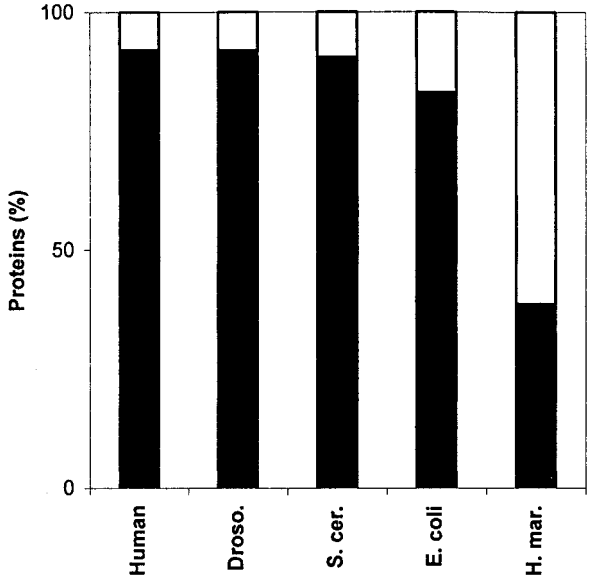

FIG. 2.—Proteins with and without at least one cysteine residue in various species. The black part of the stacked column represents proteins with cysteine, while the white part represents proteins without a single cysteine residue. The total of all proteins analyzed was considered 100% (Y-axis).

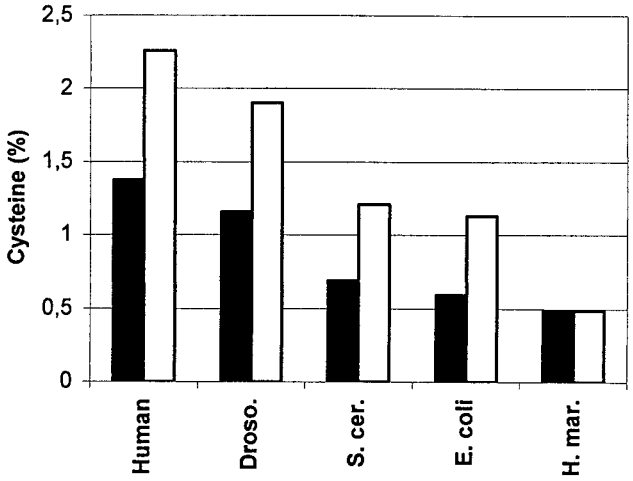

FIG. 3.—The cysteine contents of ribosomal proteins (black columns) of various species. The general cysteine contents of proteins (empty columns) of various species are also shown. Cysteine content is expressed as a percentage of the 20 coded amino acids (Y- axis). The data are GC-corrected.

solic proteins may or may not translate to lower protein cysteine contents for this reason. However, one hypothetical reason behind the elevated quantities of cysteine in proteins of eukaryotes, particularly in mammals, might be that complex organisms may contain a wider variety of cysteine-rich extracellular proteins. Therefore, we studied if known ribosomal proteins of mammals contain more cysteine than those of less developed organisms. It appears that ribosomal proteins contain fewer cysteine residues in humans, fruit flies, yeast, and *E. coli* when compared with the average cysteine contents of their proteins (fig. 3). This is not the case for *H. maresmortui.* However, the general trend, i.e., less developed organisms containing fewer cysteine residues, remains valid.

We also selected 10 human ribosomal proteins for which the related sequences may be retrieved for yeast, *E. coli,* and *H. marismortui* (table 1). We found that the 10 human ribosomal proteins combined contained 2,404 amino acids and 30 cysteines among them. Thus, their cysteine content was 1.25%, significantly less than the average cysteine content of human proteins (2.26%). However, the average cysteine content of their sibling ribosomal proteins in yeast was 0.42%, followed by *E. coli* at 0.19% and *H. maresmortui* at 0.11%. Data were not affected significantly by correction for the GC contents of various species. The ribosomal proteins listed in table 1 showed 60%–78% identity/similarity between humans and yeast (RL6 is an exception; it shows insignificant identity/similarity between humans and yeast). Despite this, two out of three cysteines were new introductions in human ribosomal proteins.

Cysteines in human proteins were found in place of alanine ($n = 8$), threonine ($n = 4$), serine ($n = 4$), valine ($n = 4$), isoleucine ($n = 1$), glycine ($n = 1$,) and glutamine ($n = 1$) in yeast. Frequently, these replacements were found within well-conserved regions of ribosomal proteins.

In conclusion, ribosomal proteins are relatively poor in cysteine, but the overall relationship among their

Table 1
Cysteine in Ribosomal Proteins of Various Species

	HUMAN		*SACCHAROMYCES CEREVISIAE*		*ESCHERICHIA COLI*		*HALOARCULA MARISMORTUI*	
SEQUENCE	Length	Cys	Length	Cys	Length	Cys	Length	Cys
RS3	243	3	239	1	232	0	304	0
RS4	221	5	253	1	205	1	175	0
RS6	249	3	236	1	135	0	116	0
RS8	207	4	199	2	129	1	129	1
RL3	402	4	386	3	209	0	337	0
RL6	286	2	253	0	176	1	177	0
RL13	210	1	199	0	142	0	145	1
RL15	203	2	203	1	144	0	154	0
RL18	187	2	186	1	117	0	186	0
RL19	196	2	188	0	114	0	148	0
Total.	2,404	28	2,342	10	1,603	3	1,871	2

Table 2
Occurrence of Phe, Tyr, Ile, and Cys Amino Acid Residues in the Proteins of Various Species (%)

	Phe	Tyr	Ile	Cys
Human	3.8	3.6	4.6	2.3
Drosophila	3.4	2.9	5.0	1.9
Saccharomyces cerevisiae	4.5	3.3	6.5	1.3
Escherichia coli	3.9	2.9	6.0	1.1
Haloarcula maresmortui	2.6	2.0	4.4	0.5
Thermus aquaticus	3.6	3.1	3.8	0.4

cysteine contents in various species (fig. 3) is similar to the general trend displayed in figure 1. From these data, it appears that an evolutionary trend favors the incorporation of more cysteine residues into proteins of more complex organisms.

GC Contents Versus AU Contents of Coding DNAs and Their Relation to the Occurrence of Cysteines

The simplest explanation for the low cysteine contents of proteins in members of *Archea* would be that the GC content of the DNA correlates with the occurrence of cysteine in proteins of various species. Since many members of the *Archea* have high-GC DNA, the occurrence of amino acids which are encoded by UA-rich triplets is not preferred. However, cysteine may be coded by the triplets UGU and UGC, respectively. One may find a strong bias toward the use of C rather than U in the third codon position for encoding cysteine in members of *Archea* and other GC-rich species (data not shown). Since the two important codon letters for cysteine are U and G, the GC contents of the first two codon positions were used to correct our data as described in *Materials and Methods.*

Furthermore, we studied whether the occurrences of other amino acids encoded by UA- or GC-rich triplets follow trends similar to that of cysteine in various species. Table 2 shows that neither phenylalanine (codons UUU and UUC) nor tyrosine (UAU and UAC) nor isoleucine (AUU, AUC, and AUA) follow trends similar to that of cysteine among the species studied. The same holds true for other amino acids (data not shown).

The Nonrandom Occurrence of Cysteine Residues in Proteins of Various Species

Next, we investigated some aspects of the spatial distribution of cysteines in proteins of various species. To answer the question of whether there are some preferential positions of two cysteines relative to each other, we carried out systemic searches in representative protein data files of various species. As a probe, a C-$(X)_n$-C search string was applied. Here, X stands for any amino acid, whereas n is an integer between 0 and 22. Because the data files for different species were of very different sizes, the numbers representing the occurrence of C-$(X)_n$-C sequences in the $n = 0$–22 range were averaged and considered 1 for each species.

Figure 4 shows the occurrence of C-$(X)_n$-C sequences in various species in the $n = 0$–8 range. We found that there was a preferred occurrence for the C-$(X)_2$-C sequence in every species investigated except for plants. Thus, human sequences (other mammals show similar patterns) contain more than 1.5 times as many C-$(X)_2$-C motifs as C-X-C or C-$(X)_3$-C motifs. Essentially the same pattern was observed in *D. melanogaster* and *C. elegans.* Approximately every fifth protein contained one or more C-$(X)_2$-C sequences in these species (555 out of 2,681 human proteins and 131 out of 624 fruit fly proteins). Interestingly, plant proteins did not contain outstanding numbers of C-$(X)_2$-C sequences, as shown in the example of maize. In order to elucidate whether all plants are like maize, we also studied wheat, rice, and tomato proteins. We found similar results (data not shown), confirming that plant proteins do not have high representations of C-$(X)_2$-C sequences. However, proteins of the chloroplasts of various plants show C-$(X)_2$-C patterns similar to those of bacteria (data not shown).

The free-living organism *S. cerevisiae,* which has low cysteine content compared with humans (fig. 1), has more than 5% of its total cysteine content arranged in C-$(X)_2$-C motifs.

The relative number of C-$(X)_2$-C motifs was further increased in *E. coli* and in other eubacteria, but the most spectacular increase was in the members of the *Archea.* Out of 1,063 cysteines found in 553 known proteins, 232 cysteine residues were found within C-$(X)_2$-C motifs. Thus, over 21% of the cysteines are arranged in this particular pattern.

In order to elucidate whether any other amino acid is arranged in a pattern similar to that of cysteine, we investigated the other 19 amino acids within the *S. cerevisiae* and *E. coli* data files. Our results indicated that no other amino acid showed a distribution similar to that of cysteine in proteins of *S. cerevisiae* or *E. coli* (data not shown).

Next, we analyzed various C-$(X)_2$-C motifs in humans, yeast, *E. coli,* and members of *Archea.* As was anticipated, some C-$(X)_2$-C motifs were more abundant than others. In humans, there are many glutamic acid (and other charged amino acid) residues in the place of X. Zinc finger proteins, metallothionens, and dehydrogenases may contain several of these motifs. However, 555 out of 2,681 investigated human proteins contain at least one C-$(X)_2$-C motif, and it is beyond the scope of the present report to detail and discuss their possible interrelatedness and similarities to their siblings. Suffice it here to claim that any possible combination of amino acids was found at least once. Also, the most abundant C-$(X)_2$-C motifs in data files of humans, yeast, *E. coli,* and *Archea* are different, albeit some characteristic compositional features are definitely present. One such case is that of the noted preferred occurrence of polar charged amino acids in mammals. Yeast, *E. coli,* and *Archea* also seem to contain a number C-P-X-C or C-X-P-C residues.

Discussion

Cysteine is the only coded amino acid which carries a reactive sulfhydryl group. By forming intra- and

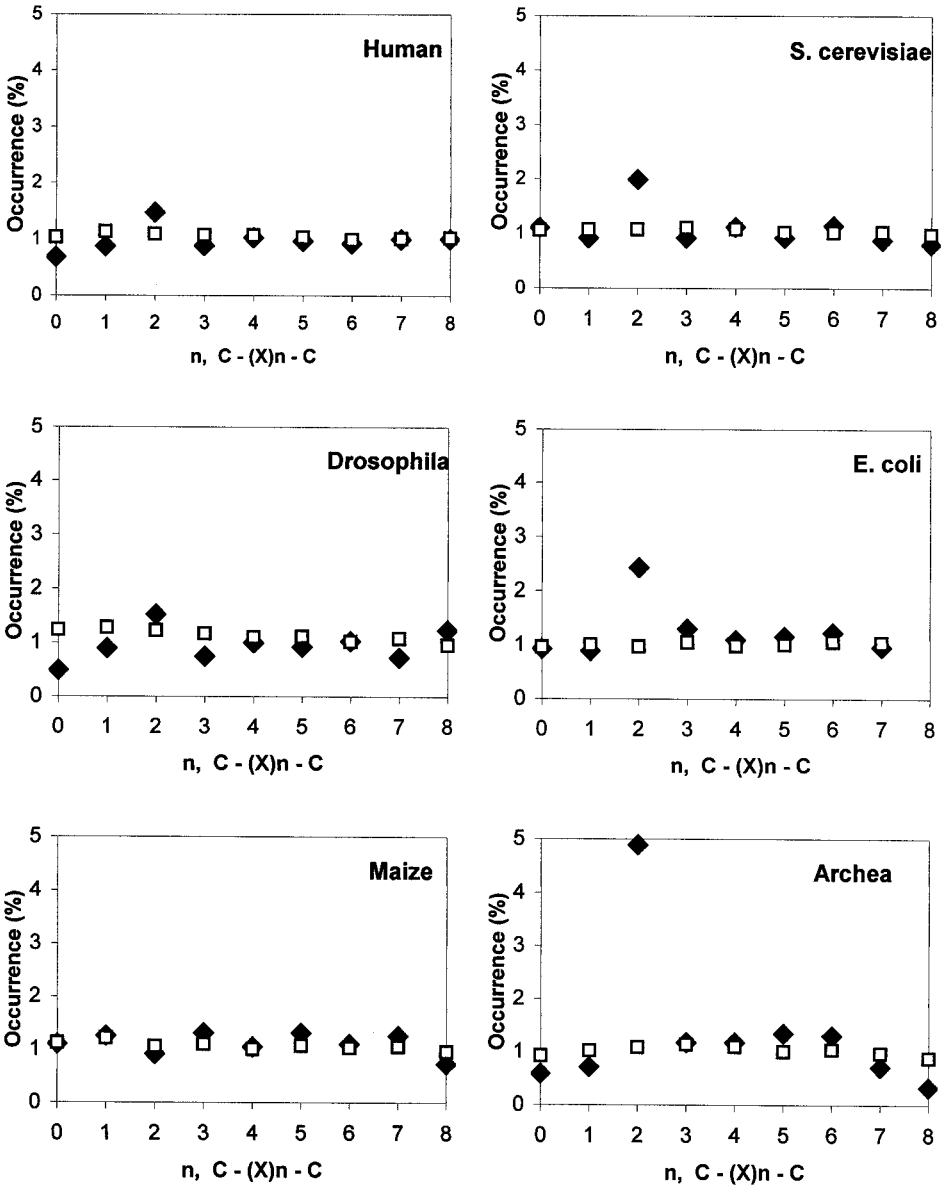

FIG. 4.—The relative positions of cysteines (filled squares) and glycines (empty squares) within a stretch of amino acids ≤8 ($n \leq 8$). The occurrences of cysteine and glycine residues were analyzed for $n = 0$–22, and the average occurrence was considered 1. Total occurrence values were 24,484 for humans, 7,475 for *Drosophila*, 877 for maize, 9,602 for *S. cerevisae*, 4,515 for *E. coli*, and 545 for *H. maresmortui*. The actual occurrence is displayed on the Y-axis; *n* is displayed on the X-axis. The data are GC-corrected.

interchain disulfide links, it plays special role in protein structure (Noiva 1994; Raina and Missiakas 1997). Apparently, it is similarly important that most cysteines of cytosolic proteins are maintained in the reduced thiol form (Derman et al. 1993).

In the present work, we studied the occurrence of cysteine in various species. We found that the representation of cysteine in various organisms has one common feature: cysteine is underrepresented in proteins of all species investigated (fig. 1). However, the cysteine contents of proteins of different species may be as low as 0.4%–0.5% in *Archea*, whereas proteins of mammals characteristically contain about 2.26% cysteine residues.

The higher cysteine contents in proteins of more complex organisms may be explained in part by the higher numbers of disulfide link–rich noncytosolic proteins. Naturally, the relative absence of disulfide links indicates but does not prove that cytosolic proteins are relatively poor in cysteine. Nevertheless, our results indicate that ribosomal proteins are poor in cysteine when compared with the average cysteine contents of proteins found in identical species (fig. 3).

However, the increase in the cysteine contents of proteins of complex organisms species is valid for ribosomal proteins too. We demonstrated that related ribosomal proteins tend to include more cysteines in the more developed species (fig. 3 and table 1). For example, despite the extensive similarity between human and yeast ribosomal proteins, the former contains almost three times more cysteine residues. Therefore, one can-

not fail to note that there is a relationship between increased complexity and increased representation of cysteine (fig. 1).

Importantly, this trend is unique to cysteine, as other amino acids do not follow the same trend. We explain the above-described phenomena by assuming that evolution takes advantage of the increased use of cysteine. The unique protein structure–shaping capability of cysteine is probably used more frequently in developed organisms.

The increased use of cysteine in developed organisms may be investigated in light of theories about codon allocation. In brief, it is possible that even primitive ancient organisms used a code table very similar to that used in modern organisms. Alternatively, the ancient code table may have coded for fewer than 20 amino acids, and the present code allocations may be the result of progressive code divisions (Wong 1975, 1976, 1988). During this hypothetical process, the codons allocated to a given amino acid were shared with other amino acids. Arguments for and against these theories are beyond the scope of the present report. On one hand, a problem with the ''code-sharing theory'' is that the new allocation of a codon—even if the replacing amino acid is a physicochemically related sibling—will cause widespread structure and function changes.

On the other hand, the fact that some organisms contain ''anomalous'' code allocations supports the code evolution and code-sharing hypotheses (Osawa et al. 1990, 1992; Jukes and Osawa 1993). Also, Doring and Marliere (1998) showed that mutant tRNAs that incorporate cysteine at positions corresponding to the isoleucine or methionine may be maintained in *E. coli.* The toxicity of cysteine miscoding was low, proving that cysteine is an acceptable substitution at most protein positions.

Our results show that the introduction of cysteine was gradual during evolution (fig. 1), and progressive mutations resulted in a shift toward the increased use of cysteine. Consequently, our observation provides no direct support for codon sharing, but it points toward the possibility that cysteine might have been a ''late arrival'' compared with other amino acids.

We also found a peculiar nonrandomness in the spacing of cysteines; namely, two cysteines spaced by two other amino acids occur with exceptionally high frequency. The C-$(X)_2$-C domain is overrepresented in the proteins of a number of species. This overrepresentation is more and more pronounced among primitive organisms, with the result that over 21% of the total cysteines are within C-$(X)_2$-C motifs in members of the *Archea* order.

Our explanation for this phenomenon is that the first and fourth amino acid residues in the alpha helix or in the beta turn conformations are closest to each other, respectively. Amino acids that are most abundant in β turns are found to be assigned to a well-defined part of the genetic code. Since this group of amino acids contains biosynthetic precursors to other amino acids, it is possible that these amino acids and β turns were more abundant during early evolution (Jurka and Smith 1987).

C-$(X)_2$-C motifs are important domains of metal-binding proteins and oxidoreductases (Shuber et al. 1986; Ammendola et al. 1992). Cammack, Hall, and Rao (1971) suggested that ferredoxins played an important role in early evolution.

The very high abundance of C-$(X)_2$-C motifs in primitive organisms is likely to be linked to the relative importance of short-range intrapolypeptide chain interactions, as opposed to long-range intrapolypeptide chain interactions and interpolypeptide chain interactions. While the former was more important in ancient organisms, the latter two types of interactions become more and more important later in evolution. This may explain the observation that the increase in the cysteine content resulted in a relative decrease in the number of cysteines present in C-$(X)_2$-C motifs.

Despite the relative decrease in the number of cysteines in C-$(X)_2$-C motifs, we found 555 such motifs in the 2,681 human proteins studied. Zinc finger or zinc finger–like transcription factors are used in a number of proteins which emerged in multicellular organisms. These include a variety of hormone receptors. Consequently, evolution resulted in many new uses for proteins with C-$(X)_2$-C domains, but even more proteins acquired cysteines in various other positions.

Acknowledgments

We thank Drs. David M. Bedwell, Miklós Kellermayer, István Simon, and Denys N. Wheatley for critcal reading of the manuscript. We thank András Rab for technical assisstance. This work was supported by a grant from the Hungarian Ministry of Health (T-04 035/99).

LITERATURE CITED

AMMENDOLA, S., C. A. RAIA, C. CARUSO, L. CAMARDELLA, S. D'AURIA, M. DE ROSA, and M. ROSSI. 1992. Thermostable NAD(+)-dependent alcohol dehydrogenase from Sulfolobus solfataricus: gene and protein sequence determination and relationship to other alcohol dehydrogenases. Biochemistry **31**:12514–12523.

BAUMANN, U., and J. ORO. 1993. Three stages in the evolution of the genetic code. Biosystems **29**:133–141.

CAMMACK, R., D. HALL, and K. RAO. 1971. Ferredoxins: are they living fossils? New Sci. **23**:696–698.

DERMAN, A. I., W. A. PRINZ, D. BELIN, and J. BECKWITH. 1993. Mutations that allow disulfide bond formation in the cytoplasm of Escherichia coli. Science **262**:1744–1747.

DORING, V., and P. MARLIERE. 1998. Thermostable NAD(+)-dependent alcohol dehydrogenase from *Sulfolobus solfataricus*: gene and protein sequence determination and relationship to other alcohol dehydrogenases. Genetics **150**:543–551.

FENG, D. F., M. S. JOHNSON, and R. F. DOOLITTLE. 1984. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. J. Mol. Evol. **21**:112–125.

JUKES, T. H., and S. OSAWA. 1993. Evolutionary changes in the genetic code. Comp. Biochem. Physiol. B Comp. Biochem. **106**:489–494.

JUNGCK, J. R. 1978. The genetic code as a periodic table. J. Mol. Evol. **2**:211–224.

JURKA, J., and T. F. SMITH. 1987. β turns in early evolution: chirality, genetic code, and biosynthetic pathways. Cold Spring Harbor Symposia on Quantitative Biology. Vol. LII. Cold Spring Harbor Symposium on Quantitative Biology, Cold Spring Harbor, NY.

LOFERER, H., and H. HENNECKE. 1994. Protein disulphide oxidoreductases in bacteria. Trends Biochem. Sci. **19**:169–171.

MISETA, A. 1989. The role of protein associated amino acid precursor molecules in the organization of genetic codons. Physiol. Chem. Phys. **21**:237–242.

MYERS, E. W., and W. MILLER. 1988. Optimal alignments in linear space. Comput. Appl. Biosci. **4**:11–17.

NOIVA, R. 1994. Enzymatic catalysis of disulfide formation. Protein Expr. Purif. **5**:1–13.

OSAWA, S., T. H. JUKES, K. WATANABE, and A. MUTO. 1992. Recent evidence for evolution of the genetic code. Microbiol. Rev. **56**:229–264.

OSAWA, S., A. MUTO, T. H. JUKES, and T. OHAMA. 1990. Evolutionary changes in the genetic code. Proc. R. Soc. Lond. B Biol. Sci. **241**:19–28.

RAINA, S., and D. MISSIAKAS. 1997. Making and breaking disulfide bonds. Annu. Rev. Microbiol. **51**:179–202.

SHUBER, A. P., E. C. ORR, M. A. RECNY, P. F. SCHENDEL, H. D. MAY, N. L. SCHAUER, and J. G. FERRY. 1986. Cloning, expression, and nucleotide sequence of the formate dehydrogenase genes from methanobacterium formicicum. J. Biol. Chem. **261**:12942–12947.

STEWART, E. J., F. ASLUND, and J. BECKWITH. 1998. Disulfide bond formation in the Escherichia coli cytoplasm: an in vivo role reversal for the thioredoxins. EMBO J. **17**:5543–5550.

WEBER, A. L., and J. C. LACEY. 1978. Genetic code correlations: amino acids and their anticodon nucleotides. J. Mol. Evol. **2**:199–210.

WOESE, C. R., and R. GUPTA. 1981. Are archaebacteria merely derived 'prokaryotes'? Nature **289**:95–96.

WOESE, C. R., O. KANDLER, and M. L. WHEELIS. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA **87**:4576–4579.

WONG, J. T. F. 1975. A co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. USA **72**:1909–1912.

———. 1976. The evolution of the universal genetic code. Proc. Natl. Acad. Sci. USA **73**:2336–2340.

———. 1988. Evolution of the genetic code. Microbiol. Sci. **5**:174–181.

ZHONG, L., E. S. ARN-ER, J. LJUNG, F. ASLUND, and A. HOLMGREN. 1998. Rat and calf thioredoxin reductase are homologous to glutathione reductase with a carboxyl-terminal elongation containing a conserved catalytically active penultimate selenocysteine residue. J. Biol. Chem. **273**:8581–8591.

WILLIAM TAYLOR, reviewing editor

Accepted April 16, 2000