Nucleic Acids and Molecular Biology

For further volumes:
http://www.springer.com/series/881

John F. Atkins · Raymond F. Gesteland
Editors

Recoding: Expansion of Decoding Rules Enriches Gene Expression

Springer

Editors
John F. Atkins
BioSciences Institute
University College Cork
Ireland
and
Department of Human Genetics
University of Utah
and Genetics Department
Trinity College Dublin, Ireland
atkins@genetics.utah.edu

Raymond F. Gesteland
Department of Human Genetics
University of Utah
15N. 2030E.
Salt Late City
UT 84112-5330
USA
ray.gesteland@genetics.utah.edu

ISBN 978-0-387-89381-5 e-ISBN 978-0-387-89382-2
DOI 10.1007/978-0-387-89382-2
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009938958

Foreword

The literature on recoding is scattered, so this superb book fills a need by providing up-to-date, comprehensive, authoritative reviews of the many kinds of recoding phenomena.

Between 1961 and 1966 my colleagues and I deciphered the genetic code in *Escherichia coli* and showed that the genetic code is the same in *E. coli*, *Xenopus laevis,* and guinea pig tissues. These results showed that the code has been conserved during evolution and strongly suggested that the code appeared very early during biological evolution, that all forms of life on earth descended from a common ancestor, and thus that all forms of life on this planet are related to one another. The problem of biological time was solved by encoding information in DNA and retrieving the information for each new generation, for it is easier to make a new organism than it is to repair an aging, malfunctioning one.

Subsequently, small modifications of the standard genetic code were found in certain organisms and in mitochondria. Mitochondrial DNA only encodes about 10–13 proteins, so some modifications of the genetic code are tolerated that probably would be lethal if applied to the thousands of kinds of proteins encoded by genomic DNA.

In 1986 the 21st amino acid, selenocysteine, which responds to the terminator codon, UGA, when a stem-loop structure in mRNA is downstream of the UGA codon and is recognized by a protein was discovered. In 2002 the 22nd amino acid, pyrrolysine, which responds to the terminator codon, UAG, was discovered. Pyrrolysine is found only in a few species of bacteria.

During the last 40 years a great deal of information has been obtained that shows that some mRNA molecules contain signals in addition to the 64 kinds of RNA codons that modify the translation of codons. These signals may involve intramolecular hydrogen bonding between nucleotides in mRNA such as the formation of hairpin-like stem-loop structures or pseudoknots, certain nucleotide sequences followed by mRNA secondary structure that delay codon translation, or hydrogen bonding between mRNA and ribosomal RNA of the translating ribosomes. These signals add considerable complexity to the translation of mRNA. For example, these signals can alter the reading frame of specific species of mRNA at specific sites within the partially translated mRNA. The signals can specify whether reading frame 1 should be changed to reading frame 2 or to reading frame 3 at specific

codons during the translation of the mRNAs. The reading frame can be altered by skipping one nucleotide in the 3' direction, or by going back one nucleotide or two nucleotides in the 5' direction. There is also a mechanism that enables the ribosome to skip 50 bases. Another mechanism evolved that allows ribosomes to translate a specific species of mRNA to a certain point and then continue translation of different molecules of RNA.

Some remarkable and quite beautiful recoding mechanisms have been discovered that function as regulators of gene expression. For example, *E. coli* release factor 2 (RF2) mRNA contains near the beginning of the mRNA a slippery nucleotide sequence before the terminator codon, UGA, followed by a pseudoknot in the mRNA. When the concentration of RF2 protein is high, RF2 protein recognizes the UGA codon and terminates, i.e., aborts the synthesis of RF2 protein. However, when the concentration of RF2 protein is low, one base is skipped in the 3' direction resulting in a shift to reading frame 2 thus enabling the synthesis of full-length RF2 protein. Thus, a frameshift in mRNA translation is used to regulate the translation of RF2 mRNA. Programmed frameshifts are required for the translation of many species of viral RNA, including HIV. Programmed frameshifts also are involved in the translation of some species of mRNA derived from genomic DNA.

Many human genetic diseases have been found that result from mutations that convert a codon for an amino acid to a terminator codon that prematurely terminates the synthesis of the protein. One approach that has been explored is to treat these patients with small molecules such as the aminoglycoside, gentamicin, or other molecules that result in some misreading of codons. This enables premature terminator codons to be translated sometimes as amino acid codons thereby resulting in the synthesis of some full-length proteins.

Another approach that currently is being explored is the use of oligonucleotides that base pair with newly synthesized RNA and prevent defective regions of mRNA from being incorporated into mRNA via alternative splicing. If either approach is successful, many genetic diseases would be alleviated.

Many additional recoding phenomena are described in this book. The book will be useful to investigators in many fields, ranging from molecular biologists to clinical researchers who are interested in the genetic code, regulation of gene expression, or mechanisms of protein synthesis and codon translation.

Marshall Nirenberg

Preface

By 1966 the general nature of readout of the genetic code and codon identity had been established. What was not appreciated then was that decoding is dynamic. Decoding can be altered in an mRNA-specific manner and in a remarkable variety of ways.

The specific meaning of individual codons can be redefined in response to signals in an mRNA. Or a proportion of translating ribosomes can be diverted to a different reading frame at a specific site. And ribosomes can be directed to bypass a block of nucleotides or even to resume on a different mRNA. This book chronicles and analyzes these "recoding" phenomena both to understand the contribution they make to the complexity of gene expression and to understand the mechanisms involved, illuminating the features of ribosomes and mRNA.

These unusual genetic decoding events tell us that the readout of the code itself has been subject to the wiliness of selection, increasing the repertoire of ways to utilize the richness of information encoded in DNA or RNA. A coding sequence in mRNA can specify additional protein products not predicted from standard readout of the classical open reading frame. In some cases the recoding event is a control point for a regulatory circuit. In certain other cases, the key feature is specification of the "special" amino acids selenocysteine and pyrrolysine. Not surprisingly the world of viruses and small mobile chromosomal elements is rich with examples of recoding since their genomes are compact and every mechanism is used to maximize gene density. But, with one viral exception, the cases known so far of specification of the "special" amino acids are for cellular gene decoding.

Deciphering recoding has led to the realization that there is an extra layer of information in messenger RNA that can change the program for its own individual readout. These instructions include a site where the nonstandard decoding event occurs and an assortment of types of signals that greatly stimulate the proportion of ribosomes that perform the recoding event. These stimulatory signals can be 3' or 5' of the recoding site or both. The recoding signals located 3' can be nearby, or distant from the recoding site, and are often in the form of intra-mRNA structures (e.g., single stem-loops or pseudoknots) that somehow influence the ribosome. There are even translation factors that are specialized to specifically interact with some of these signals. Another set of signals involves mRNA pairing with the rRNA of translating ribosomes; in the established cases, the mRNA segment involved is 5'

of the recoding site. Yet another signal can be a particular sequence of amino acids in the growing nascent peptide acting within the peptide exit tunnel of the translating ribosome. How the ribosome senses and responds to this variety of signals is still quite unclear but is now becoming amenable to study due to the major advances in knowledge of ribosome structure and an emerging understanding of ribosome conformational changes during the translation cycle.

Redefinition. Carboxy terminal extensions of proteins can be programmed when the meaning of a UAG or UGA stop codon is redefined so that a proportion of ribosomes accepts a near-cognate aminoacyl-tRNA, such as that charged with glutamine (for UAG) or tryptophan (for UGA) instead of a release factor. Translation then continues in the zero frame to synthesize a "readthrough" protein which often contains an additional domain or two. UGA within an open reading frame can also be redefined in a different way, to specify the non-universal, 21st amino acid, selenocysteine, often located at the crucial active site of the enzyme product. Dramatically, multiple UGAs are redefined in selenoprotein P mRNA (10 in human and apparently 28 in sea urchin) for the purpose of transporting selenium. Redefinition of the UGAs in these mRNAs is clearly programmed because it is messenger specific; other UGAs in the same cell specify termination. However, in methanogens when UAG specifies the 22nd amino acid, pyrrolysine, there may be an ambiguous reassignment of the meaning of UAG. But, the specific context of an mRNA may enhance the specification of pyrrolysine.

In the inverse of stop codon redefinition, a sense codon in a specific context can mediate termination. In the case of the StopGo (also called "Stop-Carry on") phenomenon the specific sense codon specifies an amino acid, the protein chain is terminated, and translation continues on to make a second protein from the single ORF.

So far there is no known case of a simple programmed change in the meaning of a standard sense codon – switching one amino acid for another (though there is dynamic redefinition of an exceptional codon for tryptophan at some, but not other, positions in a particular mRNA in the ciliate *Euplotes*).

Redirection of linear readout. Ribosomal frameshifting links two overlapping ORFs, with a variety of mechanisms, a mix of functional results, and with a variety of mRNA-specific signals.

Most programmed frameshifting involves single nucleotide, −1 or +1 shifts (some −2 shifts are known). At least most of these cases involve a dissociation of anticodon:codon pairing, followed by tRNA:mRNA realignment and anticodon re-pairing to mRNA in a new frame (but the situation of Ty3 frameshifting in yeast appears different and in several cases of +1 frameshifting the initial pairing of the tRNA involved is not as stringent as generally occurs).

The known cases of programmed +1 frameshifting involve a slow-to-decode codon in the ribosomal A-site, either a stop codon or a sense codon for which the relevant aminoacyl-tRNA is limiting (a "hungry" codon). There is competition between the peptidyl-tRNA realigning forward and the tRNA or release factor for the zero frame A-site codon. Thus the first nucleotide of the A-site codon can be pivotal for frameshifting-mediated regulatory circuits.

Programmed −1 frameshifting generally yields a fixed ratio of shift to non-shift products: the product whose synthesis involved a frameshift event and the product of standard decoding. The most common type of −1 frameshifting involves tandem dissociation of the anticodon:mRNA pairing of tRNAs in both the P- and A-sites, followed by realignment and re-pairing of both mRNAs in the −1 frame, although re-pairing of only the A-site tRNA is likely to be involved in some cases.

A greatly exaggerated version of dissociation and re-pairing occurs when re-pairing of peptidyl-tRNA to mRNA occurs not at an overlapping codon but at a downstream triplet on the same mRNA, thus bypassing the mRNA sequence in-between. In the best characterized case, 50 nucleotides are bypassed by about half the ribosomes reading the message apparently due to the formation of mRNA structure within the bypassing ribosomes.

In an even more extreme case of redirection, coding resumption occurs on a specific, unique "mRNA," tmRNA. In this case a protein, SmpB, is crucial for resume site selection. tmRNA function was initially thought to be just an elegant mechanism for rescuing ribosomes stuck at the 3' end of aberrant mRNAs that lacked a terminator and for facilitating the destruction of the associated incomplete proteins. However, it is now apparent that tmRNA's role is more extensive as in some cases it is involved in regulation. Also there is emerging evidence of distant 5' nucleotide sequence in several mRNAs that influence tmRNA action.

Examples of Function. Many of the viruses that utilize recoding are of great medical or economic importance, and their mobile chromosomal gene counterparts have had a significant evolutionary impact. The panoply of decoding versatility and sophistication by compact genomes is common and accomplishes diverse goals. For instance, in some plant RNA viruses, frameshifting may be part of the strategy for preventing a logjam of opposing ribosomes and RNA dependent, RNA polymerase acting on the same RNA. In another example, recoding generates the retroviral GagPol polyprotein that results in the precursor form of reverse transcriptase being included in the virion by virtue of its linkage to a small proportion of Gag. This crucial linkage of Gag and Pol could also be accomplished by RNA splicing. But, this would be deleterious because the location of the RNA packaging site would result in virion packaging of subgenomic RNA yielding defective viruses.

Interestingly, the type of recoding utilized by murine leukemia virus for this purpose is programmed readthrough whereas that utilized by HIV is programmed frameshifting – two recoding solutions to the same problem.

Another case of using different types of nonstandard mechanisms to accomplish the same result is the expression of two DNA polymerase subunits from a single bacterial chromosomal *dnaX* gene. In *Escherichia coli*, decoding the standard ORF yields a product containing two carboxy terminal domains that are lacking in the product resulting from a ribosomal frameshift event two-thirds of the way through the ORF. This foreshortened protein likely has a role in translesion polymerase that helps deal with transition through lesions or obstacles on template DNA. Its synthesis is mediated by 50% efficient ribosomal frameshifting with ribosomes in the new frame quickly encountering a stop codon. In contrast, in *Thermus thermophilus*,

foreshortened products are derived from translation of the transcripts that result from transcriptional slippage at a run of A residues in the DNA. The population of mRNAs with varying numbers of extra nucleotides at the slippage site result in ribosome termination at now in-frame stop codons.

Evolution of recoding involves selection for both the position and the nature of the recoding site with its requisite stimulatory signals. In the absence of stimulatory signals, sites at which frameshifting or readthrough occur at low levels are, of course present. The current evidence suggests that, at least in bacteria, the most shift-prone sites that are not utilized for recoding are largely confined to poorly expressed mRNAs. For the sites whose "shifty" nature is dependent on scarcity of a particular tRNA, overexpression of an mRNA can lead to an increase in frameshifting raising a cautionary note for expression of high levels of proteins, often in nonhomologous systems, for biotechnological applications.

Scarcity of charged tRNAs can also be caused by amino acid starvation, a not uncommon state for bacteria. Starvation-induced frameshifting might be utilized to retune metabolism in response to the new growth state, so far this has not been shown.

Another consequence of recoding that needs further investigation is a possible under-appreciated role for frameshift-, bypassing-, and readthrough-derived events that do not exist to produce functional products. Ribosomes entering a region of mRNA not accessible by standard translation could have significant consequences on mRNA structure perhaps altering mRNA half-life. Alternatively, frameshifting within a coding sequence that yields early termination in a new frame could also affect mRNA half-life.

Recoding and Human Disease. Much remains unknown about the possible role of nonstandard translation in aging, viral infection, and certain autoimmune diseases. But the beginnings are there.

The stability of some of the proteins derived from ORFs not accessed by standard decoding is of particular interest from an immunological perspective. Preferential display on MHC class I molecules of peptides derived from short-lived proteins for activation of CD8+ T lymphocytes, this is important for the rapid CD8+ T-cell response to viral infection. Though the exact pathway for creating the array of peptides for display is not clear, models invoke rapidly degraded translation products. Some of these could be created by release of short nascent peptides due to ribosomal frameshifting.

Also, frameshifting may influence the severity of some of the triplet repeat diseases. The expanded string of repeats induces frameshifting leading to some product with poly-alanine in place of poly-glutamine.

Other genetic diseases involve frameshift mutations or substitutions that generate premature stop codons. If these new in-frame stop codons happen to be in a favorable context, small molecule drugs that alter translational fidelity can be used to phenotypically partially correct the mutations by stimulating synthesis of even a small portion of full-length product. This could alleviate the symptoms. Clinical trials in cystic fibrosis and Duchenne's muscular dystrophy are in an advanced stage.

It may also be possible to phenotypically correct certain frameshift mutants. Compensatory frameshifting can be stimulated by supplying a small RNA molecule to create a stimulatory signal in the mutant mRNA. Additions to tissue culture cells of such an RNA to create a signal just downstream of a frameshift mutant have yielded some positive results in optimal circumstances, but delivery problems remain.

Recoding events themselves may be targets for beneficial intervention. Since the ratio of Gag to GagPol is critical for HIV propagation, the efficiency of the frameshift event required for GagPol synthesis is a target for drug development. However, success depends on the host not having crucial similar targets. This is just one of the reasons for curiosity about the number of chromosomal genes that utilize the different types of frameshifting.

Foot and mouth disease virus appears to be a case in hand where it appears that the host cell does not use the unique StopGo recoding mechanism that the virus needs for propagation. This StopGo mechanism could be a target for antiviral development.

The path to recoding studies. The origin of knowledge about recoding has several different threads. In the mid-1960s, it was thought that decoding was so rigidly triplet that deviations from it would not be found, i.e., compensatory leakiness of frameshift mutations would not be detectable. And it was thought that mutants of translation components which would violate triplet decoding could not be found, i.e., external suppressors for frameshift mutants would not be isolatable. By 1972, both propositions were known to be incorrect.

Later that decade, an RNA phage-encoded product whose synthesis involved a frameshift event was detected. Also the balance of WT tRNAs was shown to be important for one type of frameshifting, and the relevance of noncognate codon:anticodon interaction was recognized. Nevertheless, the impact of these studies and of the discovery of a DNA phage frameshift product in 1983 was limited.

It was not until 1985–1987 that there were big breakthroughs in the detection of the utilization of specific frameshifting for gene expression. These cases are described in this book.

Redefinition of the meaning of one of the stop codons, UGA, was first discovered in the decoding of the coat protein gene of the RNA phage Qβ in the early 1970s. A proportion of translating ribosomes read through the stop codon by inserting an amino acid at the corresponding position in the protein. Not long afterward, essential readthrough was also shown for some plant viruses to make their RNA polymerase and for murine leukemia virus to make the GagPol precursor protein. This was accepted only slowly since the discovery of RNA splicing in 1977 provided a convenient explanation for accessing alternate open reading frames.

That selenocysteine was directly encoded by specific UGA stop codons, was discovered in 1986 at approximately the same time as the discovery of the initial cases of programmed frameshifting. The common features of reprogramming led to coining of the term “recoding” in 1992.

Recoding versus Reassignment. There seems to be a clear distinction between mRNA, site-specific, reassignment of codon meaning, and the complete reassignment, as for example in certain mitochondria. However, it is usual in biology for boundaries not to be sharp. Ambiguity arises where reassignment has not been fully refined as suggested above in the case of encoding pyrrolysine by UAG codons. For instance, a codon may be especially slow-to-decode, as with AGU and AGA in certain mitochondria. Perhaps surprisingly, the effects of such a codon in a fortuitous context may make a shift-prone site. Such a case may be evident in the common ancestor of the mitochondria of birds and turtles some 200 million years ago. It is thought that an extra nucleotide was present at an internal site in the coding sequence with frameshifting at a fortuitous "shifty" site restoring essential in-frame decoding. The extra nucleotide, and its associated compensatory frameshifting, is inferred to have been lost in many of the descendents of this common ancestor except in the mitochondrial decoding of the majority of extant birds and tortoises.

A parallel situation with an extra nucleotide occurs in a proportion of tracts of nine or more as in certain AT-rich endosymbionts such as *Buchnera aphidicola* which is associated with Aphids. However, in this case, the reading frame is restored by compensatory transcriptional slippage.

In the ciliate, *Euplotes*, UGA is reassigned so that it does not specify termination. It has been proposed that coincident changes in the release factor cause UAA, especially with a 3'A, to become unusually slow-to-decode. There is efficient frameshifting at AAA UAA A in *Euplotes* and required frameshifting occurs at this "terminator" sequence in a remarkable proportion of identified genes. Together with the mitochondrial frameshifting, *Euplotes* decoding illustrates more overlap between recoding and reassignment than encountered in other organisms.

Ancient decoding. Are there any cases of redefined meaning of a codon that are actually ancestral in an evolutionary sense? Consider UGA. Since special signals are required to change the meaning of UGA to specify selenocysteine, it is easiest to consider the standard termination meaning as ancestral. However, in early decoding there may not have been discrimination between cysteine and selenocysteine and perhaps at a stage before divergence of the common ancestor of bacteria, archaea, and eukaryotes, both amino acids were specified by UGN codons. In one version of this scenario, a next step was limitation of cysteine decoding to UGU and UGC, with UGA encoding selenocysteine. As the original anaerobic atmosphere changed to an aerobic one with the advent of an oxygen-rich atmosphere some 2.4 billion years ago, there could have been selection against oxygen-labile selenocysteine except where it was especially advantageous. Perhaps this "restriction stage" is when selenocysteine-recoding signals started to arise, and non-tagged UGA codons later acquired the termination meaning. Such a model is in marked contrast to the obvious one in which the termination meaning was ancestral.

In modern bacteria UGA specifies selenocysteine only if it is followed by a specific stem-loop structure in the mRNA. It is a reasonable supposition, although no more than that, that a 3' nearby stem-loop structure became important for selenocysteine specification in the common ancestor of bacteria, archaea, and eukaryotes.

In modern eukaryotes a specific structure in the 3' untranslated region is required. However, some eukaryotic mRNAs that encode selenocysteine-containing proteins also have some "remnant" of a stimulatory structure just 3' adjacent to the UGA. This element likely preceded the emergence of specific structures in the 3' UTR.

At a much earlier time than selenocysteine specification, during the evolution of decoding itself, it seems likely that primitive readout was incapable of being anything other than slipshod. At this time polyamines may have been playing a protein-like role in primitive ribosomes. The result likely was a plethora of products serving as food for selection. As triplet decoding and codon assignment became locked in, was there a parallel refinement of alternative decoding? Or did the currently observed alternative decoding evolve later as a sophisticated refinement after a period of tediously standard decoding?

Frameshifting for expression of bacterial release factor 2 decoding also has an ancient origin. Its hallmark is stimulation of the frameshift event by pairing between mRNA and rRNA during translation. We can wonder whether this interaction between mRNA and rRNA in ribosomes in the act of translating might not itself have an ancient origin. Could interactions of this type have helped to grip the message?

In modern day ribosomes, it is anticodon pairing that holds the mRNA in place. Detachment and realignment lead to frameshifting, at least in most cases. There is an appealing if somewhat controversial suggestion that standard frame maintenance is maintained by pairing two tRNAs at all times. In this scenario, anticodon pairing by E-site tRNA does not dissociate until A-site aminoacyl-tRNA pairing is established. So strong ribosomal gripping of tRNA would lead to the in-frame grip of the mRNA. However, the E-site appears to be a late addition in ribosome evolutionary history since it is protein-rich. Therefore, before it existed, what served to clasp mRNA? One candidate is the rRNA:mRNA Shine–Dalgarno pairing which was discovered because of its role in initiation of protein synthesis in bacteria. Programmed frameshifting studies have revealed that this interaction is not unique to initiation in that the anti-Shine–Dalgarno sequence of translating ribosomes can scan the mRNA being decoded for potential complimentarity. After such a rRNA:mRNA hybrid forms, the ribosome continues translation for up to 10 nucleotides before the hybrid ruptures. Whether interactions of this type played a role in primordial protein synthesis is of course unknown. But, if so, rather than the primordial coding sequences having been G-rich, perhaps there could have been blocks of coding sequences spanned by G-rich noncoding "anchors" that decoding could bypass. Setting aside such speculative "excesses," recoding studies are clearly contributing to our knowledge of standard decoding and scanning by the anti-Shine–Dalgarno sequences of translating ribosomes is one of several cases in point.

Transcription slippage (also called pseudo-templated transcription or stuttering)

Realignment during transcription parallels translational realignment. A few examples are mentioned above where transcription slippage substitutes for cases of programmed frameshifting. In these cases there has been selection for high-level

transcription slippage at specific sites. Such slippage yields mRNAs with inserts of one or more nucleotides – in a bacterial case a diminishing series of mRNAs with up to 15 additional nucleotides and a small minority with deletions of one or a few nucleotides. Standard translation of these mRNAs yields unique products. Instead of the detachment of triplet anticodon pairing, dissociation of the nascent RNA hybrid with template DNA in the transcription bubble is involved. The identity of flanking sequence can delimit the number of extra nucleotides inserted to 1. But whether the flanking sequence can also enhance the frequency, possibly even by the ability of the nascent RNA chain to form a short stem, remains to be seen.

Editing of preformed transcripts can also have consequences similar to several types of recoding. For instance, mRNA editing that changes a stop codon to a sense codon can give the equivalent of stop codon readthrough. Similarities even extend to variable efficiencies of the process and to the importance of mRNA structure. Editing to change the identity of one sense codon to another in a proportion of the mRNAs, constitutes a type of diversity for which there is only one specialized recoding counterpart. It will be fascinating to discover to what extent nonstandard transcription and RNA editing parallel and substitute for their translational counterparts.

Future. As this book attests, our knowledge of recoding has a firm basis but much remains to be done. Together with studies of mutants of ribosomal components, advances in structural information about translation components now are offering the prospect of an understanding of how ribosomes sense and respond to recoding signals. The deluge of sequence information is providing exciting bioinformatic opportunities for comparative analyses to reveal the extent of recoding and transcription slippage. And a dramatic recent advance in determining ribosome location en masse at sub-codon resolution by sequencing vast numbers of mRNA segments protected within ribosomes at a specific time, has great potential in this regard.

Knowledge of the "dark matter" of the genome, those transcribed regions that do not encode mRNA, tRNA, or rRNA, is rapidly showing the complex roles of small RNAs in gene expression. Are some cases of recoding influenced by them?

We look forward to discovering the answers to these and questions not yet asked.

Acknowledgment: We thank Ken Keiler for instigating the American Society of Microbiologists' session that inspired Andrea Macaluso (Springer) to propose this book, our past colleagues, especially Bob (R.B.) Weiss and Alan Herr, and our current colleagues. We also thank Marshal Nirenberg, the pioneer of codon identification, for his generous contribution.

John Atkins and Ray Gesteland

Contents

Contributors

Michael Anikin Department of Cell Biology, School of Osteopathic Medicine, University of Medicine and Dentistry of New Jersey, Stratford, NJ 08084, USA.

Pavel V. Baranov Biochemistry Department, University College Cork, Cork, Ireland.

David M. Bedwell Department of Microbiology, University of Alabama at Birmingham, Birmingham, AL 35294-2170, USA.

Michaël Bekaert School of Biology and Environmental Science, University College Dublin, Ireland.

Marla J. Berry Department of Cell and Molecular Biology, John A. Burns School of Medicine, University of Hawaii at Manoa, Honolulu, HI 96813, USA.

Léa Brakier-Gingras Département de Biochimie, Université de Montréal, Montréal, Québec, H3T 1J4, Canada.

Ian Brierley Division of Virology, Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QP, UK.

Jeremy D. Brown Institute for Cell & Molecular Biosciences, The Medical School, Newcastle University, Newcastle upon Tyne NE2 4HH, UK.

Jonathan D. Dinman Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA.

Dominic Dulude Département de Biochimie, Université de Montréal, Montréal, Québec, H3T 1J4, Canada; Centre de Recherche, Hôpital Sainte-Justine, Montréal, Québec, H3T 1C5, Canada.

Philip J. Farabaugh Department of Biological Sciences and Program in Molecular and Cell Biology, University of Maryland Baltimore County, Baltimore, MD 21250, USA.

Olivier Fayet Centre National de la Recherche Scientifique, Laboratoire de Microbiologie et Génétique Moléculaires, Université de Toulouse, F-31000 Toulouse, France.

Andrew E. Firth BioSciences Institute, University College Cork, Cork, Ireland.

David P. Giedroc Department of Chemistry, Indiana University, Bloomington, IN 47405-7102, USA.

Robert J. C. Gilbert Division of Structural Biology, Henry Wellcome Building for Genomic Medicine, University of Oxford, Oxford OX3 7BN, UK.

Vadin N. Gladyshev Department of Biochemistry and Redox Biology Center, University of Nebraska, Lincoln, NE 68588, USA.

Olga Gurvich Cork Cancer Centre, BioSciences Institute, University College Cork, Cork, Ireland.

Dolph L. Hatfield Molecular Biology of Selenium Section, Laboratory of Cancer Prevention, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA.

Roger W. Hendrix Pittsburgh Bacteriophage Institute & Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA.

Michael T. Howard Department of Human Genetics, University of Utah, Salt Lake City, UT 84112-5330, USA.

Ivaylo P. Ivanov BioSciences Institute, University College Cork, Cork, Ireland; Department of Human Genetics, University of Utah, Salt Lake City, UT 84112-5330, USA.

Kim M. Keeling Department of Microbiology and Gregory Fleming James Cystic Fibrosis Research Center, University of Alabama at Birmingham, Birmingham, AL 35294-2170, USA.

Kenneth C. Keiler Department of Biochemistry and Molecular Biology, Penn State University, 401 Althouse Laboratory, University Park, PA 16802, USA.

Joseph A. Krzycki Department of Microbiology, Ohio State University, Columbus, Ohio 43210, USA.

Dennis M. Lee Department of Biochemistry and Molecular Biology, Penn State University, 401 Althouse Laboratory, University Park, PA 16802, USA.

Senya Matsufuji Department of Molecular Biology, The Jikei University School of Medicine, 3-25-8 Nishi-shinbashi, Minato-ku, Tokyo 105-8461, Japan.

William T. McAllister Department of Cell Biology, School of Osteopathic Medicine, University of Medicine and Dentistry of New Jersey, Stratford, NJ 08084, USA.

W. Allen Miller Plant Pathology Department, and Biochemistry, Biophysics & Molecular Biology Departments, Iowa State University, Ames, IA 50011, USA.

Vadim Molodtsov Department of Cell Biology, School of Osteopathic Medicine, University of Medicine and Dentistry of New Jersey, Stratford, NJ 08084, USA.

Olivier Namy IGM, CNRS, UMR 8621, F 91405 Orsay, France and Université Paris-Sud, F 91405 Orsay, France.

Knud H. Nierhaus Max-Planck-Institut für Molekulare Genetik, D-14195 Berlin, Germany.

Michael O'Connor School of Biological Sciences, University of Missouri-Kansas City, Kansas City, MO 64110, USA.

Markus Pech Max-Planck-Institut für Molekulare Genetik, D-14195 Berlin, Germany.

Simon Pennell Division of Molecular Structure, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK.

Marie-Françoise Prère Centre National de la Recherche Scientifique, UMR5100, Laboratoire de Microbiologie et Génétique Moléculaires, Université de Toulouse, F-31000 Toulouse, France.

Jean-Pierre Rousset IGM, CNRS, UMR 8621, Orsay, F 91405 France, Université Paris-Sud, Orsay, France.

Martin D. Ryan Centre for Biomolecular Sciences, Biomolecular Sciences Building, North Haugh, University of St. Andrews, St. Andrews KY16 9ST, UK.

Dmitry Temiakov Department of Cell Biology, School of Osteopathic Medicine, University of Medicine and Dentistry of New Jersey, Stratford, NJ 08084, USA.

Oliver Vesper Max-Planck-Institut für Molekulare Genetik, D-14195 Berlin, Germany.

Norma M. Wills Department of Human Genetics, University of Utah, Salt Lake City, UT 84112-5330, USA.

Daniel N. Wilson Gene Center, Ludwig-Maximilians-Universität München, D-81377 München, Germany.

Hiroshi Yamamoto Max-Planck-Institut für Molekulare Genetik, D-14195 Berlin, Germany.

Part I
Redefinition

Chapter 1
Selenocysteine Biosynthesis, Selenoproteins, and Selenoproteomes

Vadim N. Gladyshev and Dolph L. Hatfield

Abstract Selenocysteine (Sec), the 21st amino acid in the genetic code, is encoded by UGA. The pathway of Sec biosynthesis in eukaryotes has only recently been discovered. Sec is constructed on its tRNA that is initially aminoacylated with serine and modified to a phosphoseryl-tRNA intermediate with the help of several dedicated enzymes. More than 50 selenoprotein families are now known with most selenoproteins being oxidoreductases. Development of bioinformatics tools led to the identification of entire sets of selenoproteins in organisms, selenoproteomes, which in turn helped explain biological and biomedical effects of dietary selenium and identify new functions of selenium in biology. Roles of selenium and selenoproteins in health have also been addressed through sophisticated transgenic/knockout models that targeted removal or modulation of Sec tRNA expression.

Contents

V.N. Gladyshev (✉)
Department of Biochemistry and Redox Biology Center, University of Nebraska, Lincoln, NE 68588 USA
e-mail: vgladyshev@rics.bwh.harvard.edu

J.F. Atkins, R.F. Gesteland (eds.), *Recoding: Expansion of Decoding Rules Enriches Gene Expression*, Nucleic Acids and Molecular Biology 24,
DOI 10.1007/978-0-387-89382-2_1,

1.1 UGA is Recoded for Sec

1.1.1 Variations in the Genetic Code

The genetic code was deciphered and shown to be universal by the mid-1960s (see Nirenberg et al. 1966 and references therein). All 64 code words in the code were assigned to amino acids or a specialized function. One code word, AUG, was recognized to have a dual function serving to dictate the initiation of protein synthesis and to code for the insertion of methionine at internal protein positions. Three code words, UAG, UAA, and UGA, were assigned specialized roles of dictating the cessation of protein synthesis. It was assumed at that time that there was no more room in the code for another (or other) amino acid(s) and the possibility that code words other than AUG might have dual functions was not considered.

There have been several major variations reported in the genetic code, however, since the mid-1960s. It was initially recognized that not all organelles use the same genetic language, and subsequently, that some organisms use a different genetic language. For example, variations in the universal genetic code were observed in mitochondria and chloroplasts (reviewed in Jukes and Osawa 1990; Yokobori et al. 2001) and in organisms such as mycoplasma that use UGA to code for tryptophan instead of termination (Yamao et al. 1985), *Euplotes* that use UGA to code for cysteine instead of termination (Meyer et al. 1991), and several species of *Candida* that use CUG to code for serine instead of leucine (reviewed in Pesole et al. 1995). Furthermore, some bacteria and archaea use GUG and/or UUG as start codons instead of the universal codon, AUG (Bell and Jackson 1988). Interestingly, evidence in the mid-1980s suggested that the termination codon, UGA, likely had a dual function. The gene sequences of the selenium-containing proteins, glutathione peroxidase 1 (GPx1) in mammals (Chambers et al. 1986) and formate dehydrogenase in *Escherichia coli* (Zinoni et al. 1986), showed that both genes had an in-frame TGA codon in their open reading frames that aligned with Sec in the corresponding proteins. These correlations suggested that UGA coded for Sec, but this assignment could not be made without further experimental evidence as the available data at that time had shown that the serine moiety (Sunde and Evenson, 1987) initially attached to a minor Sec tRNA that decoded UGA was converted to phosphoseryl-tRNA by a kinase (Hatfield, Diamond and Dudock 1982). Thus, it was possible that phosphoserine was incorporated into protein and then posttranslationally modified to Sec making phosphoserine the 21st amino acid in the genetic code. This point was clarified when Sec was indeed shown to be biosynthesized on its tRNA in both bacterial (Leinfelder et al. 1989) and mammalian cells (Lee et al. 1989a). These two studies

provided the first direct evidence that Sec was the 21st amino acid and that UGA was therefore recoded for Sec in those organisms that synthesize selenoproteins. The expanded genetic code that includes Sec is shown in Fig. 1.1

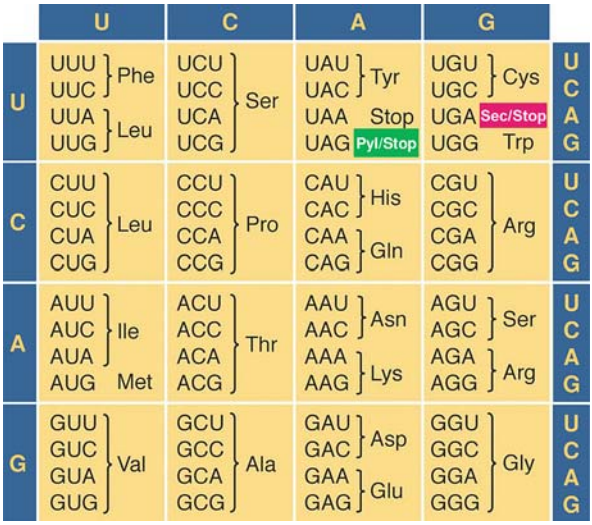

Fig. 1.1 The genetic code. Sec, encoded by UGA, is highlighted to show that it is the 21st amino acid in the genetic code. A 22nd amino acid, pyrrolysine (Pyl), encoded by UAG, is also shown

It should also be noted that pyrrolysine was recently added to the genetic code as the 22nd amino acid (see Fig. 1.1) (Srinivasan et al. 2002; Hao et al. 2002) which is described in Chapter 3 by Krzycki. The possibility that a 23rd amino acid may also occur in the code has been considered, and although not likely to occur, has not been completely ruled out (Lobanov et al. 2006a). If a 23rd amino acid exists in the code, it would be much less widespread than Sec and may be limited to only a few organisms. Another variation in the genetic code was recently found wherein a single code word can code for two different amino acids, not only in the same organism but also within the same gene (Turanov et al. 2009). UGA was shown to specify the incorporation of Cys and Sec in a single mRNA in the *Euplotes* genus and the structural arrangements of the mRNA preserve the location-dependent dual function of the UGA codon.

1.2 Biosynthesis of Sec

A number of factors had been identified in higher vertebrates over the years that play a role in the biosynthesis of Sec and its insertion into protein. The components involved in the biosynthesis of Sec are discussed below, while the chapter by Berry and Howard (Chapter 2) focuses on those components involved with the incorporation of this amino acid into protein. The principle factors that have

been associated with Sec biosynthesis in eukaryotes are Sec tRNA, seryl-tRNA synthetase (SerS), phosphoseryl-tRNA kinase (PSTK), Sec synthase (SecS), and selenophosphate synthetases 1 and 2 (SPS1 and 2). They are described in greater detail below.

1.2.1 Unique Features of Sec tRNA

Sec tRNA is undoubtedly the most unique tRNA identified to date. For example, its transcription begins, unlike any known tRNA, at the first nucleotide within the coding region of its gene (Lee et al. 1987), while all other tRNAs are transcribed with a leader sequence that must be processed. The upstream regulatory sites that govern the transcription of Sec tRNA are unique for tRNA (reviewed in detail elsewhere (Hatfield et al. 1999)). The mature form of the tRNA has a triphosphate on its 5′-end (Lee et al. 1987). It is the longest tRNA sequenced, ranging in length from 90 to 93 nucleotides in some lower eukaryotes (Mourier et al. 2005; Lobanov et al. 2006b) to 95 in *E. coli* and more than a 100 nucleotides in various other prokaryotes (Heider and Bock, 1993). Sec tRNAs in higher vertebrates contain only five modified nucleosides, whereas up to 15–17 modified nucleosides have been identified in other tRNAs. The fact that Sec tRNA is initially aminoacylated with serine, but is the tRNA for Sec, has resulted in it being designated as Sec $tRNA^{[Ser]Sec}$ (Hatfield et al. 1994). The secondary structure of $tRNA^{[Ser]Sec}$ found in mammals and *Plasmodium falciparum* is shown as a cloverleaf model in Fig. 1.2.

The modified nucleosides in $tRNA^{[Ser]Sec}$ are 1-methyladenosine (m^1A) at position 58, pseudouridine (ψU) at position 55, N^6-isopentenyladenosine (i^6A) at position 37, and either 5-methoxycarbonylmethyluridine (mcm^5U) or methoxycarbonylmethyl-2′-*O*-methyluridine (mcm^5Um) at position 34, which is the wobble position of tRNA (Hatfield et al. 2006). The synthesis of the methyl group at position 34 is the last step in the maturation of Sec $tRNA^{[Ser]Sec}$ and this 2′-*O*-methyluridine is designated Um34. Interestingly, the synthesis of Um34 is stringently dependent on primary structure and on intact secondary and tertiary structures of $tRNA^{[Ser]Sec}$; i.e., the addition of Um34 cannot occur without the prior synthesis of m^1A, ψU, i^6A, and mcm^5U, and disruption of the secondary or tertiary structure of the tRNA inhibits its attachment (Kim et al. 2000). Furthermore, Um34 formation is dependent on selenium status (reviewed in Hatfield et al. 2006). Under conditions of selenium deficiency, the ratio of mcm^5U/mcm^5Um shifts dramatically in mammalian organs, tissues, and cells from the latter to the former isoforms, and vice versa under conditions of selenium sufficiency (Chittum et al. 1997). Finally, the addition of Um34 to Sec $tRNA^{[Ser]Sec}$ results in striking changes in secondary and tertiary structures.

The above observations relating to the synthesis of Um34 led us to propose that this maturation step was a highly specialized event yielding mcm^5Um, an isoform with a different function in selenoprotein synthesis than its precursor, mcm^5U (Moustafa et al. 2001). This hypothesis was later confirmed as discussed below in the section on **Sec.**

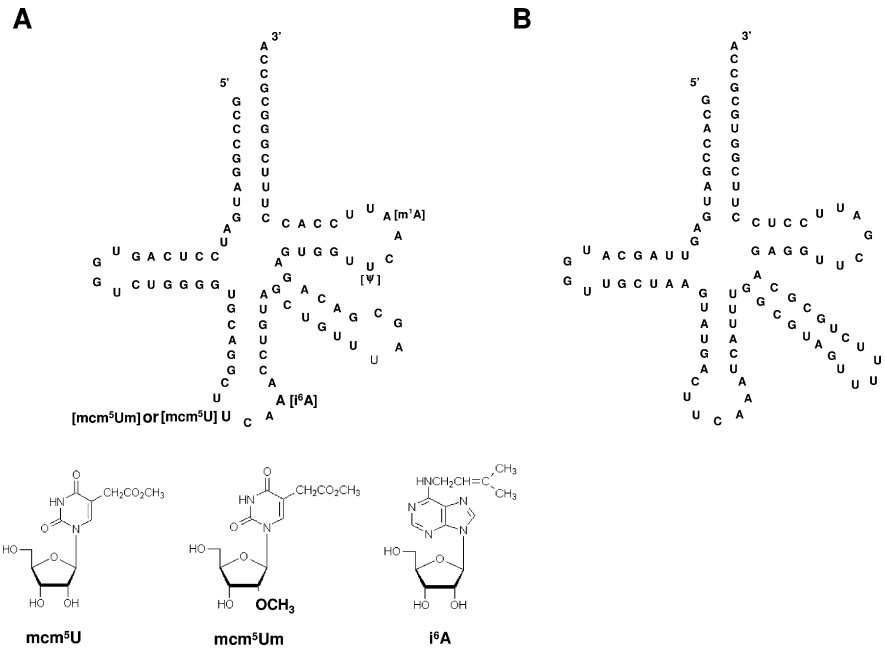

Fig. 1.2 Mammalian and *P. falciparum* Sec tRNA$^{[Ser]Sec}$. (**A**) Mammalian tRNA$^{[Ser]Sec}$. The structures of the modified bases in the anticodon loop of mammalian tRNA$^{[Ser]Sec}$, i^6A, mcm^5U, and mcm^5Um, are also shown. (**B**) *P. falciparum* tRNA$^{[Ser]Sec}$ is 93 nucleotides long and mammalian tRNA$^{[Ser]Sec}$ is 90 nucleotides long and the extra bases occur in the long extra arm (see text). The mammalian tRNA$^{[Ser]Sec}$ structure was determined by sequencing the tRNA (Hatfield et al. 2006), while the *P. falciparum* tRNA$^{[Ser]Sec}$ structure is based on sequencing its gene (Lobanov et al. 2006b), wherein the CCA 3′-terminus, which is added posttranscriptionally, is shown in the figure

1.2.2 tRNA Knockout and Transgenic Mouse Models

Another novel feature of tRNA$^{[Ser]Sec}$ is that it has nine paired bases in the acceptor stem and four in the TψC stem, i.e., it exists in a 9/4 cloverleaf form (Böck et al. 1991; Hubert et al. 1998). Other tRNAs have seven paired bases in the acceptor stem and five paired bases in the TψC stem, i.e., they exist in a 7/5 cloverleaf model. An additional novel feature of tRNA$^{[Ser]Sec}$ is that the D-stem may contain six base pairs while other tRNAs have three to four base pairs in this stem. There are numerous other characteristics of tRNA$^{[Ser]Sec}$ that distinguish it from other tRNAs and these have been reviewed in detail elsewhere (Hatfield et al. 1999).

1.2.3 Aminoacylation of Sec tRNA$^{[Ser]Sec}$

Sec tRNA$^{[Ser]Sec}$ is aminoacylated with serine by SerS which is the initial step in the biosynthetic pathway of Sec (Lee et al. 1989a; Leinfelder et al. 1989). The identity elements in Sec tRNA$^{[Ser]Sec}$ for its aminoacylation therefore must correspond to

those in SerS. The identity elements in mammalian tRNA$^{[Ser]Sec}$ have been identified and the major areas are the discriminator base and the long extra arm which have essential roles in aminoacylation (Wu et al. 1993; Ohama et al. 1994). Other regions of tRNA$^{[Ser]Sec}$ that have identity roles are located in the acceptor, TψC, and D-stems (Amberg et al. 1996). Once the tRNA is aminoacylated with serine, the serine moiety serves as the backbone for the synthesis of Sec in prokaryotes and eukaryotes (reviewed in Hatfield and Gladyshev 2002).

1.2.4 Phosphoseryl-tRNA$^{[Ser]Sec}$kinase

A kinase activity that phosphorylates a minor seryl-tRNA to form phosphoseryl-tRNA was identified many years ago in rooster liver by Maenpaa and Bernfield (1970). About the same time, a minor seryl-tRNA in bovine, rabbit, and chicken livers that recognized specifically the nonsense codon, UGA, was reported (Hatfield and Portugal, 1970). Subsequently, the phosphoseryl-tRNA identified in rooster liver and the UGA decoding seryl-tRNA were later shown to be selenocysteyl-tRNA$^{[Ser]Sec}$ (Lee et al. 1989a, b). The significance of the phosphoseryl-tRNA$^{[Ser]Sec}$ kinase (PSTK) that phosphorylated seryl-tRNA$^{[Ser]Sec}$ to form phosphoseryl-tRNA$^{[Ser]Sec}$ was not assessed until PSTK was isolated and characterized. The kinase activity remained elusive for many years, but was finally identified by combining bioinformatics and biochemistry approaches (Carlson et al. 2004a). That is, we examined completely sequenced genomes for kinase genes occurring in archaea that synthesized selenoproteins, but absent in archaea that lacked selenoproteins, and identified four candidates. The completely sequenced genomes of *Caenorhabditis elegans* and *Drosophila* that were known to synthesize selenoproteins were then searched for homologous sequences to these four kinase genes that were in turn not present in the genome of *Saccharomyces cerevisiae* which did not make selenoproteins. A single candidate kinase was detected using this strategy. Since a gene was present in the mouse genome with homology to the candidate *pstk* gene, it was cloned, its product expressed, characterized, and identified as PSTK (Carlson et al. 2004a). PSTK used seryl-tRNA$^{[Ser]Sec}$ and ATP as substrates and Mg^{++} as a cofactor to yield *O*-phosphoseryl-tRNA$^{[Ser]Sec}$ and ADP. At the time this work was reported, the role of PSTK and its product, phosphoseryl-tRNA$^{[Ser]Sec}$, had not been determined.

1.2.5 Sec Synthase (SecS) and Selenophosphate Synthetase (SPS)

SecS, which was designated SelA in prokaryotes, was initially identified and characterized in *E. coli* by Bock and collaborators (Böck et al. 1991). *E. coli* seryl-tRNA$^{[Ser]Sec}$ served as a substrate for SelA and was converted to an intermediate, which is most likely dehydroalanyl-tRNA$^{[Ser]Sec}$ (reviewed in Böck et al. 1991). The active selenium donor, monoselenophosphate (SeP), is synthesized from

selenide and ATP by SPS (SelD) in prokaryotes (Glass et al. 1993). The intermediate, dehydroalanyl-tRNA$^{[Ser]Sec}$, while still bound to SelA, accepts SeP to generate selenocysteyl-tRNA$^{[Ser]Sec}$ which is now ready to incorporate Sec into protein (Böck et al. 1991).

A gene with homology to *selA* was not found in archaea or eukaryotes. However, a candidate SecS was subsequently identified in eukaryotes and archaea by comparative genomic analysis of completely sequenced eukaryotic and archaeal genomes as was carried out in detecting *pstk* (Xu et al. 2006). The survey searching for a eukaryotic and archaeal SecS resulted in the identification of genes co-occurring with known components in the Sec insertion machinery and, in addition, a candidate SecS was detected in mammals. This protein had previously been found in cell extracts from patients with an autoimmune chronic hepatitis as an autoimmune factor that co-precipitated with tRNA$^{[Ser]Sec}$ (Gelpi et al. 1992). This factor was designated as the soluble liver antigen (SLA). SLA was found to be a PLP-dependent transferase (Kernebeck et al. 2001) and also to bind other components involved in Sec metabolism (Xu et al. 2005; Small-Howard et al. 2006). SLA occurred in all eukaryotic and archaeal selenoprotein synthesizing organisms that were examined by comparative genomic analysis, but not in those organisms not synthesizing selenoproteins, nor in any prokaryotic organism whether it did or did not make selenoproteins (Xu et al. 2006).

The mouse gene for SLA (*SecS*) was cloned, the protein expressed, and the function of SLA established by experimental analysis (Xu et al. 2006). *O*-phosphoseryl-tRNA$^{[Ser]Sec}$ was dephosphorylated by SLA to yield Pi and a product that bound to the enzyme. The product that remained bound to SLA, which was an intermediate in the biosynthesis of Sec, was likely not seryl-tRNA$^{[Ser]Sec}$ as seryl-tRNA$^{[Ser]Sec}$ did not itself bind to SLA. Dehydroalanine is likely the intermediate generated by mammalian SecS (Xu et al. 2006), which was the same intermediate identified in *E. coli* (Böck et al. 1991).

selD has two homologous genes in mammals, designated *sps1* and *sps2* (Kim and Stadtman 1995; Low, Harney and Berry 1995; Guimaraes et al. 1996) that were initially proposed to serve as SPS. The product of *sps*2, which is SPS2, is a selenoprotein and can therefore serve as an autoregulator of selenoprotein synthesis (Guimaraes et al. 1996; Kim et al. 1997), as it is indeed the enzyme that synthesizes SeP in mammals (Xu et al. 2006). In studies that further elucidated the roles of SPS1 and SPS2 in mammals, the Sec moiety in SPS2 was mutated to Cys, wherein the mutant was found to have low enzyme activity (Guimaraes et al. 1996; Kim et al. 1997, 1999), but was capable of complementing *selD* minus *E. coli* cells transfected with the mutant mammalian *sps*2 (Kim et al. 1999). Other studies involved complementing *selD* minus *E. coli* cells that had been transfected with either *sps*1 or Sec$^-$ *sps*2, and they suggested that SPS1 has a role in recycling Sec via a selenium salvage pathway, whereas SPS2 was involved in the synthesis of SeP (Tamura et al. 2004). However, these studies did not directly demonstrate the roles of SPS1 and SPS2 in Sec biosynthesis.

To further clarify the roles of SPS1 and SPS2 in Sec biosynthesis, *C. elegans* SPS2, which naturally has Cys instead of Sec at its active site, mouse SPS2

containing a Sec→Cys mutation, *E. coli* SelD and mouse SPS1 were prepared and their abilities to generate SeP from selenide and ATP were determined (Xu et al. 2006). Each SPS synthesized SeP with the exception of mouse SPS1 demonstrating that SPS2, and not SPS1, was SPS in higher animals.

It should first be noted, however, that none of the earlier studies had shown that SeP could serve directly as the selenium donor which would unequivocally demonstrate that SeP was the active selenium donor. SeP was therefore synthesized chemically and added to Sec biosynthesis reactions (Xu et al. 2006). SeP and *O*-phosphoseryl-tRNA$^{[Ser]Sec}$ incubated in the presence of mouse SecS did indeed generate Sec. Reactions containing seryl-tRNA$^{[Ser]Sec}$ in place of *O*-phosphoseryl-tRNA$^{[Ser]Sec}$, with or without SeP, or containing another protein in place of SecS, did not form Sec. These reactions unambiguously proved that SeP is the active selenium donor in Sec biosynthesis (Xu et al. 2006). Reactions containing mouse SecS, *O*-phosphoseryl-tRNA$^{[Ser]Sec}$, mouse mutant Sec→Cys SPS2, selenide and ATP produced selenocysteyl-tRNA$^{\{Ser\}Sec}$, but seryl-tRNA$^{[Ser]Sec}$ in place of *O*-phosphoseryl-tRNA$^{[Ser]Sec}$, or mouse SPS1 in place of SPS2, did not. Thus, SPS2 synthesizes SeP and SPS1 must have another role that may or may not be related to selenoprotein biosynthesis (Lobanov, Hatfield and Gladyshev 2008a). In addition to unequivocally demonstrating that SeP is the active donor and that SPS2, and not SPS1, is the SPS in higher animals, the above in vitro studies showed that SLA is the mammalian SecS and *O*-phosphoseryl-tRNA$^{[Ser]Sec}$ is the correct intermediate in the pathway. At the same time these studies on elucidating the Sec biosynthetic pathway were being carried out, the archaeal SLA gene (*SecS*) was cloned, expressed, and the gene product shown to convert *O*-phosphoseryl-tRNA$^{[Ser]Sec}$ to selenocysteyl-tRNA$^{[Ser]Sec}$ (Yuan et al. 2006).

The roles of SPS1 and SPS2 were further elucidated intracellularly in a complementary study (Xu et al. 2007). SPS1 and SPS2 were knocked down using RNA interference technology in NIH 3T3 cells and the effect of their loss on selenoprotein synthesis examined. Selenoprotein synthesis was abolished completely by the removal of SPS2, but was unaffected by removal of SPS1. The knockdown cells were then used for transfection with SPS2, SelD, or SPS1. Either SPS2 or SelD complemented the loss of SPS2, but SPS1 did not. These "in vivo" studies showed that SPS2, which synthesizes SeP (Xu et al. 2006), is essential for selenoprotein biosynthesis, but SPS1 is not (Xu et al. 2007). Furthermore, SPS1 has been found to occur in animals in which the SPS2 and the other Sec insertion machinery have been lost providing additional evidence that SPS1 has roles other than in Sec biosynthesis and its insertion into protein (Lobanov et al. 2008a).

1.2.6 The Sec biosynthetic pathway

The entire Sec biosynthetic pathway in eukaryotes is shown in Fig. 1.3. The pathway begins with the aminoacylation of tRNA$^{[Ser]Sec}$ with serine by SerS (Sunde and Evenson, 1987; Lee et al. 1989a; Leinfelder et al. 1989). PSTK phosphorylates the serine moiety to form *O*-phosphoseryl-tRNA$^{[Ser]Sec}$ (Carlson et al. 2004a) which

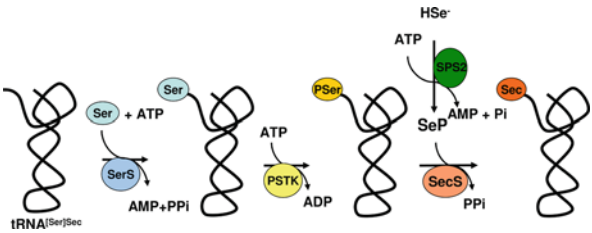

Fig. 1.3 Biosynthesis of Sec in eukaryotes and archaea. Abbreviations of the factors involved in Sec biosynthesis are defined in the text

then serves as a substrate for SecS that hydrolyzes the phosphate group to form the acceptor molecule for SeP, likely dehydroalanyl-tRNA$^{[Ser]Sec}$, that remains bound to SecS (Xu et al. 2006). SPS2 synthesizes SeP, the active selenium donor, using selenide and ATP as substrates and with the addition of SeP to the intermediate attached to SecS, the synthesis of Sec is complete. This pathway established how the 21st and last known eukaryotic amino acid in the genetic code whose biosynthesis had not been established, is synthesized.

Although PSTK is not found in eubacteria, SelA can use *O*-phosphoseryl-tRNA$^{[Ser]Sec}$ as a substrate (Xu et al. 2006). The major difference in the biosynthesis of Sec in eubacteria and in eukaryotes and archaea is the extra step involving *O*-phosphoseryl-tRNA$^{[Ser]Sec}$ which is synthesized using seryl-tRNA$^{[Ser]Sec}$ and ATP by PSTK. *O*-phosphoseryl-tRNA$^{[Ser]Sec}$ then serves as a substrate for SecS in eukaryotes and archaea, whereas seryl-tRNA$^{[Ser]Sec}$ is a substrate for eubacterial SecS. Selenocysteyl-tRNA$^{[Ser]Sec}$ is now poised to be incorporated into selenoproteins (see Chapter 2 by Berry and Howard).

1.3 Identification of Selenoproteins in Sequence Databases

The major form of selenium in cells is Sec residues in proteins. This is illustrated, for example, by the analyses of mice, in which the tRNA$^{[Ser]Sec}$ gene is disrupted in liver (Carlson et al. 2004b). In these animals, liver selenium content is significantly reduced. Similarly, selenoproteins account for most of selenium in body fluids. For example, the main selenoprotein in plasma of mammals is selenoprotein P (SelP), which accounts for more than half of selenium in plasma (Burk and Hill 2005). Selenium may also occur in selenoproteins in the form of a cofactor. For example, in several bacterial selenium-containing molybdoenzymes, such as xanthine dehydrogenase and nicotinic acid hydroxylase, a Se-Mo cofactor is formed in the active site (Gladyshev et al. 1994). This labile cofactor is easily destroyed releasing both elements. The possibility that similar enzymes exist in eukaryotes has not been addressed.

Sec-containing proteins are often misannotated in sequence databases. This is because their TGA codons are interpreted as stop signals by available annotation

tools (Gladyshev et al. 2004). It is obviously impossible to identify selenoprotein genes by only searching for TGA codons. However, selenoprotein genes have an RNA structure known as the Sec insertion sequence (SECIS) element (see Chapter 2 for details). SECIS elements are highly specific for selenoprotein genes and possess a sufficiently complex secondary structure (Chapter 2). Initial bioinformatics analyses of selenoprotein genes focused on SECIS elements. In these studies, selenoprotein genes were identified using the following strategy: (1) detection of SECIS elements by searching for conserved stem-loop structures satisfying SECIS consensus sequence and structure; (2) analyzing regions upstream of SECIS elements for coding regions of selenoprotein genes; and (3) computational and experimental analyses of candidate selenoproteins (Kryukov et al. 1999; Lescure et al. 1999; Castellano et al. 2001). This strategy immediately resulted in the identification of several novel selenoproteins. Subsequently, it was applied to entire genomes, identifying full sets of selenoproteins (selenoproteomes) in a variety of organisms. For large and complex genomes, searches were carried with pairs of closely related genomes (e.g., *D. melanogaster* and *D. pseudoobscura*, or human and mouse genomes) by detecting conserved pairs of SECIS elements located upstream of a pair of selenoprotein orthologs (Kryukov et al. 2003). In particular, this strategy was useful in the analysis of the human genome: these analyses were assisted by the availability of mouse and rat sequenced genomes.

A second strategy was also developed wherein selenoproteins can be identified by searching for cysteine (Cys) homologs (Kryukov et al. 2003, 2004; Fomenko et al. 2007). This strategy is based on the observation that most selenoprotein genes have homologs, in which Cys replaces Sec. Thus, protein sequence databases (e.g., NCBI protein database, and ORFs from genome and environmental genome projects) were searched against large nucleotide sequence databases (genomes, ESTs, metagenomics projects, etc.) to identify nucleotide sequences containing an in-frame TGA codon, which, when translated, aligned with Cys-containing protein homologs such that the resulting Sec/Cys pairs were flanked by conserved sequences. It should be noted that such Cys/Sec homology strategy is completely independent of the searches for SECIS elements and thus provided a SECIS-independent tool for selenoprotein detection. In addition, since both strategies (i.e., SECIS based and Sec/Cys pair based) identified identical or nearly identical sets of selenoprotein genes in various genomes, both tools should be viewed as satisfactory and complementary for selenoprotein analyses in sequence databases. Moreover, this observation suggested that the two procedures can identify nearly all or all selenoproteins in sequence databases as well as in completely sequenced genomes.

1.4 Selenoproteins

While the first selenoproteins were discovered in 1973, until recently only a handful of such proteins were known. In fact, the majority of known selenoproteins have been discovered within the last 6 years. Currently, more than 50 selenoprotein families are known (Fig. 1.4). Our laboratories have described many of these proteins and

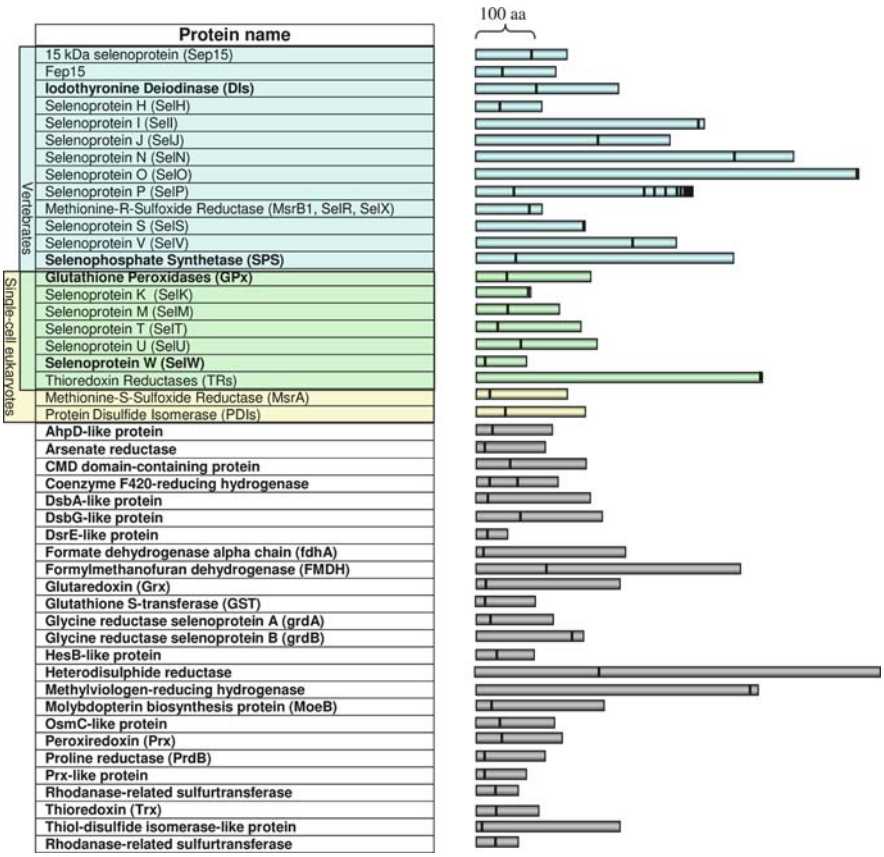

Fig. 1.4 Selenoprotein families. Selenoproteins in vertebrate and single-celled eukaryotes are highlighted, and those selenoproteins shown in bold are also present in bacteria. Other selenoproteins (*lower part of the figure*) are prokaryotic. On the *right side* of the figure, relative lengths of selenoproteins and location of Sec are shown

the reader is referred to the corresponding primary literature (Martin-Romero et al. 2001; Kryukov et al. 2003; Lobanov et al. 2006b, c, 2007; Mix et al. 2007; Lobanov et al. 2008a, b). As several detailed reviews covering selenoproteins and selenoprotein functions have been published recently (Gromer et al. 2005; Schweizer and Schomburg 2005; Hatfield et al. 2006; Holmgren 2006; Moghadaszadeh and Beggs 2006; Papp et al. 2007; Brigelius-Flohe 2008; Gromadzinska et al. 2008; Margis et al. 2008; Schweizer et al. 2008), we do not cover individual selenoproteins here.

1.4.1 Overview of Selenoprotein Functions

Those selenoproteins for which functions have been established are oxidoreductases with Sec located in catalytic sites and serving redox function (Kryukov et al. 2004; Zhang and Gladyshev 2008). By analogy, it may be predicted that many