Finding a needle in a haystack: The Eukaryotic Selenoproteome

Charles E. Chapple

PhD thesis

Barcelona, 2008

UNIVERSITAT POMPEU FABRA
Departament de Ciències Experimentals i de la Salut

The eukaryotic selenoproteome

Charles E. Chapple

Memòria presentada per optar al grau de Doctor
en Biologia per la Universitat Pompeu Fabra.

Aquesta Tesi Doctoral ha estat realitzada sota la direcció del
Dr. Roderic Guigó Serra al Departament de Ciències Experimentals
i de la Salut de la Universitat Pompeu Fabra

Roderic Guigó Serra **Charles E. Chapple**

Barcelona, 2009

The research in this thesis has been carried out at the Genome Bioinformatics Lab (GBL) within the Grup de Recerca en Informtica Biomdica (GRIB) at the Parc de Recerca Biomèdica de Barcelona (PRBB), a consortium of the Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra (UPF) and Centre de Regulació Genòmica (CRG).

The research carried out in this thesis has been supported by a predoctoral fellowship from the Spanish Ministry of Education and Science (Ministerio de Educación y Ciencia) to C.E. Chapple and grants from Ministerio de Ciencia y Tecnología to R. Guigó.

To my parents for giving me the chance,
To miju, who did not let me waste it...

Contents

List of Figures

About this document

Tips on reading this thesis

Some hints on the format used in this document.

On the text

Footnotes those comments not needed for following the main flow of the text, and that may interrupt it, are placed in an *ad hoc*[1] footnote section; and

Gene names are written in lowercase *italics* (eg. *sps2*); and

Protein names are written in a standard face and, if appropriate, in uppercase (eg. SPS2).

On the main chapters

Colors : Each of the basic chapters has a different color for headings and table of contents entries. This is to facilitate browsing the document.

Introduction
Results
Methods
Discussion

Results : includes the original research papers; and

Methods : contains additional information on the methods and software used during this PhD.

[1] For the particular purpose.

On the appendices section

Abbreviations : a list of the abbreviations used in the text and their meanings.

Glossary : Certain terms used throughout the text are explained. Additionally, some of the proteins and other factors commonly referred to in the text are described.

List of publications : a complete list of published articles and posters of which I am an author.

On the PDF file

Table of contents entries are links to the corresponding page.

References appear in blue and are clickable links which will take you to the appropriate reference. For example: (Crick, 1968)

URLs in red are linked to the corresponding electronic resource and will be opened in your web browser. For example: `http://www.google.com`

Abstract

Selenoproteins are a diverse family of proteins containing the trace element Selenium (Se) in the form of the non-canonical amino acid selenocysteine (Sec). Selenocysteine, the 21st amino acid, is similar to cysteine (Cys) but with Se replacing Sulphur. In many cases the homologous gene of a known selenoprotein is present with cysteine in the place of Sec in a different genome. Selenoproteins are believed to be the effectors of the biological functions of Selenium and have been implicated in male infertility, cancer and heart diseases, viral expression and ageing.

Selenocysteine is coded by the opal STOP codon (TGA). A number of factors combine to achieve the co-translational recoding of TGA to Sec. The 3′ **U**n**t**ranslated **r**egions (UTRs) of eukaryotic selenoprotein transcripts contain a stem-loop structure called a **Sec I**nsertion **S**equence (SECIS) element. This is recognised by the **S**ecis **B**inding **P**rotein 2 (SBP2), which binds to both the SECIS element and the ribosome. SBP2, in turn, recruits the Sec-specific Elongation Factor EFsec, and the selenocysteine transfer RNA, $tRNA^{Sec}$.

The dual meaning of the TGA codon means that selenoprotein genes are often mispredicted by the standard annotation pipelines. The correct prediction of these genes, therefore, requires the development of specific methods.

In the past few years we have contributed significally to the description of the eukaryotic selenoproteome[2] with the discovery of novel families (Castellano et al., 2005), the elaboration of novel methods (Taskov et al., 2005; Chapple et al., 2009) and the annotation of different genomes (Jaillon et al., 2004; Drosophila 12 genomes Consortium, 2007; Bovine Genome Sequencing and Analysis Consortium, 2009). Finally, and perhaps most importantly, we have identified the first animal to lack selenoprotein genes (Drosophila 12 genomes Consortium, 2007; Chapple and Guigó, 2008). This last finding is particularly surprising because it had previously been believed that selenoproteins were essential for animal life.

[2]The set of selenoproteins in a given organism.

Resum

Les selenoproteïnes constitueïxen una família diversa de proteïnes, caracteritzada per la presència del Seleni (Se), en forma de l'amino àcid atípic, la selenocisteïna (Sec). La selenocisteïna, coneguda com l'amino àcid 21, és similar a la cisteïna (Cys) per amb un àtom de seleni en lloc de sofre (S). Les selenoproteïnes són els responsables majoritaris dels efectes biolgics del seleni i s'ha observat que poden estar implicades en la infertilitat masculina, el càncer, algunes malalties coronàries, l'activació de virus latents i l'envelliment.

La selenocisteïna es codifica pel codó UGA, normalment codó de parada (STOP). Per a la recodificació correcta del UGA són necessaris diversos factors. A la part 3' de la regiò no traduïta (UTR) dels transcrits dels gens de selenoproteïnes en organismes eucariotes s'hi troba una estructura de "stem-loop" anomenada SECIS. La proteïna SBP2 interactua amb el SECIS, així com amb el ribosoma, i forma un complex amb el factor d'elongació EFsec i el tRNA de la selenocisteïna, el $tRNA^{Sec}$.

Donat que el codó TGA normalment significa fi de la traducció, les formes tradicionals de búsqueda de gens no el reconeixen com a codó codificant. Per aquesta raó ha estat necessari desenvolupar una metodologia específica per a la predicció de gens de selenoproteïnes.

En els últims anys, hem contribuït a la descripció del selenoproteoma eucariota amb el descubriment de noves famílies (Castellano et al., 2005), amb l'elaboració de nous mètodes (Taskov et al., 2005; Chapple et al., 2009) i l'anotació de diferents genomes (Jaillon et al., 2004; Drosophila 12 genomes Consortium, 2007; Bovine Genome Sequencing and Analysis Consortium, 2009). Finalment, hem identificat el primer animal que no té selenoproteïnes (Drosophila 12 genomes Consortium, 2007; Chapple and Guigó, 2008), un descubriment soprenent donat que, fins el moment, es creia que les selenoproteïnes eren essencials per la vida animal.

CHAPTER 1

Introduction

Summary

In this chapter I introduce the field of selenoproteins, and the basic biological concepts needed to understand the research presented here. I present what is known of the factors necessary for selenoprotein expression as well as a brief history of Selenium and its biology. The genetic code and its evolution are discussed. Finally, I also touch on eukaryotic genes and their prediction and RNA structures and their prediction.

Contents

Overview

SELENOPROTEINS are a diverse group of proteins set apart by their incorporation of the amino acid selenocysteine. Selenocysteine is known as the 21st amino acid and it is designated as Sec or U in the three and one-letter codes, respectively. It is coded for by the opal STOP codon (UGA) which is cotranslationally recoded in the presence of a stem-loop structure on the 3′ untranslated region (UTR) of selenoprotein gene transcripts (the SECIS element). This recoding depends on a set of both *trans* and *cis* factors which will be described in detail further on.

Selenoproteins have been found in all three domains of life (Eukarya, Bacteria and Archaea) but show a scattered phylogenetic distribution. With the data available today it is still far from clear just what path has led to the evolution of the selenocysteine coding trait. In this chapter I will give an overview of the concepts necessary to understand the work presented here.

1.1 The Genetic code

Nearly 45 years ago, in 1961, Crick and colleagues deduced the general nature of the genetic code from the results of crosses between mutants in the rIIB cistron of the T4 bacteriophage (Crick et al., 1961). They found that the genetic code is a language with three letter words, the codons. That is, the genetic code is read in triplets, each of which defines one amino acid (words) or the end of translation (STOP, punctuation, if you will). This seminal paper showed that (quoted directly from Crick et al. (1961)):

- A group of three bases (or, less likely, a multiple of three bases) codes one amino-acid.
- The code is not of the overlapping type[1].
- The sequence of the bases is read from a fixed starting point. This determines how the long sequences of bases are to be correctly read off as triplets. There are no special "commas" to show how to select the right triplets. If the starting point is displaced by one base, then the reading of the triplets is displaced, and thus becomes incorrect.
- The code is probably "degenerate"; that is, in general, one particular amino-acid can be coded by one of several triplets of bases.

Ironically, the authors veer away from the notion of "nonsense" codons which had already been proposed by Crick and colleagues in an earlier paper (Crick et al., 1957). In this paper Crick and colleagues propose a solution to the "coding problem" by assuming that some of the triplets are "sense" and some "nonsense". Although it later turned out that this assumption was wrong, we have kept the term "nonsense" to refer to STOP codons, of which we will be hearing more later on.

[1]The authors are careful to note that the non-overlapping nature of the code comes not from their work but from that of Wittman and of Tsugita and Fraenkel-Conrat (Tsugita and Fraenkel-Conrat, 1962)

1.1.1 The standard code

This newly discovered genetic code was, in a fit of hubris, given the grandiose name of "The universal genetic code". Although a reasonable assumption at the time, given that it was found shared between organisms as distant to each other as yeast, vertebrates and the tobacco mosaic virus, we now know that it is but one of the many genetic codes present in nature. This code is now referred to as the "Standard code". The standard code is shown in figure 1.1

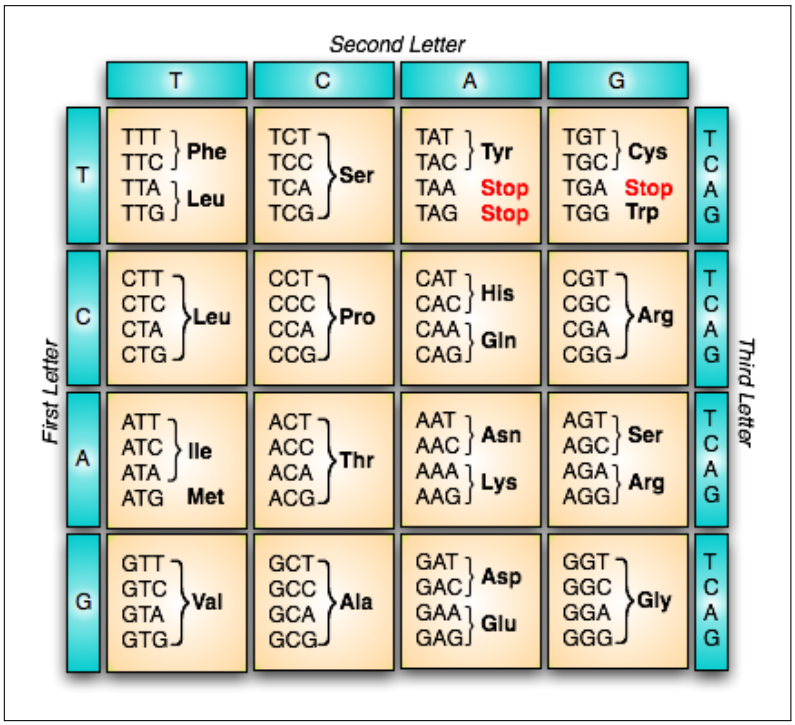

Figure 1.1: **The standard genetic code**. Note the three standard stop codons in red. Figure from http://plato.stanford.edu/entries/information-biological/.

1.1.2 The nonstandard codes

Barrel et al were first to challenge the universality of the standard code in 1979 (Barrell et al., 1979). They showed that human mitochondria have a different code, with UGA coding for tryptophan, AUA for methionine instead of isoleucine and AGA and AGG as

terminators instead of coding for arginine. AUU codes for isoleucine during elongation but can code for methionine for initiation.

Variant genetic codes can be defined as minor, or partial, and major, or complete. Partial changes are those that affect specific codons under specific circumstances. An example of a partial deviation from the standard code is the recoding of UGA to code for selenocysteine. Major changes are those that change the standard meaning of a codon in the genome in question. An example of a major change is the mitochondrial code described above.

Variations from the standard code often involve stop codon reassignment[2] (see Table 1.1). For example, in the yeast and invertebrate mitochondrial code as well as in molds, protozoans mycosplasmata kinetoplasts echinoderms and the flatworm, UGA codes for tryptophan instead of STOP. In the ciliates, UAA and UAG both code for glutamine instead of STOP. A complete listing of alternate genetic codes can be found at the ncbi web page (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c)

Codon (Standard Meaning)	Deviations
UGA (STOP)	Trp (Codes 2,3,4,5,9,13,14,21) Cys (Code 10)
UAG (STOP)	Leu (Codes 16,22) Gln (Codes 6,10)
UAA (STOP)	Gln (Code 6) Tyr (Code 14)
CUG (Leu)	Thr (Code 3) Ser (Code 12)
AGG (Arg)	STOP (Code 2) Ser (Codes 5,9,14,21) Gly (Code 13)
AUA (Ile)	Met (Code 21)
AGA (Arg)	STOP (Code 2) Ser (Codes 5,9,14,21) Gly (Code 13)
UCA (Ser)	STOP (Code 22)
CGA (Arg)	absent (Code 3)
UUA (Leu)	STOP (Code 23)
CUU (Leu)	Thr (Code 3)
CUC (Leu)	Thr (Code 3)
CGC (Arg)	absent (Code 3)
CUA (Leu)	Thr (Code 3)
AAA (Lys)	Asn (Code 9,14)
AUA (Ile)	Met (Code 2,3,5,13)

Table 1.1: **Major Variations from the standard genetic code.** See Table 1.2 for a listing of the different genetic codes. Compiled from the information available on the ncbi page http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c

[2]They are, however, by no means restricted to stop codon reassignment. Sense codons are also changed in some species, but STOP codons are most often changed.

Genetic Codes	
Code 2	Vertebrate Mitochondrial Code
Code 3	Yeast Mitochondrial Code
Code 4	Mold, Protozoan, and Coelenterate Mitochondrial Codes and the Mycoplasma/Spiroplasma Code
Code 5	Invertebrate Mitochondrial Code
Code 6	Ciliate, Dasycladacean and Hexamita Nuclear Code
Code 9	Echinoderm and Flatworm Mitochondrial Code
Code 10	Euplotid Nuclear Code
Code 12	Alternative Yeast Nuclear Code
Code 13	Ascidian Mitochondrial Code
Code 14	Alternative Flatworm Mitochondrial Code
Code 10	Blepharisma Nuclear Code
Code 16	Chlorophycean Mitochondrial Code
Code 21	Trematode Mitochondrial Code
Code 22	Scenedesmus obliquus mitochondrial Code
Code 23	Thraustochytrium Mitochondrial Code

Table 1.2: **The genetic codes referred to in Table 1.1 .** Compiled from the information available on the ncbi page `http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c`

Of the partial code variations we will deal extensively with the recoding of UGA to Sec later. A similar recoding occurs in the case of pyrrolysine (Pyl), the 22nd amino acid. Unlike selenocysteine which is found in all three domains of life, pyrrolysine appears limited to the Methanosarcinacea and the Gram-positive *Desulfitobacterium hafniensea*. In these organisms the amber codon UAG is recoded to incorporate this residue in monomethylamine methyltransferases. Another major difference with selenocysteine is that where Sec is formed directly on its cognate tRNA, and is never a free metabolite, pyrrolysine follows the lead of the common set of amino acids. So, pyrrolysine is the 21st amino acid from nature that is charged directly onto a dedicated tRNA by a cognate aminoacyl-tRNA synthetase. For a recent review on the genetic encoding of pyrrolysine please see Krzycki (2005).

Other cases of STOP codon readthrough have been observed, for example the *kelch, oaf* and *hdc* genes in *Drosophila melanogaster*. Such cases involve naturally occurring suppressor tRNAs, tRNAs that can recognize STOP codons in addition to their cognate sense codons. For a review of suppressor tRNAs in eukaryotes see (Beier and Grimm, 2001).

1.1.3 Evolution of the code

There are three major theories to explain the origin of the genetic code. Here, I will briefly outline each of them. For more detail on the origin and evolution of the genetic code please see Giulio (2005); Koonin and Novozhilov (2009)

The stereochemical theory (Crick, 1968) claims that the origin of the genetic code can be traced to the stereochemical interactions between codons or anticodons and their

amino acids. If true, this model implies that the genetic code is not accidental, as it is an emergent quality of the system itself.

The adaptive theory (Epstein, 1966; Woese, 1965) postulates that the structure of the genetic code is such as to maximize robustness, that is, to minimize the effect that errors would have on the code's function.

The coevolution theory (Wong, 1975) Proposes that the structure of the genetic code was determined by the sequence of evolutionary emergence of new amino acids within the primordial biochemical system.

There are also three major theories to explain the evolution of the genetic code:

The ambiguous intermediate theory posits that codon reassignment occurs through an intermediate stage where a particular codon is ambiguously decoded by both the cognate tRNA and a mutant tRNA. This could be followed by the eventual deletion of the original tRNA ad the subsequent takeover of the codon by the mutant.

The codon capture theory (Osawa et al., 1992) suggests that under evolutionary pressure to decrease genomic GC-content, GC rich codons could disappear from a genome. Then, because of random genetic drift, these codons would reappear and be reassigned to another amino acid.

The genome streamlining theory (Andersson and Kurland, 1995) hypothesis states that selective pressure to minimize mitochondrial genomes yields reassignments of specific codons, in particular, one of the three stop codons.

None of these theories are mutually exclusive, in fact I expect that the "truth" lies in a combination of many if not all of them acting on the genome over evolutionary time. Probably different forces took precedence at different times in evolution.

1.2 Selenium

Selenium(Se), is a chemical element with the atomic number 34, represented by the chemical symbol Se, and with an atomic mass of 78.96. It takes its name from Selene ($\Sigma\epsilon\lambda\eta\nu\eta$), the Greek name for the moon[3]. Pure selenium rarely occurs in nature, but when it does it can be in several different forms, the most stable of which is a semi-metal (semiconductor) form which is used in photocells. Chemically, selenium is related to sulphur and tellurium. It was first discovered by Jöns Jakob Berzelius in 1817 (Birringer et al., 2002). Although it had long been known to cause disease in livestock and humans its beneficial effects were unknown until the 1950s when it was reclassified as an essential trace element.

[3]Hence the artistic font on the cover...

Figure 1.2: **Selenium hunk**. From http://www.periodictable.com/Items/034.12/index.html.

Selenium and human health

Selenium is an essential trace nutrient which has been shown to have various beneficial effects on human health. It is needed for the proper function of the immune system, for sperm motility, and has been shown to inhibit HIV progression to AIDS. Through its presence in selenoproteins Selenium acts as an antioxidant (e.g. SelW) and is necessary for the production of active thyroid hormone (e.g. TR1). Selenium also has also been shown to reduce cancer risk and cardiovascular disease. Conversely, higher concentrations of Selenium are toxic and can give rise to a condition known as selenosis. So, in Selenium as in so many things in life, moderation is essential. An excellent review of Selenium and its role in human health can be found in (Rayman, 2000).

Specifically, Selenium has been implicated in a variety of human diseases and disorders. Selenium deficiency has been associated with Keshan disease, an endemic cardiomyopathy, and Kashin-Beck disease, a deforming arthritis and has been connected to loss of immunocompetence (Rayman, 2000). Conversely, if not surprisingly, Selenium supplementation has immunostimulant effects (Rayman, 2000).

Selenium deficiency has also been linked to occurrence, virulence, or disease progression of some viral infections (Beck et al., 2003; Rayman, 2000) including HIV.[4] Indeed evidence exists that some viruses encode selenoproteins themselves. The fowlpox virus has been shown to contain a selenoprotein homolog of glutathione peroxidase 4 (Mix et al., 2007) as have others, including potentially serious human pathogens like HIV-1 and hepatitis C virus, coxsackievirus B3, HIV-2, and the measles virus (Zhang et al., 1999).

Ironically, most Selenium in our diet comes from plants, the only eukaryotic kingdom to (so far) lack selenoproteins[5]. However, various meat, fish and dairy products also contain

[4]The exact connection between Selenium and HIV is, however, controversial (e.g. Passaretti and Gupta (2007); Hurwitz et al. (2007); Dillon and Stapleton (2007)).

[5]With the exception of certain green algae (Obata and Shiraiwa, 2005; Lobanov et al., 2006)

selenium. A comprehensive review of Selenium sources in the human diet can be found in (Navarro-Alarcon and Cabrera-Vique, 2008). Selenium content in soil varies significantly from region to region (see Figure 1.3). The Selenium content of plants depends on that of the soil in which they were grown. In an interesting historical anecdote, the forces sent to reinforce the troops at the Alamo were delayed by (among other things) Selenium poisoning of the horses induced by the excess Selenium in the plants on the way.[6] Finally, if any doubt remains as to the importance of Selenium, "Selenium may have been the main cause for the extinction of dinosaurs and other animal species at the end of the Mesozoic era" (Koch, 1967).

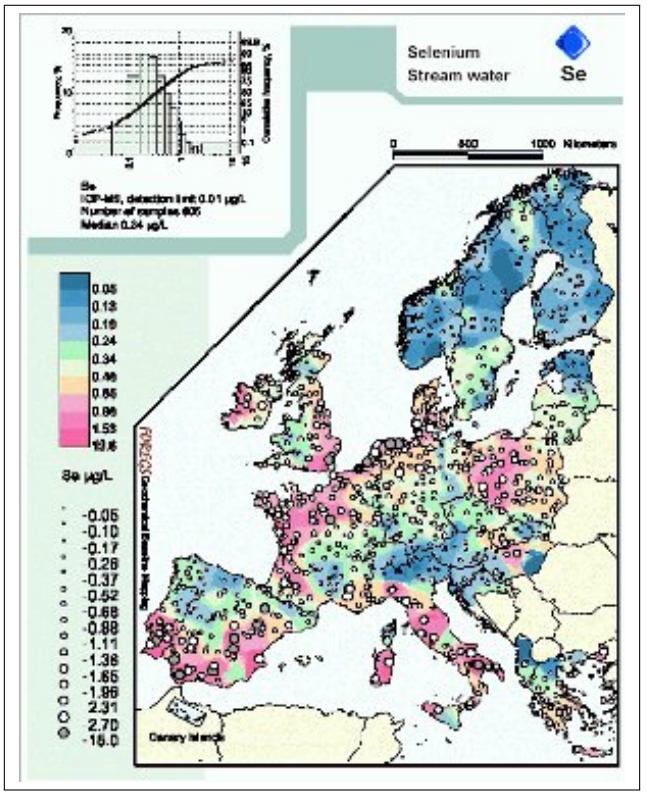

Figure 1.3: Concentration of Selenium in stream waters of 26 European countries
Source: Analysis: BGR, Hannover; Map: Geological Survey of Finland (GTK), Helsinki
Taken from http://www.bgr.bund.de/cln_092/nn_324514/EN/Themen/Wasser/Bilder/Was__foregs__projektbeschr__abb2__g__en.html

[6] I heard this at the 2006 Selenium meeting and although I remember a reference given, I have not been able to find one.

1.3 Selenoproteins

Selenoproteins are a diverse group of proteins characterized by the presence of the 21st amino acid, selenocysteine (Sec). Selenium can also be post-translationally bound to some proteins but these are not considered selenoproteins. Like all amino acids, Sec is cotranslationally inserted into the growing polypeptide chain. Unlike other amino acids however, Sec needs an array of cis and trans acting factors to enable and direct its correct incorporation.

The first selenoproteins to be discovered were protein A of the glycine reductase system (Turner and Stadtman, 1973) and formate dehydrogenase (Andreesen and Ljungdahl, 1973), both from *Clostridium sp.* and a mammalian enzyme, glutathione peroxidase (GPx) (Flohe et al., 1973; Rotruck et al., 1973).

1.3.1 Selenocysteine and selenoproteins

Selenocysteine, the 21st amino acid, is a cysteine analog with Selenium replacing Sulfur (see Figure 1.4). Selenocysteine was first identified as a non-canonical amino acid in the bacterial protein formate dehydrogenase (Cone et al., 1976) and the mammalian GPx (Forstrom et al., 1978). However, the mechanism of selenocysteine insertion was still unknown.

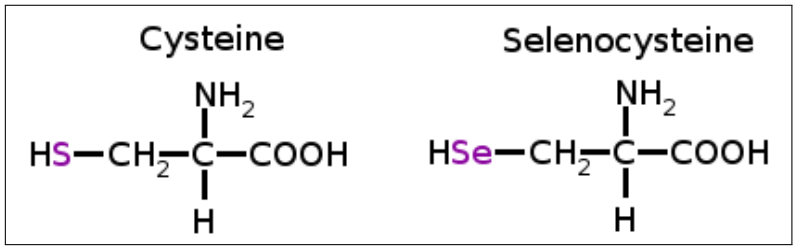

Figure 1.4: **Cysteine and selenocysteine**.

In 1986 Chambers and colleagues (Chambers et al., 1986) isolated and sequenced a cDNA encoding mouse GPx2 and showed that the selenocysteine residue was encoded by a TGA codon. In the same year, TGA was shown to direct selenocysteine incorporation in formate dehydrogenase (Zinoni et al., 1986).

1.3.2 Eukaryortic selenoprotein families

The following is a list of known eukaryotic selenoprotein families and what is known of their function. Alternate names are given in parentheses. An excellent review of the human selenoproteins and the source of much of the information below is (Gromer et al., 2003).

15kDa (Sep15): This protein is directed to the endoplasmatic reticulum (ER), where it tightly binds UDP-glucose glycoprotein glucosyltransferase, a protein whose function is quality control of protein folding.

Deiodinases : These proteins catalyze activation or inactivation (or both) of thyroid hormones (T3 and T4).

DI1 (DIO1): Homodimeric plasma membrane protein that primarily deiodinates the 5′-position of the phenolic ring of T4.

DI2 (DIO2): Its primary function is the conversion of T4 into T3 in specific target tissues.

DI3 (DIO3): Deiodinates the 5-position of the tyrosyl ring, thereby inactivating T3 and T4.

Fep15 : Absent in mammals, it can be detected only in fish and is present in these organisms only in the selenoprotein form (Novoselov et al., 2006).

Glutathione peroxidases : Glutathione peroxidases reduce and thereby detoxify different types of peroxides to their respective alcohols at the expense of (typically) glutathione.

- **GPx1** (cGPx) : Ubiquitous homotetrameric cytosolic enzyme.
- **GPx2** (GI-GPX): Found in the liver and within the gastrointestinal system (but absent in heart and kidney).
- **GPx3** (p-GPx) : The physiological function of this homotetrameric glycoprotein is not convincingly resolved.
- **GPx4** (ph-GPx): Exhibits the broadest substrate specificity of all glutathione peroxidases and can even reduce phospholipid hydroperoxides.
- **GPx6** [7]: Believed to have a function in olfaction.

SelH : Contains a CXXU motif (redox box) suggestive of redox function.

SelI : Possibly an integral membrane protein.

SelJ : SelJ (as a sec-containing protein) is specific to actinopterygian fishes and sea urchin and has a suggested structural role (Castellano et al., 2005). On a personal note, this selenoprotein gave me my first moment of scientific discovery since it was I who first found and named it when analyzing the genome of *Tetraodon nigroviridis* (Jaillon et al., 2004).

[7] GPx5 GPx7 and GPx8 are not selenoproteins.

SelK : Membrane protein, possibly involved in redox control.

SelL : Restricted to marine organisms, this protein may have a redox function (Shchedrina et al., 2007).

SelM : Contains a CXXU motif, suggestive of redox function. Also characterized (in human) by an atypical SECIS element, with Cytosines replacing the conserved Adenosines on the apical loop.

MsrA : This protein is a methionine sulfoxide reductase(Novoselov et al., 2002).

SelN : Is retained within the ER. Mutations in SelN are associated with various myopathies.

SelO : Characterized by an atypical SECIS with Cytosines replacing the conserved Adenosines of the apical loop in all species investigated. Contains a CXXU motif, suggestive of redox function.

SelP : Is the major selenoprotein in plasma, accounting for >50% of total plasma selenium. Is the only known selenoprotein with >1 selenocysteine residue (apart from some splice variants of SelN which have 2), with 10 Sec residues in the human homolog. It is believed to act as a Selenium transporter.

SelR (SelX, methionine-R-sulfoxide reductase 1, MsrB1): cytosolic and nucleic protein with a role in the protection of the cell from oxidative stress.

SelS (Tanis): Apparently involved in hepatic glucose metabolism and retrotranslocation of ER proteins.

Selenophosphate synthetase (SPS2): Catalyses the formation of mono selenophosphate (SeP) from selenide and ATP.

SelT : Contains a CXXU motif, suggestive of redox function.

Thioredoxin reductases : these proteins reduce oxidized thioredoxin (Trx(SH_2)) at the expenses of NADPH.

- **TR1** (TrxR1): Ubiquitous cytoplasmatic housekeeping enzyme involved in many aspects of redox regulation.
- **TR2** (TrxR2): Is located in mitochondria.
- **TR3** (TGR): Testis-specific enzyme located in the ER. Unlike TR1 and TR2, it can reduce glutathione disulfide.

SelU : This selenoprotein of unknown function was discovered by our group (Castellano et al., 2004).

SelV : Contains a CXXU motif, suggestive of redox function.

SelW : Homologous to SelV, this small protein is highly expressed in muscle and may have an antioxidant role.

In addition to the eukaryotic proteins listed above which are found in various orders, a few order or species-specific selenoproteins have also been identified:

SelTryp : Protist-specific selenoprotein is of unknown function. It was first found in *Trypanosoma brucei* (Lobanov et al., 2006) but has since been identified in other protists as well (C. E. Chapple, unpublished data).

Novel selenoproteins identified in *O. tauri* (Lobanov et al., 2007):

MSP : Membrane selenoprotein of unknown function.

Hypothetical protein 1,2 and 3

Homologs to known bacterial selenoproteins identified in *O. tauri* :

Methyltransferase

Peroxiredoxin

Thioredoxin-fold protein

EhSep1,2,3 (PDI): *Emiliana huxleii* specific selenoproteins. EhSep2 seems to be a protein disulfide isomerase (PDI) (Obata and Shiraiwa, 2005).

No selenoproteins and no $tRNA^{Sec}$ have been found in higher plants although a few have been identified in other green algae and diatoms (Novoselov et al., 2002; Obata and Shiraiwa, 2005; Shrimali et al., 2005; Lobanov et al., 2006).

1.3.3 Selenoprotein distribution

At the dawn of the selenoprotein era, it was a commonly held belief that the number of selenoproteins encoded in a given genome was proportional to its complexity. That is, the more complex an organisms the more selenoproteins its genome will encode. This view, although supported by the data at the time, has now been shown to be wrong. The largest published eukaryotic selenoproteome is that of the green alga *Ostreococcus lucimarinus* (Lobanov et al., 2007) with 29 selenoproteins.

1.3.4 Selenocysteine incorporation into eukaryotic proteins

The mechanism of selenocysteine incorporation into proteins was first understood in prokaryotes. Zinoni *et al* showed that Sec incorporation was dependent on a stem-loop structure found immediately downstream of the in-frame UGA of *Escherichia coli* formate dehydrogenase H (Zinoni et al., 1986). During the next few years the mechanism of Selenium incorporation into bacterial proteins was determined by the groups of Stadtman and Böck (For a review see Böck et al. (2006)).

They showed that this stem-loop structure (hereafter referred to as the **Sec I**nsertion **S**equence, SECIS, or SECIS element)[8] is recognised and bound by the specialized translation factor SelB which also binds $tRNA^{Sec}$ and directs it to the ribosome.

[8]The term SECIS was originally coined by Berry and coworkers (Berry et al., 1991) for eukaryotic SECIS elements but has been extended to the prokaryotic SECISes

As is so often the case, the situation in eukaryotes is more complex. The roles of SelB are split between two proteins, SBP2 which binds the SECIS element, the ribosome and the specialized elongation factor EFsec which binds tRNA^{Sec}. In the following sections I will give a more detailed description of these and other factors necessary for selenoprotein biosynthesis in eukaryotes.

1.3.5 Selenoprotein mRNA translation

Apart from the in-frame UGA codon, selenoprotein mRNA translation is no different to that of any standard protein. The ribosome moves stepwise along the mRNA chain decoding codons and elongating the polypeptide chain and, only when the translational machinery meets an in-frame UGA codon, does the specialized decoding apparatus for selenocysteine insertion come into play. For a review of selenoprotein mRNA translation see Allmang and Krol (2006).

Currently, there are two models of eukaryotic selenoprotein mRNA translation. Following their observation that SBP2 cannot simultaneously bind both the ribosome and the SECIS element, Copeland and coworkers (Kinzy et al., 2005) propose that a subset of ribosomes with prebound SBP2 are somehow selected for selenoprotein translation. The interaction of the ribosome-bound SBP2 with the SECIS element produces a conformational change in the ribosomal A site, allowing delivery of the EFsec/Sec-tRNA^{Sec}. Ribosomal protein L30 would then displace SBP2 from the SECIS RNA to relocate it to its original position in the ribosome (see Figure 1.5 A).

The second model, by Driscoll and coworkers (Chavatte et al., 2005), has SBP2 binding the SECIS and subsequently serving as a platform to recruit the EFsec/Sec-tRNA^{Sec} complex. During translation, the approach of the ribosome will lead L30 to displace SBP2, inducing a conformational change that triggers the release of the Sec-tRNA^{Sec} and GTP hydrolysis (see Figure 1.5 B).

1.3.6 Necessary factors for eukaryotic selenoprotein biosynthesis

In eukaryotes, a number of both cis- and trans- acting factors [9] are needed for the correct production of selenocysteine and its incorporation into selenoproteins. Each of these will be addressed briefly in the following sections. For a recent review of the molecular partners involved in selenoprotein biosynthesis please see Allmang et al. (2009).

- **tRNA**^{Sec} : The specific tRNA for Selenocysteine.
- **PSTK** : Phosphoseryl tRNA kinase, phosphorylates the Ser-tRNA^{Sec}.
- **SecS** : Eukaryotic selenocysteine Synthetase (previously SLA/LP), converts Ser-tRNA^{Sec} to Sec-tRNA^{Sec}.
- **SPS1 and SPS2** : Selenophosphate Synthetases

[9] I use the term factor here in a very general sense to mean any protein or nucleotide element or signal

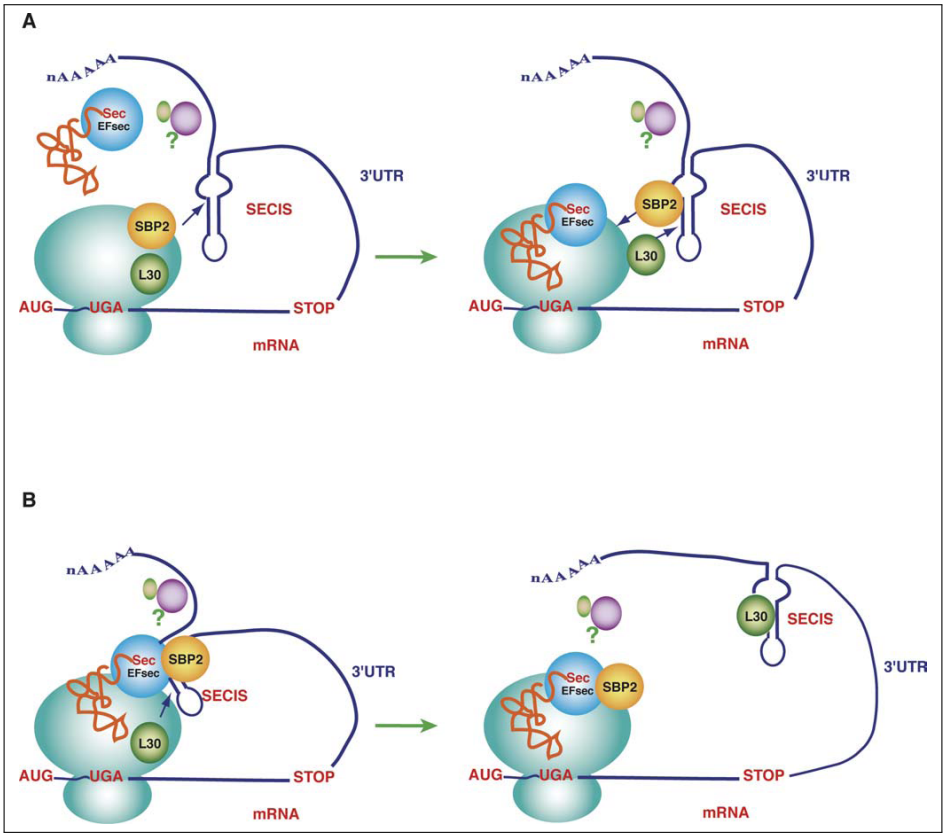

Figure 1.5: (A) SBP2 travels with ribosomes, interacts with the SECIS RNA and the EFsec/Sec-tRNA^{Sec} to deliver this complex to the A site of the ribosome. L30 displaces the SECIS-bound SBP2. (B) The EFsec/Sec-tRNA^{Sec} complex is recruited at the SECIS RNA by SBP2. Ribosome-bound L30 displaces SBP2. In both models, L30 must leave the SECIS RNA to reset the system. Black arrows indicate factor reshuffling; as yet unidentified factors, possibly involved in the mechanism, are indicated with a question mark. Image taken from (Allmang and Krol, 2006)

- **secp43** : Forms part of the SBP2/tRNA^{Sec}/EFsec complex but its exact role is unclear.
- **SBP2** : **S**ecis **B**inding **P**rotein 2, binds the SECIS element and the ribosome.
- **EFsec** : Elongation factor specific for selenocysteine.
- **Ribosomal protein L30** : A component of the ribosome which has also been shown to bind the SECIS element.
- **SECIS** : **S**elenocysteine **I**nsertion **S**equence, a stem-loop structure on the 3′ UTR of selenoprotein mRNAs.
- **The UGA codon** : Signal specifying selenocysteine insertion into the growing polypeptide chain.

Transfer RNA: tRNA^{Sec}

tRNA^{Sec} is the specific tRNA that provides Sec for selenoprotein mRNA translation. A number of features unique to this tRNA have been identified. tRNA^{Sec} exists in two isoforms in mammals. One seems to be responsible for the synthesis of selenoproteins with housekeeping functions. The other, which differs by only a single methyl group in position 23 (designated Um34) appears to be responsible for the synthesis of stress-related selenoproteins that are less dependent on Selenium concentration (Carlson et al., 2006).

Although tRNA^{Sec} has the typical cloverleaf structure of canonical tRNAs, both prokaryotic and eukaryotic tRNA^{Sec}s are considerably longer and can stretch up to 100nt; (Tormay et al., 1994)). This is mostly due to the presence of a long variable arm as well as an extended acceptor stem in all selenocysteine-inserting tRNAs (see Figure 1.6).

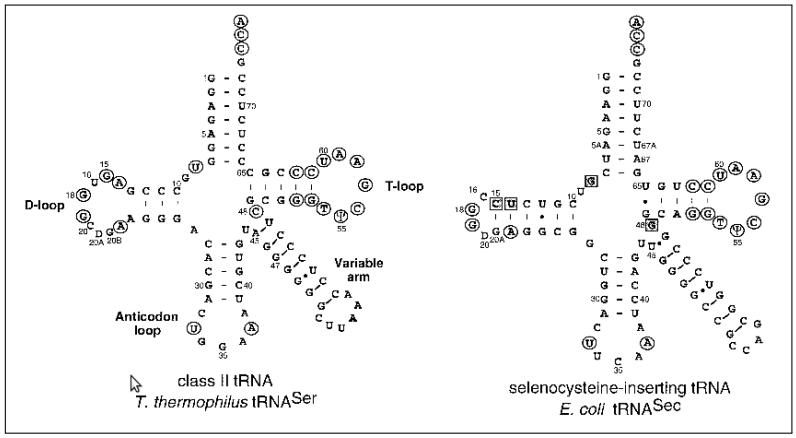

Figure 1.6: **Structures of two prokaryotic tRNAs.** Image taken from (Commans and Böck, 1999)

PSTK

Phosphso**s**eryl **t**RNA **k**inase was identified in 2004 and shown to convert seryl-tRNA to phosphoseryl-tRNA, a likely intermediate to selenocysteil-tRNA (Carlson et al., 2004). PSTK, along with EFsec, are the only proteins involved in selenoprotein biosynthesis that have been consistently found in all organisms known to code for Sec, and missing in all organisms believed to lack this trait (Chapple and Guigó, 2008). However, unlike EFsec which shares sequence similarity to EFtu, PSTK shows little conservation with other proteins, making it an easy marker for the presence of selenoprotein genes in a given genome.

SecS

SecS was originally called SLA/LP but has recently been shown (Ganichkin et al., 2008) to be the eukaryotic selenocysteine synthase. SecS requires selenophosphate and O-phosphoseryl-tRNA$^{[Ser]Sec}$ as substrates to generate selenocysteyl-tRNA$^{[Ser]Sec}$. The recently determined (Xu et al., 2007) mechanism of Sec synthesis by SecS is shown in Figure 1.7.

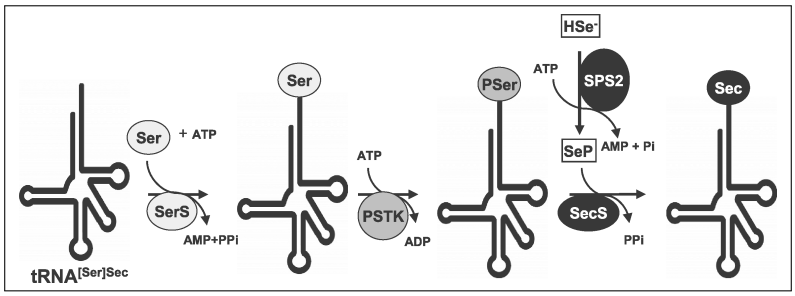

Figure 1.7: **Sec biosynthesis in eukaryotes.** Image taken from (Xu et al., 2007)

Selenophosphate synthetases, SPS1 and SPS2

SPS1 and SPS2 are the homologs of bacterial selenophosphate synthetase, SelD. Interestingly, SPS2 is itself a selenoprotein in many organisms. Although both SPS1 (Low et al., 1995) and SPS2 (Guimaraes et al., 1996) proteins have been known for quite some time, their exact roles are only now becoming clear. The authors of a recent publication (Xu et al., 2007) demonstrated that SPS2 is in fact the protein that catalyses the formation of mono selenophosphate (SeP) from selenide and ATP. The role of SPS1 in the selenoprotein biosynthesis pathway is unknown.

This corroborates finding by us and others (Chapple and Guigó, 2008; Lobanov et al., 2008) that show SPS1 to be present and highly conserved in species lacking selenoprotein genes.

secp43

The exact role of secp43 is unclear. This protein has only recently been recognised as a component of the eukaryotic selenoprotein biosynthesis machinery. It has been shown to co-exist in a complex with Sec-tRNA^{Sec}/EFsec, to interact with SPS1 *in vitro* and *in vivo* and to be involved in the redistribution of these proteins to the nucleus (Xu et al., 2005; Small-Howard et al., 2006). It also seems to be involved in selenoprotein synthesis and the

2′ methylation of the tRNASec U34 position, being thus a good Um34 methylase candidate (Small-Howard et al., 2006).

SBP2

Secis **B**inding **P**rotein 2, like bSelB, recognizes and binds to both the SECIS element and the ribosome and is essential for the correct recoding of the UGA codon. Unlike SelB however, it does not directly interact with the $tRNA^{Sec}$. SBP2 was first identified in rat in the late 1990s (Lesoon et al., 1997; Copeland and Driscoll, 1999) and later in human (Lescure et al., 2002). Various functional studies have defined three domains on the protein.

- **N-term Putative Regulatory Domain** : Amino acids 1-399[10]. The function of this domain is unknown and it is in fact dispensable for UGA recoding (Driscoll, 2006). Indeed, recent work has shown that it is not present in many (if not all) invertebrate SBP2 sequences (Chapple and Guigó, 2008; Takeuchi et al., 2009)

- **SECIS binding domain** : Amino acids 399-517 (Copeland et al., 2001). Within this domain there is an L7Ae RNA-binding module.

- **Ribosome Binding Domain** : Amino acids 470-508 (Copeland et al., 2001). The ribosome binding domain forms part of the SECIS binding domain described above.

The exact mechanism of SBP2 function is still unclear. SBP2 seems to exist in two forms, the short and the long, with the short form lacking the N-terminal domain of vertebrate SBP2s. The long form of SBP2 has so far only been identified in vertebrates whereas the short only in invertebrates (specifically in insects (Chapple and Guigó, 2008; Takeuchi et al., 2009) and in nematodes (C. E. Chapple unpublished data). In a recently published article we identified a short Lysine-rich domain (amino acids 507-534 (Takeuchi et al., 2009)) which, in the *Drosophilia melanogaster* SBP2 sequence, confers specificity for typeI SECIS elements. In addition, recent work has shown that the SBP2 L7Ae domain also allows interaction with EFsec (Donovan et al., 2008).

EFsec

Eukaryotic elongation factor EFsec (previously known as mSelB) is the selenocysteine specific elongation factor. Unlike other amino acids, Sec enjoys the distinction of having its own elongation factor which controls its insertion into the growing polypeptide chain during translation. Like *pstk*, *efsec* can be used as a marker for the presence of selenoprotein genes in a given genome (Chapple and Guigó, 2008).

[10]Numbering refers to rat SBP2 unless otherwise specified.

L30

Ribosomal protein L30 is a constituent of eukaryal and archaeal ribosomes. Like SBP2 it is a protein of the L7Ae family and has been shown to bind SECIS elements *in vivo* and *in vitro* (Chavatte et al., 2005). One of the models for selenoprotein mRNA translation has it competing with SBP2 for SECIS binding (See page 15).

The SECIS element

The **Se**leno**c**ysteine **I**nsertion **S**equence, SECIS, or SECIS element is a stem-loop structure on selenoprotein mRNAs which is necessary for the correct recoding of the UGA codon. SECIS elements were first identified by Marla Berry and collaborators in 1991 (Berry et al., 1991). They showed that a sequence in the 3′ UTR of human and rat DI1 mRNAs was required for the correct insertion of selenocysteine by the UGA codon but was dispensable for cysteine-mutants. There are two kinds of eukaryotic SECIS elements, typeI and typeII which differ in their structure (see Figure 1.8).

The structure of the typeI SECIS element was determined by Krol and coworkers in 1996 (Walczak et al., 1996). It is a stem-loop structure, with two helices (I and II) separated by an internal loop, with an apical loop surmounting helix II. Another type of SECIS element (typeII) with an additional helix III and a shorter apical loop was discovered a few years later. The naming of these elements is a historical accident and does not reflect their relative abundance in nature. In fact, typeII SECISes are far more common (Chapple et al., 2009). A plethora of studies have now given us a solid understanding of SECIS structure (e.g. Gu et al., 1997; Grundner-Culemann et al., 1999; Fagegaltier et al., 2000; Chapple et al., 2009) which is summarized in Figure 1.8.

Although the SECIS structure is conserved, there is little sequence conservation beyond the consecutive non-Watson-Crick base pairs UGAN/KGAW constituting the quartet, an unpaired A 5′ to UGAN and a run of As in the apical loop/internal loop 2 (Walczak et al., 1996; Fagegaltier et al., 2000) and of these only the UGA/GA of the quartet is invariable (e.g. Novoselov et al., 2007; Lobanov et al., 2006).

The SECIS element is found just after the UGA in prokaryotes and so is actually translated into protein. In eukaryotes, on the other hand, it is located on the 3′ UTR and can be more than 1kb downstream from the UGA codon (Berry et al., 1991).

The UGA codon

The UGA codon was first recognised as the third STOP codon by Crick and colleagues in 1967 (Brenner et al., 1967). As mentioned on page 11, UGA was shown to be the codon coding for the Sec residue in GPx2 (Chambers et al., 1986) and formate dehydrogenase (Zinoni et al., 1986). This came as a surprise as it showed that a single codon can have

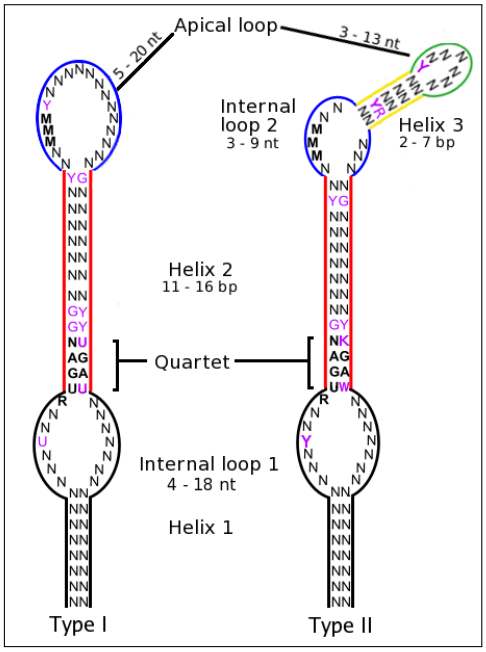

Figure 1.8: **Eukaryotic SECIS element consensus sequence.** Novel conserved residues identified in Chapple et al. (2009) are shown in magenta. Figure taken from (Chapple et al., 2009), Section , page 70. For an explanation of the ambiguity codes please see Appendix 6, page 181.

a dual (or, as it turns out, even triple[11]) meaning, STOP and Sec. Please see Table 1.1 for alternate meanings of the UGA codon.

Other protein factors

Recent work has identified nucleolin and NSEP1 as involved in selenoprotein biosynthesis but the their exact involvement in unknown. Nucleolin is involved in many and varied cellular pathways including, apparently, the formation of protein complexes on selenoprotein mRNAs during translation (Squires et al., 2007; Wu et al., 2000).

NSEP1, the **N**uclease **S**ensitive **E**lement Binding **P**rotein 1, was recently identified shown to be structurally associated with the selenoprotein translation complex and functionally involved in the translation of selenoproteins in mammalian cells (Shen et al., 2006).

1.4 Eukaryotic genes

The existence of genes was first suggested by the father of modern genetics, Gregor Mendel (1822-1884), who, in the 1860s, studied inheritance in peaplants and hypothesized a factor that conveys traits from parent to offspring. The definition of a gene has changed over the years. Where once it was thought to be a discrete nucleotide sequence that gives rise to a single protein, the discovery of non-coding RNAs, splicing, trans-splicing and

[11] A recent paper (Turanov et al., 2009) showed that in the ciliate *Euplotes crassus* UGA can mean both Cys and Sec in the same mRNA.

chimeric transcripts have led to a more permissive definition. A recent review (Gerstein et al., 2007) of our understanding of the gene concept especially in the light of the complex patterns of transcription and regulation uncovered by the ENCODE project (The ENCODE Project Consortium, 2007) gives the following definition:

> A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.

This definition, although more correct in many ways, is perhaps more complex than is necessary for the work described here. I will therefore adopt a more classical definition of a gene:

> A gene is a sequence of nucleotides (DNA) which is transcribed to RNA.

Most current definitions of a gene also include its associated regulatory regions (promoters, enhancers, splicing signals etc). I have kept a more limited definition because my work does not touch on the regulatory elements of a gene but only in that portion of it that is transcribed to RNA. Given the above definition, the eukaryotic gene can be summarized as shown in figure 1.9.

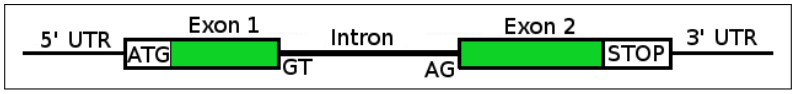

Figure 1.9: **Simplified eukaryotic gene structure**

1.5 Eukaryotic gene prediction

Gene prediction is still one of the more challenging fields in computational biology. Ideally, researchers should be able to infer all the possible genes in a given DNA sequence automatically. Although we are ever approaching this ideal, available methods are still not perfect.

The general problem of gene prediction consists of identifying those stretches of genomic sequence that will be transcribed to RNA and give rise to a protein (or non-coding RNA) product. Genes form only a small percentage of a genome. In the human genome, for example, exons (those regions of a gene, be they coding or non-coding, that survive to the mature mRNA) account for only 1.2% of the total genomic DNA (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001)[12]. Therefore, correctly identifying these exons and assembling them into genes is an extremely complex task.

This section is largely derived from a recent review of gene-finding strategies (Harrow et al., 2009) to which the reader is referred for more details. Briefly, information on

[12]However, the recent ENCODE paper (The ENCODE Project Consortium, 2007) has shown that as much as 93% of bases may actually be transcribed to RNA

the location of possible genes on a given genomic DNA sequence can come from various sources: conservation with other, informant, genomes (Figure 1.10(1)); sequence signals such as START and STOP codons, splice sites etc (Figure 1.10(2)); statistical properties that differentiate coding from non-coding sequences. (Figure 1.10(3)); and known transcript and protein sequences from other genomes (Figure 1.10(4)).

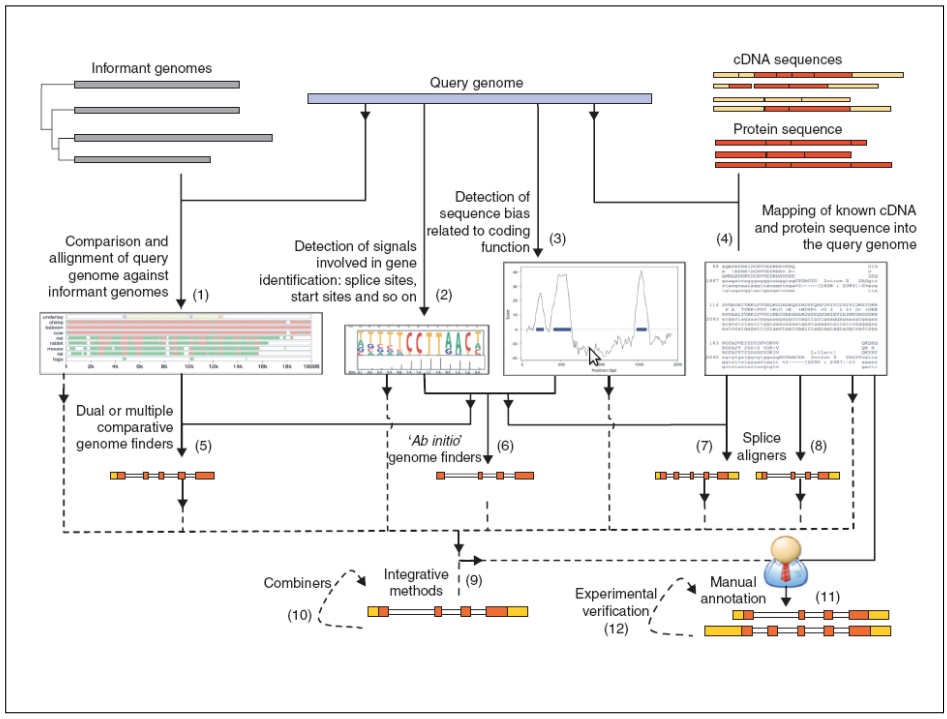

Figure 1.10: **Gene-finding strategies.** Please see text for details. Figure taken from (Harrow et al., 2009)

Ab initio gene finders such as `GeneID` (Guigó et al., 1992; Parra et al., 2000), and `genscan` (Burge and Karlin, 1997), use "intrinsic" evidence (e.g. codon bias, third base periodicity, see Cruveiller et al. (2003) for a review of these features) to produce gene predictions[13]. These programs are the only choice available in the absence of known transcript or protein sequences or phylogenetically related genomes. In the cases where such informant sequences are available, the intrinsic information can be combined with patterns of genomic sequence conservation using programs often referred to as comparative (or dual- or multi-genome) gene finders such as `SGP` (Parra et al., 2003) or `twinscan` (Korf et al., 2001). The most sophisticated of these combine information from many genomes, taking into account the phylogenetic distances involved, when scoring their predictions. Usually, when cDNA or EST sequences are available, these take priority over other sources of information.

These informant sequences are then mapped to the target genome using a variety of tools, including simple sequence-similarity searches. The initial mapping is subsequently refined using more sophisticated, "splice alignment" algorithms, capable of modeling in-

[13]Today, most *ab initio* gene finders can also use external sources to inform their predictions

trons (Figure 1.10(8)). Another approach is to give available transcript or protein information to *ab initio* gene finders which will then build their prediction while taking these external sequences into account (Figure 1.10(7)).

Often, however, available cDNA and protein evidence is only partial. In such cases, the initial reliable gene and transcript set may be extended with more hypothetical models derived from *ab initio* or comparative gene finders, or from the genome mapping of cDNA and protein sequences from other species. This multi-step process has been automated by various pipelines (Figure 1.10(9)).

More recent programs combine the output of many gene finders (Figure 1.10(10)). Such "combiners" rely on the assumption that predictions shared between various programs are likelier to be true. Predictions are therefore weighed according to the particular features of the program producing them.

Despite the variety of tools and methods described above, the most reliable gene models are still those obtained after manual curation of automatic predictions (Figure 1.10(11)).

1.6 RNA structure prediction

Over the last few years, the discovery of many small non-coding RNAs with diverse functions has highlighted the importance of the RNA world. Where RNA was relegated to the role of simple messenger, it has now become clear that it is in fact a major player in the world of molecular biology and genetics. The RNA world is one of the most interesting and active fields of research in biology today for a recent review please see (Mattick and Makunin, 2006). In this section I will focus only on the computational prediction of RNA structures.

1.6.1 RNA structure

Like that of all macromolecules, RNA structure is hierarchical. The four levels of structure of RNA molecules are the following:

Primary structure is simply the nucleotide sequence of the molecule.

Secondary structure is the way the molecule folds itself in two dimensions, loops, helices etc (see Figure 1.11a). RNA molecules are more permissive than DNA, not only AU and GC but also GA and GU pairing is allowed.

Tertiary structure is the way the RNA secondary structure folds in on itself. Tertiary structure depends on the three-dimensional positioning of the atoms in the RNA molecule and on the interactions between them. See Figure 1.11b for a few well known tertiary structures.

Quartenary structure is the interaction with other molecules.

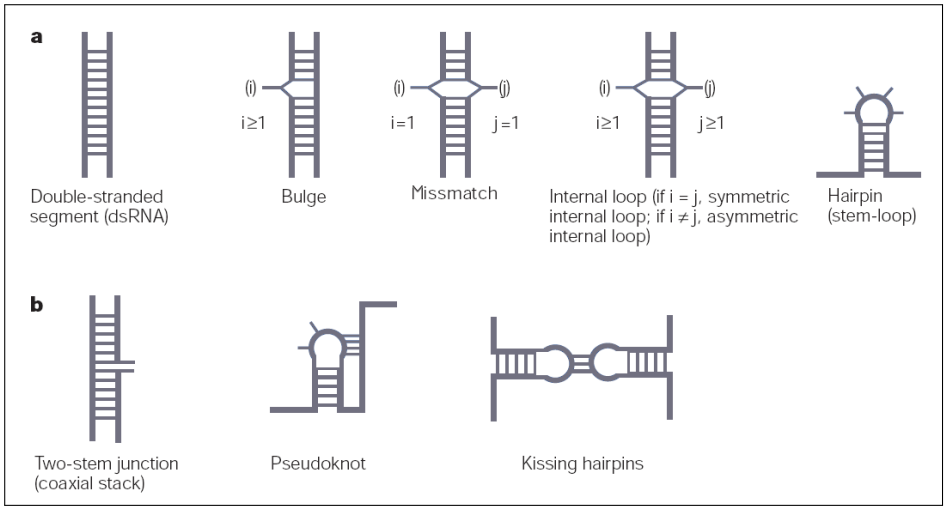

Figure 1.11: **Common RNA structures**. Examples of commonly found RNA secondary (a) and tertiary (b) structures. Image adapted from Tian et al. (2004)

1.6.2 RNA structure prediction

Because of the variety of possible conformations an RNA molecule can adopt (see Figure 1.11), predicting the structure of an RNA molecule is a computationally hard problem. It has been shown (Zuker and Sankoff, 1984) that the number of possible structures for a sequence grows exponentially with length, N:

$$\text{Possible secondary structures} \approx (1.8)^N$$

This means that even for sequences as short as 100nt, the number of possible structures is approximately 10^{25}. Given that a modern computer processor can calculate the free energy for about 10,000 structures in a second, this calculation would require 10^{21} seconds or 10^{13} years[14]! In order to circumvent this problem, a number of dynamic programming algorithms have been devised.

The most common are based on the computation of the lowest free energy ($DeltaG$) structure. Examples of these programs are mfold (Zuker, 2003) and RNAFold (Schuster et al., 1994). These programs are based on the assumption that an RNA molecule will fold into the most stable conformation possible. However, this is not always the case since RNA molecules can adopt non-optimal and even multiple conformations. Therefore, prediction of suboptimal structures is also necessary. Although programs like mfold can return suboptimal structures, the nature of the dynamic algorithm is such[15] that certain structures will be missed. Attempting to calculate all possible sub-optimal structures is impractical for the reasons explained above. A possible way around this is the use of statistical sampling to select more likely structures from the predictions. A recent program implementing this approach is Sfold (Ding et al., 2004).

[14]Taken from Mathews (2006)

[15]For a review of RNA folding algorithms see Mathews (2006)

Whatever the approach used, RNA structure prediction is still imperfect with only limited accuracy at the tertiary level. Happily however, secondary structure prediction of relatively short sequences is quite robust as can be seen by its extensive application in the prediction of SECIS elements (see Results).

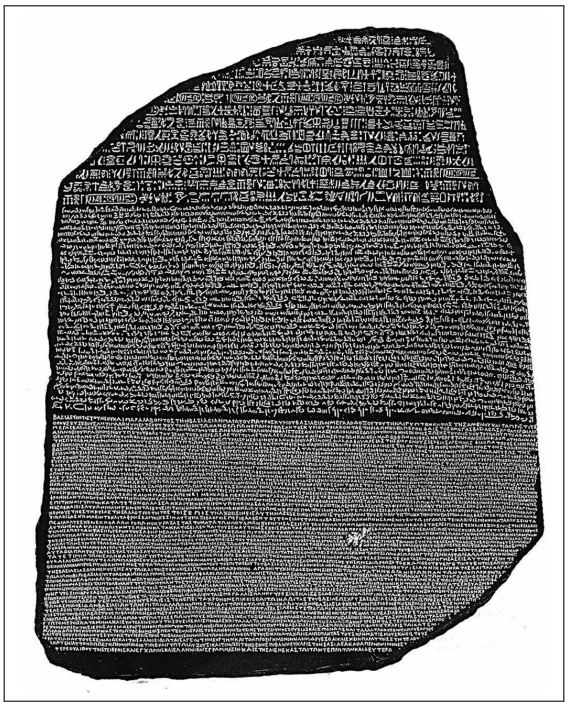

Figure 1.12: **The Rosetta stone**. The text on the stone is a decree from Ptolemy V, describing the repealing of various taxes and instructions to erect statues in temples. From http://www.uncp.edu/home/rwb/lecture_ancient_civ.htm.

1.7 Comparative genomics

Comparative genomics is possibly the most powerful tool in genome research today. The field of comparative genomics relies upon the assumption that functional elements (genes, proteins, sequence signals etc) are conserved across species. That is, a predicted gene model, for example, found conserved in multiple genomes is more likely to represent a *bona fide* gene than one found in only a single species. Although by no means certain, this assumption is a very useful tool. For an example of such cross-species conservation, please see the alignment figures from Chapple and Guigó (2008), Results section , page 55.

In keeping with what has become a tradition in the Guigó group, I will use the Rosetta Stone to illustrate the concept of comparative genomics. The Rosetta Stone (see Figure 1.12) is an Ancient Egyptian artifact which was instrumental in advancing modern understand-

ing of hieroglyphic writing. It carries three versions of the same passage, two in Egyptian language scripts and one in ancient Greek. A french scholar, Jean-François Champollion, used the Greek text which he could understand as a starting point for the translation of the two –then unknown– Egyptian scripts.

In a similar way, biologists can use information previously known in one species to help their search in another. We have already seen an application of this concept for comparative gene finding (see page 22). This conservation of elements and mechanisms (biological pathways etc) is also the underlying assumption behind the study of model organisms. cancer studies done on the mouse can yield informative results for humans as well.

Comparative genomics tools and approaches have been extensively used throughout the present work. The first step in any prediction of a novel selenoprotein gene is to search sequence databases for conservation with other species. In the ORF approach (see Methods, section 3.3, page 109) developed for the *C. elegans* and *C. briggsae* selenoproteomes (Taskov et al., 2005) we used annotated genes in one nematode genome ti inform our predictions in the other. The novel selenoprotein gene *selj* (Castellano et al., 2005) predicted in the *Tetraodon* genome was recognized as a *bona fide* novel gene based on, among other things, its conservation across actinopterygian fish species. For our work on insect genomes (Chapple and Guigó, 2008) we also extensively relied on comparative genomics approaches using the well-annotated genomes of *Drosophila melanogaster* and *A. gambiae* to inform our searches in the more recently sequenced species.

CHAPTER 2

Results

Summary

In this section, I include all my published articles which are directly relevant to my thesis. The articles are presented chronologically and in the way they appeared in the original journals. The 12 fly genome paper is too long to be included here and can be found in Appendix 6. The tetraodon genome paper cannot be used here as it has already been included in another thesis, it can be found in Appendix 6. Our paper on the taxonomic diversity of a metagenomic library is not relevant to the subject of my PhD and is included as Appendix 6.

Contents

2.1 Taskov et al, 2005

Taskov K, Chapple C, Kryukov GV, Castellano S, Lobanov AV, Korotkov KV, Guigó R, Gladyshev VN.
Nematode selenoproteome: the use of the selenocysteine insertion system to decode one codon in an animal genome?
Nucleic Acids Res. 2005 Apr 20;33(7):2227-38. Print 2005.

In this paper we identified the first organisms to have retained the entire selenoprotein machinery for the benefit of just a single selenoprotein gene. It was a collaboration between Vadim Gladyshev's group and our own. I carried out all the research for the SECIS independent approach (see Figure 1 of the manuscript).

_ Article abstract :
http://nar.oxfordjournals.org/cgi/content/short/33/7/2227
_ Full text :
http://nar.oxfordjournals.org/cgi/content/full/33/7/2227
_ PDF :http://nar.oxfordjournals.org/cgi/reprint/33/7/2227.pdf

2.2 Castellano et al, 2005

Castellano S, Lobanov AV, Chapple C, Novoselov SV, Albrecht M, Hua D, Lescure A, Lengauer T, Krol A, Gladyshev VN, Guigó R.
Diversity and functional plasticity of eukaryotic selenoproteins: identification and characterization of the SelJ family.
Proc Natl Acad Sci U S A. 2005 Nov 8;102(45):16188-93. Epub 2005 Oct 31.
Comment in:
Proc Natl Acad Sci U S A. 2005 Nov 8;102(45):16123-4.

In this paper we describe the novel selenoprotein SelJ.

_ Article abstract :
http://www.pnas.org/content/102/45/16188.abstract?sid=10f9ac36-34c3-42b3-acd0-a4d31e3f3303
_ Full text :
http://www.pnas.org/content/102/45/16188.full?sid=10f9ac36-34c3-_
PDF :
http://www.pnas.org/content/102/45/16188.full.pdf+html?sid=10f9ac36-34c3-42b3-acd0-a4d31e3f3303
_ Supplementary material:
http://www.pnas.org/content/102/45/16188/suppl/
DC1additional data

2.3 Drosophila 12 Genomes Consortium, 2008

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM,Oliver B *et al.*
Evolution of genes and genomes on the Drosophila phylogeny.
Nature. 2007 Nov 8;450(7167):203-18.
Comment in:
Nature. 2007 Nov 8;450(7167):184-5.

In this paper, the Drosophila 12 genomes consortium[1] published a comparative analysis of the genomes of 12 Drosophyla flies. One of the more interesting findings of this paper is that one species of Drosophila (D. willistoni has completely lost the ability to code for selenoproteins.
Because of the length of this paper, only the first page and the two relevant paragraphs are included here. The entire article can be found at 6, page 163

_ Article abstract :
http://www.nature.com/nature/journal/v450/n7167/abs/nature06341.html
_ Full text :
http://www.nature.com/nature/journal/v450/n7167/full/nature06341.html
_ PDF :
http://www.nature.com/nature/journal/v450/n7167/full/nature06341.html
_ Supplementary material:
http://www.nature.com/nature/journal/v450/n7167/suppinfo/nature06341.htmladditional data

2.4 Chapple and Guig´o R., 2008

Chapple CE, Guigó R.
Relaxation of selective constraints causes independent selenoprotein extinction in insect genomes.
PLoS One. 2008 Aug 13;3(8):e2968.

In this article we reported the first selenoprotein lacking animals and the general depletion of selenoproteins in the Insecta.

_ Full text :
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0002968
_ PDF :
http://www.plosone.org/article/fetchObjectAttachment.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0002968&representation=PDF
_ Supplementary material :
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0002968#s5

2.5 Chapple et al, 2009

Chapple CE, Guigó R, Krol A.
SECISaln, a web-based tool for the creation of structurebased alignments of eukaryotic SECIS elements.
Bioinformatics. 2009 Mar 1;25(5):674-5. Epub 2009 Jan 29.

In this paper, we present the web-based tool SECISaln which provides for the first time an extensive structure-based sequence alignment of SECIS elements resulting from the well-defined secondary structure of the SECIS RNA and the increased size of the eukaryotic selenoproteome. We used SECISaln to improve our knowledge of SECIS secondary structure and to discover novel, conserved nucleotide positions. The work presented here was carried out both in Barcelona and in Dr. Krol's laboratory in Strasbourg, France.

_ Article abstract :
http://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/5/674
_ Full text :
http://bioinformatics.oxfordjournals.org/cgi/content/full/25/5/674
_ PDF :
http://bioinformatics.oxfordjournals.org/cgi/reprint/25/5/674
_ Supplementary material :
http://bioinformatics.oxfordjournals.org/cgi/content/full/btp020/DC1

2.6 Takeuchi et al, 2009

Takeuchi A, Schmitt D, Chapple C, Babaylova E, Karpova G, Guigo R, Krol A, Allmang C.
A short motif in Drosophila SECIS Binding Protein 2 provides differential binding affinity to SECIS RNA hairpins.
Nucleic Acids Res. 2009 Apr;37(7):2126-41. Epub 2009 Feb 17.

In this paper we report a novel domain in the protein SBP2 which confers differential preferences for the two types of SECIS element.

_ Article abstract :
http://nar.oxfordjournals.org/cgi/content/abstract/gkp078v1
_ Full text :
http://nar.oxfordjournals.org/cgi/content/full/gkp078v1
_ PDF : http://nar.oxfordjournals.org/cgi/screenpdf/gkp078v1
_ Supplementary material :
http://nar.oxfordjournals.org/cgi/content/full/gkp078/DC1

2.7 Bovine Genome Sequencing and Analysis Consortium, 2009

Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM *et al*. *The genome sequence of taurine cattle: a window to ruminant biology and evolution.*
Science. 2009 Apr 24;324(5926):522-8.
Comment in:
Science. 2009 Apr 24;324(5926):478-9.
Science. 2009 Jun 19;324(5934):1515.

In this paper (Bovine Genome Sequencing and Analysis Consortium, 2009), the bovine genome sequencing and analysis consortium2 reports the sequencing of the cow genome.

_ Article abstract :
http://www.sciencemag.org/cgi/content/abstract/324/5926/522
_ Full text :
http://www.sciencemag.org/cgi/content/full/324/5926/522
_ PDF : http://www.sciencemag.org/cgi/reprint/324/5926/522.pdf
_ Supplementary material :
http://www.sciencemag.org/cgi/content/full/324/5926/522/DC1

CHAPTER 3

Methods

Summary

In this chapter, I will give a brief description of the general tools used throughout this work, as well as a more detailed one of the special software developed specifically for selenoproteins. I will also give a brief description of the BLAST algorithm and the basic ideas underlying sequence alignments since these are the two non selenoprotein specific tools on which I have relied the most.

Contents

3.1 General Tools

3.1.1 BLAST

THE **B**ASIC **L**OCAL **A**LIGNMENT **T**OOL, or BLAST as it is commonly known, is probably the most essential tool in the field of comparative genomics and bioinformatics in general. Developed by Altschul et al the 1990s (Altschul et al., 1990, 1997), BLAST searches a nucleotide or protein database for sequences showing similarity to the user submiteted sequence(s).

The BLAST algorithm is relatively straightforward. In the first step, BLAST removes low complexity regions from the query sequence. It then generates a list of all possible "words" of length k (default k values are 3 for protein searches and 11 for nucleotides) in the query sequence(s). So, for example, given a protein query sequence of CPQGKF and a word length k=3, BLAST will create the following words: CPQ,PQG,QGK and GKF. Once compiled, each of these query words is compared to all possible k-length words in the subject database. BLAST then scores all word pairs, using, in the case of protein searches, a protein substitution matrix in the case of protein searches such as PAM (Dayhoff et al., 1978) or BLOSUM (Henikoff and Henikoff, 1992), which is the users choice. In the case of nucleotide searches, a match is simply scored as +5 and a mismatch as -4. Once this list of word pairs has been compiled, BLAST keeps only those pairs whose score passes a certain threshold. Each of these is used to seed an alignment. Starting from each of the high-scoring words, BLAST will attempt to extend the alignment around the matched word. The resulting alignment is called a **H**igh-scoring**S**egment **P**air (HSP).

Now, the score of each HSP is calculated (in the same way as that of the original "words") and only those HSPs whose score is higher than a predefined threshold (determined by comparing random sequences) are kept. Blast next asseses the E value of each HSP. The E value represents the probability of finding an HSP with the obtained score by chance given the query sequence length and the size of the database used.

BLAST comes in several "flavours":

`blastn`
: Compares a nucleotide query against a nucleotide database.

`blastp`
: Compares a protein query against a protein database.

`blastx`
: Translates a nucleotide query in all 6 frames and compares it against a protein database.

`tblastn`
: Compares a protein query against a nucleotide database translated in all 6 frames.

`tblastx`
: Translates a nucleotide query in all 6 frames and compares it against a against a nucleotide database likewise translated in all 6 frames.

`psi-blast`
: **P**osition **S**pecific **I**terated `BLAST` takes a protein sequence as input and uses it to run a simple blastp search of a protein database. The highest scoring hits are then combined to build a "profile" which is then used to further query the subject database. psi-blast is generally used to search for distant homologues.

3.1.2 Sequence alignment software

The field of multiple sequence alignment methods is one of the most active areas of bioinformatics. A full description of the various algorithms involved is well beyond the scope of this PhD thesis. Nevertheless, a brief description of the basic ideas and assumptions underlying multiple sequence alignments is necessary for the understanding of the work presented here.

A multiple sequence alignment is a collection of sequences which have been so arranged that homologous residues are arranged in columns. In the case of protein sequences, alignment programs take into account the chemical properties of the aligned amino acids such as hydrophobicity or acidity.

The first and by far the most important assumption made when using multiple sequence alignment is that sequence similarity implies homology. Although this is not always the case (for example in the case of conserved domains) it is by and large a valid assumption and one I will be using throughout this text.

A necessary caveat when discussing multiple sequence alignments is that because an exhaustive calculation of the optimal alignment between multiple sequences is computationally expensive, multiple alignment algorithms take some heuristic shortcuts which may result in imperfect alignments.

For the work described here I have used a variety of alignment programs: `t_coffee` (Notredame et al., 2000), `mafft` (Katoh et al., 2002), `kalign` (Lassmann and Sonnhammer, 2006) and `clustal` (Chenna et al., 2003).

3.1.3 GeneID

`Geneid` (Guigó et al., 1992; Parra et al., 2000) is an *ab initio* gene predictor developed by our group. It was one of the first such programs to be written and is still among the most widely used today. `Geneid` is designed with a hierarchical structure: first, gene-defining signals (splice sites and start and stop codons) are predicted along the query DNA sequence. Next, potential exons are constructed from these sites, and finally the optimal scoring gene prediction was assembled from the exons.

`GeneID` is at least as accurate as other gene finders (Guigó et al., 2006) and is particularly efficient at handling very large genomic sequences, both in terms of speed and usage of memory. The gene assembly step in `GeneID` is handled by a dynamic algorithm (GenAmic) which searches the space of predicted exons and returns gene structures maximizing the sum of the scores of the assembled exons.

Unlike most gene predictors, GeneID can also deal with atypical gene models. Modified versions of GeneID exist which are capable of predicting U12 introns (**?**), and exons with TGA in-frame (Castellano et al., 2001). For the case of selenoprotein genes, GeneID (or GeneidSP allows a dual meaning of the TGA codon (both STOP and Sec). Part of the user-modifiable parameters of GeneID is a list of permitted gene structures. That is, First exons can be followed by either a terminal or an internal exon, internal exons must be followed by either another internal or a terminal exon and sop forth. For selenoprotein gene prediction, GeneidSP is given the positions of SECIS elements on the target sequence, as predicted by SECISearch (see Section 3.2.3) as "external evidence". The gene model options then allow GeneidSP to include ORF(s) with an in-frame TGA in a gene model only if there is a predicted SECIS element at the appropriate distance downstream.

3.1.4 Others

RNAFold

RNAFold is an RNA folding program and part of the Vienna RNA Package (Schuster et al., 1994). The folding algorithm is based on a dynamic programming algorithm originally developed by M. Zuker and P. Stiegle. It will take a FASTA sequence as input and calculates the minimum free energy (ΔG) structure.

patscan

patscan (Dsouza et al., 1997)is a pattern finder. It will scan the input sequence for occurrences of a user defined pattern. For example, the following pattern will find a 10-15nt palindrome separated by 5-10 nucleotides:

```
p1=10...15 5...10 ~p1[1,0,1]
```

p1=10...15 captures the first pattern (p1). 5...10 allows any nucleotides (numbering from 5 to 10).

~

exonerate and genewise

Genewise (Birney et al., 2004), predicts gene structure using similar protein sequences. It is particularly useful when mapping a known protein from one species to the genome of a phylogenetically close second species. Unlike simple similarity search programs such as BLAST, genewise includes a sophisticated splice site models and so can return complete gene models. Genewise, while very accurate, is quite slow and is not practical for sequences exceeding 50000nt in length.

Exonerate (Slater and Birney, 2005), on the other hand is extremely fast. It is also a far more general tool than genewise, incorporating various modes of sequence comparison. Among others, Exonerate can be used to map proteins onto genomic or cDNA sequences, cDNAs onto genomic sequences, or to generate simple gapped or ungapped alignments.

Both programs have been used extensively throughout my PhD work. In my experience, the best results are obtained when using exonerate to zero in on the region of interest and then genewise for the final gene structure prediction.

3.2 MyTools

In this section I will describe three scripts I developed[1] specifically for and which were used extensively throughout the work described here.

3.2.1 alignthingie

alignthingie is essentially a blast outfile parser. The main advantage it has over other such programs is that it can search for HSPs where a specific residue (or sequence of residues) is aligned to another specific residue (or sequence of residues). Although it is designed to search for *-* or *-Cys alignments, alignthingie can be used both as a generalized BLAST parser or specifically to search for other aligned residues. It can return results as alignments (HSPs) or gff files for either subject or query. It also offers a variety of cutoff options.

alignthingie is freely available (under the GPL license) from http://genome.crg.es/~cchapple/alignthingie.pl. A list of the options follows:

```
COMMAND-LINE OPTIONS:

 -c : Minimum number of conserved residues (or '+') allowed around
      the matched residue (integer, def : 6)
 -C : Also check for Cs in the subject sequence which align to a
      '*' in the query
 -e : Maximum e-value allowed (integer, def : 10)
 -i : Minimum (i)dentity percentage allowed (integer, def : 0)
 -I : Maximum (I)dentity percentage allowed (integer, def : 100)
 -l : Minimum number of conserved residues (or '+') allowed on the
      (l)eft side of the matched residue (def : 3)
 -r : Minimum number of conserved residues (or '+') allowed on the
      (r)ight side of the matched residue (def : 3)
 -M : (M)aximum score value allowed (integer, def : 10000)
 -m : (m)inimum score value allowed (integer, def : 0)
 -q : String to match in query (def : '*' )
 -s : String to match in subject     (def : '*')
```

[1] Or modified extensively in the case of SECISearch

```
 -R : Length of (R)egion around matched residue to check for
      conservation (integer, def : 6, R-matched residue-R)
 -S : Use strict evalue, id, score (0.01,65,50 respectively)
      and conservation cuttoffs.
 -k : Do not check conservation
 -u : How many (u)naligned residues are allowed with respect to
      query length.(<query length> - <hsp length> <= <value passed>)

OUTPUT OPTIONS:
 -A : Return (A)ll HSPs which pass thresholds without looking for
      any specific aligned residues
 -b : Print only the (b)est (lowest e-value) hit for each query.
      If the smallest evalue is shared by more than one HSP, all
      such HSPs will be printed.
 -B : Print only the (B)est (lowest e-value) hit for each query.
      If the smallest evalue is shared by more than one HSP, only
      the first such HSP will be printed.
 -d : (d)ebugging mode, very very verbose...
 -f : No sel(f) : Skips subjects whose name matches (case-INsensitive)
      the value passed. (string)
 -F : Generalised no sel(F), takes first characters (until
      the first space) of the query and subj names and skips
      the hit if the 2 are identical.
 -g : (g)ff output. Use "-g q" for query position gff and "-g s" for
      subject gff.
 -L : Print most (L)ikely hits. ie, those with no stop codon before
      the matched residue and whose conservation on the right side
      of the match is no more than 2 less than that of the left side.
 -n : Print only the names of those queries which returned NO HSP.
 -p : Do not return hits against (p)lant species.
 -Q : (Q)uery name or list of names (text file, one name per line) to
      return HSPs for. Only those HSPs whose query is specified will
      be printed
 -T : (T)arget (subject) name or list of names (text file, one name
      per line) to return HSPs for. Only those HSPs whose subject
      is specified will be printed.
 -v : (v)erbose output, prints a . for each query processed.
 -V : More (V)erbose output, prints a . for each query processed
      and a '!' for each hit found.
 -x : Only return those hits with a redox box CXXU/*
 -X : Read a list of species names to ignore hits against them.
 -U : Query name or list of (U)nwanted queries. Quoted list of query
      names (or text file, one name per line) for which NOT
      to return HSPs.
```

3.2.2 retrieveseqs

Near the beginning of my PhD I was unable to find a quick and easy way to extract sequences from a FASTA file given a list of wanted sequence IDs. I needed to pass a number of groups of sequence names and obtain a multi-fasta file for each of the groups. I wrote this script, `retrieveseqs` which will do exactly that. It will take a list(s) of sequence names either as separate files or as options on the command line and retrieve their sequences from a multi fasta file.

`retrieveseqs` is freely available (under the GPL license) from http://genome.crg.es/~cchapple/retrieveseqs.pl. A list of the options follows:

```
USAGE: retrieveseqs.pl [-viofsn] <FASTA sequence file> <desired IDs, one

COMMAND-LINE OPTIONS:
  -v : verbose output, print a progress indicator (a "." for every
       1000 sequences processed)
  -V : as above but a "!" for every desired sequence found.
  -f : fast, takes first characters of name "(/^([^\s]*)/)" given
       until the first space as the search string
       make SURE that those chars are UNIQUE.
  -i : use when the ids in the id file are EXACTLY identical
       to those in the FASTA file
  -h : Show this help and exit.
  -o : will create one fasta file for each of the id files
  -s : will create one fasta file per id
  -n : means that the last arguments (after the sequence file)
       passed are a QUOTED list of the names desired.
```

3.2.3 SECISearch

`SECISearch` was originally developed by Gregory V. Kryukov from Vadim Gladyshev's group. I include it here because the version I use is one that I have extensively modified. My first contact with programing was when I first arrived at the Guigó lab and Roderic gave me a book on Perl and told me to make a command line version of the web-based `SECISearch`. And so I did.

SECISearch is essentially a wrapper script which will pipe input sequence data first through patscan (Dsouza et al., 1997) and then through `RNAFold` from the Vienna RNA package (Schuster et al., 1994). Finally, what really makes `SECISearch` better than just using these two programs is the imager2[2] function. This will take the postcript output of `RNAFold` and return pretty .png images with the important elements of the SECIS consensus (apical Rs and quartet) shown in bold.

The following are the modifications I have made to `SECISearch`. First of all, `SECISearch` can now run on the command line and can deal with multifasta files, allowing batch ex-

[2] All credit for this belongs with Gregory, I don't understand half of how imager2 works and have not touched it in any way.

ecution for large jobs (e.g. whole genome searching). SECISearch can now output .gff and .fasta files of each prediction made. It can also output sequences which while containing the pattern the user searched for, did not pass the thermodynamic evaluation. Finally, where the original SECISearch would check for the ΔG of the upper stem as well as that of the entire structure. My testing showed that this did not give a significant improvement so my version of SECISearch only checks the ΔG of the entire structure.

3.2.4 SECISaln

SECISaln (Chapple et al., 2009) is a program I developed in collaboration with Dr Alain Krol. SECISaln will predict a eukaryotic SECIS element in a nucleotide sequence, split it into its structural units and then align each unit against the SECISes in our database. SECISaln can distiguish between typeI and typeII SECIS elements and will align the submitted sequence against others of the same type. All sequences used by SECISaln have been collected from either GenBank or EGO.

SECISaln is not intended to replace SECISearch as a SECIS element predictor. In fact, SECISaln uses SECISearch to predict SECIS elements. The objective of this tool is to provide researchers with an easy way to compare structural features of SECIS elements. It should only be used on sequences known to contain a SECIS element. The pattern used by SECISaln to recognise SECIS elements is very permissive and would result in false positives when run on unknown sequences.

SECISaln first predicts a SECIS element on the user-submitted sequence. This is done by running SECISearch as an internal subroutine. Once the prediction has been made, SECISaln identifies the type (I or II) of the SECIS. SECISaln makes use of two intermediate files created during the SECIS prediction step one by RNAFold and one by patscan (see Figure 3.1).

```
>RNAFold output
CAGCGGGACUGGUGUUAAUGAAGGCUUGCACUGAAAACACUUGCUGUUAGUGUAGGCUGGAGUUCUCCCUGCCGUCUCGCUGCA
(((((((((.((((....(((((((((((((((...((((.....))))))))))))))))))).......)))))))))))..  (-31.15)
>patscan output
CAGC GGGACUGGUGUUA AUGAA GGCUUGCACUG AA AACACUUGCUG UUAGUGUAGGCU GGAG UUCUC CCUGCCG UCUCGCUGCA
```

Figure 3.1: patscan and RNAFold sample output files

SECISaln uses the RNAFold output to determine the folding of the predicted SECIS and the patscan output to determine the structural units. Once this has been done, SECISaln will align each of the units in the following way (see Figure 1.8, page 21 for an explanation of the structural units):

Helices 1 and 2 (5') are positioned so that the core quartets of all the SECISes are aligned. Gaps are added to the right, as necessary.

The apical (or internal for type II) loop is placed so that the conserved As are aligned. Gaps are added to the right and left, as necessary.

Helix 3 (5′) is positioned so that the fist nt of all the third helices is aligned. Gaps are added to the right, as necessary.

The apical loop of type II elements is positioned so that the fist nt of all the apical loops is aligned. Gaps are added to the right, as necessary.

Helix 3 (3′) is positioned so that the last nt of all the third helices is aligned. Gaps are added to the left, as necessary.

Helices 1 and 2 (3′) are positioned so that the core quartets of all the SECISes are aligned. Gaps are added to the left, as necessary.

`SECISaln` is freely available as a web-based server at `http://genome.crg.es/software/secisaln/`

3.3 The ORF approach

This SECIS-independent selenoprotein gene prediction method was developed for the selenoproteome of *C. elegans* (Taskov et al., 2005). At the time of this study, only one selenoprotein, Thioredoxin Reductase (TR) had been identified in the genome of the nematode *C. elegans*. In contrast, other nematodes were known to express additional selenoproteins. Vadim Gladyshev of Nebraska university suggested a collaboration with an end to determine whether this genome did indeed code for only a single selenoprotein. In this work we employed various prediction methods which are explained in the published article (see section 2.1):

The approach described in this article is the first SECIS-independent method for selenoprotein gene prediction. It builds on the already established protocol of SECIS-informed gene prediction Kryukov et al. (2003), and adds a comparative approach which, by virtue of being independent of the SECIS element, is capable of recognising genes with atypical SECIS elements. Any such genes would be missed by the traditional approach. We believe that the combination of the SECIS dependent and independent methods can provide a definitive map of a species' selenoproteome.

To find potential novel selenoproteins in the *C. elegans* and *C. briggsae* genomes, we developed the "ORF approach" which is described in more detail in (Taskov et al., 2005). Briefly, we predicted all TGA_ORFs in both genomes. A TGA_ORF is defined as a stretch of sequence of at least 60nt between two non-TGA, in-frame stop codons which also contains at least one in-frame TGA codon. The set of all TGA_ORFs should contain all selenocysteine coding TGAs in a given genome[3] (see Figure 3.2). Although this method lacks any support for exons, it does provide a suitable collection of sequences for a brute approach.

Since we were looking for genes that had been completely missed by the automated annotation pipeline, the TGA_ORFs were predicted after masking known genes in both

[3]With the possible exception of an exon so close to the start or the end of a chromosome that it is not between any other two STOPS. However, if this is even possible (given that chromosome ends are non-coding), it is very unlikely.

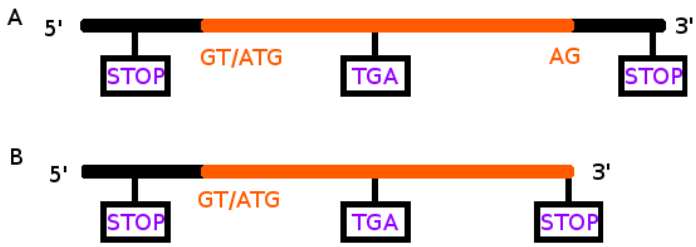

Figure 3.2: **The possible sequences selected by the ORF approach.** The TGA ORF is the sequence between the two non-TGA STOP codons. The potential TGA-containing exon is in orange. **A**: Internal or Initial exon.**B**: Terminal exon or single exon gene.

genomes. Any novel TGA containing exon should fall in regions of the genome annotated as intergenic.

The two sets of TGA_ORFS (those from *C. elegans* and those from *C. briggsae*) were compared using BLASTP. We looked for the following three combinations:

1. Sec in *C. elegans* and Sec in *C.briggsae*. *C. elegans* translated TGA-containing ORFs were compared against *C.briggsae* translated TGA-containing ORFs.
2. Sec in *C. elegans* and Cys in *C.briggsae*. *C. elegans* translated TGA-containing ORFs were compared against the set of annotated C.briggsae proteins.
3. Sec in *C. briggsae* and Cys in *C. elegans*. The set of annotated *C. elegans* proteins was compared against *C.briggsae* translated TGA-containing ORFs.

In each case, we selected HSPs with at least 3 out of 6 residues conserved on either side of the putative selenocysteine. All such HSPs were then blasted (tblastn) against non_human and non_mouse EST databases to check for conservation against other species. ORFs were further extracted that were represented by at least 5 unique alignments with an E-value <0.1 and at least 5 (out of 10) conserved residues on both sides of the aligned TGA codon. Finally, to assess whether the frame in which the ORF has been defined is the true coding frame, we used tblastx to align the ORFs against their conserved ESTs and only kept those ORFs whose highest-scoring HSP matched the previously predicted ORF with the aligned Sec codon.

CHAPTER 4

Discussion

Summary

In this chapter I will first give an historical perspective of the field of selenoproteins and selenoprotein gene prediction, describing what we knew in 2002, at the beginning of my thesis work. I will discuss the novel methods developed by our group during my PhD and the results we have obtained through them. I will then discuss the most important points of my published work as well as the advances made in the field the past few years, highlighting our contributions. Finally, I will attempt to meld this information into a coherent view of selenoprotein evolution.

Contents

4.1 Then and now...

I had the luck to enter the field of selenoprotein research at a particularly interesting time. The past few years have seen a wealth of new developments, new genomes sequenced, new proteins and factors found and novel methods developed. When the world of selenoproteins was first described to me, the situation seemed quite simple. The following are some of the things we believed:

- Selenocysteine is a selenium-containing analog of cysteine which confers greater reactivity to the enzymes incorporating it. Therefore, Sec versions of proteins exist to give the cell a more reactive species of enzyme for its times of need.
- The number of selenoproteins increased with the complexity of the organism. Higher eukaryotes have more selenoproteins than "simpler" animals, with mammals at the pinnacle with 21 selenoproteins[1](Gladyshev et al., 2001).
- Selenoproteins were essential for animal life. All animals have selenoproteins.
- Land plants lack selenoproteins.
- The SECIS element is recognised by SBP2 which binds to both the SECIS element and the ribosome as well as interacting with EFsec which is carrying the $tRNA^{Sec}$. The $tRNA^{Sec}$ kinase and Sec synthase were still missing but basically we had the major players.
- Eukaryotic SECIS elements are by definition, found on the 3' UTR.
- TypeI and typeII SECISes are equivalent.
- The SECIS sequence always has an unpaired A, followed by the quartet (AUGANGAN) and the apical As (or Cs in the special cases of SelO and SelM)

Although most of the above points still hold true to a point, we now know the situation is more complex. Indeed, as so often happens, the wealth of answers found in the past few years has raised a new thicket of questions.

- Analysis of the new genomic data available today (eg Mita et al. (2004); Jaillon et al. (2004); Honeybee Genome Sequencing Consortium (2006); Drosophila 12 genomes Consortium (2007); Bovine Genome Sequencing and Analysis Consortium (2009)) has given us a far more complete picture of the eukaryotic selenoproteome than we could in 2002.
- It has been shown that by slightly modifying the enzymatic active site, cysteine homologs can equal the reactivity of their Sec counterparts (Gromer et al., 2003). So, why exactly go top such trouble for selenocysteine?

[1]Counting each member of the GPx, DI and TR families

- It turns out that the eukaryote with the greatest number of selenoprotein genes is the green alga *Ostreococcus lucimarinus* (Lobanov et al., 2007) with 29 selenoproteins.
- Selenoproteins are not essential for animal life (Drosophila 12 genomes Consortium, 2007; Lobanov et al., 2008; Chapple and Guigó, 2008). Certain insects lack selenoproteins.
- Land plants lack selenoproteins. Well, this one is still true (so far). However quite a few algae have now been shown to code for Sec. I would be very surprised if we do not find a taxon-specific plant selenoprotein in the next few years.
- Not only have the $tRNA^{Sec}$ kinase, PSTK (Carlson et al., 2004), and Sec synthase, SecS (Ganichkin et al., 2008), been identified but also secp43 (Xu et al., 2005; Small-Howard et al., 2006), ribosomal protein L30(Chavatte et al., 2005), NSEP1 (Shen et al., 2006) and nucleolin (Squires et al., 2007) as involved in selenoprotein biosynthesis. There are far more players involved here than originally thought.
- Although all eukaryotic SECIS elements found to date do indeed reside on the 3′ UTR, experiments have shown (Novoselov et al., 2007) that SECISes can be perfectly functional even in the CDS.
- TypeI and typeII SECISes are not equivalent. Recent work by Alain Krol (Takeuchi et al., 2009), in collaboration with our group, has shown that certain SBP2 proteins have a preferential affinity for typeII SECISes.
- The current core consensus sequence(Chapple et al., 2009) for eukaryotic SECIS elements is RUGAN-MM-KGAW[2]. Of this only the UGA/GA of the quartet is truly constant and in fact, not even that. The *SelT* genes of *Toxoplasma gondii* and *Neospora canine* have a non-canonical GGA/GA sequence instead (Novoselov et al., 2007)

In the next few sections I will discuss those of our published results that have helped us reach some of the above conclusions.

Nematode Selenoproteome

One of the results to come out of my work was the discovery that the nematodes *C. elegans* and *C. briggsae* have but a single selenoprotein gene (Taskov et al., 2005). This result was taken as further evidence of the indispensable nature of selenoproteins in animals. That these two species have retained the entire selenocysteine insertion machinery (see page 15) for the benefit of a single selenocysteine residue highlights the importance of selenoproteins and lends weight to the notion that they are essential for animal life.

For the work presented in this paper, we developed a novel SECIS-independent prediction method for selenoprotein genes, the ORF approach (see Methods, Section 3.3, page 109). The ORF approach has two major limitations:

[2]See Appendix 6, page 181 for an explanation of the ambiguity codes

- It requires "sufficient" conservation after the UGA codon. Sufficient here means enough for it to be picked up by `BLAST`. This means that the ORF approach will miss any selenoproteins where the Sec codon is too close to the end of the CDS. Unfortunately, this is known to be the case in some selenoprotein genes (e.g. Thioredoxin reductases).

- It requires conservation to a known sequence in the database. If a novel sec-containing exon is predicted which, however, has no known cys- or sec-containing homologs, then this exon will be discarded.

The first limitation is a general problem in selenoprotein gene prediction. Conservation past the TGA codon is used to differentiate between real STOP and Sec codons in all but the *ab-initio* approach of `GeneID` (see Section **??**, page **??**) . The second limitation is a general problem of comparative gene prediction. Whenever conservation is used to indicate function, we are dependent on the contents of the available databases. If a given gene has not been described in any organisms, our conservation criteria will classify it as not conserved whether this is actually the case or not.

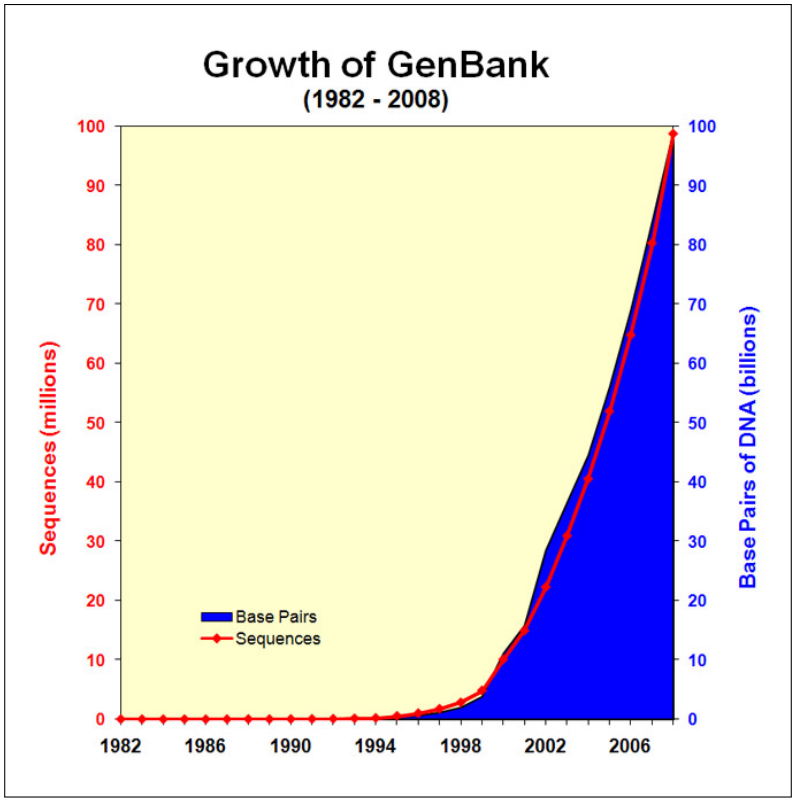

Figure 4.1: **Growth of the GenBank database**. This figure shows the growth of the GenBank database since its inception in 1982. Figure from `http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html`.

On the bright side, this problem is getting less important daily. Considering the rapid growth of sequence information (see Figure 4.1), it is safe to assume that in a few years this problem will be almost irrelevant. I say almost because biologists tend to assume that phylogenetically close species are extremely similar. This is, by and large, a perfectly valid assumption but it can lead to the dangerous belief that if we have say 95% of all organisms represented in our databases we will have 100% of all genetic information. We must always remember that near completeness is different from completeness, and there will always be missing something until we have **all** possible data collected.

In any case, I believe that despite the limitations discussed, the ORF approach is currently the best SECIS-independent method of selenoprotein gene prediction.

4.2 SelJ

Our paper (Castellano et al., 2005) presents the finding and functional characterization of the novel selenoprotein family SelJ[3]. SelJ was originally identified in the genome of the pufferfish *Tetraodon nigroviridis*.

This paper was one of the first blows that shook the belief that mammals recapitulate all eukaryotic selenoproteins. SelJ was the first selenoprotein family identified which is completely absent from mammalian species. Although it was known that MsrA was a selenoprotein in *C. reinhardtii* and SelU was one in fish, both these proteins are present as Cys homologs in mammals. SelJ by contrast, was only found in marine organisms (specifically jellyfish, anemones and sea urchins) and has no mammalian homologs.

Another surprise of this paper was that SelJ appears to have a structural role. It was the first, and so far the only, selenoprotein with such a role. Although we had no experimental evidence to support this claim, we determined that SelJ is part of the family of jellyfish J1-crystallins which in turn are a subfamily of the large family of ADP-ribosylation enzymes. This homology to these structural crystallins led us to hypothesize a structural role for SelJ as well.

Also in this paper we put forward the idea of mosaic evolution of selenoproteins which will be discussed in more detail later. Briefly, we took SelJ as further evidence of taxa specific selenoproteins whose existence is an indication that selenoproteins do not derive from a common ancestor but that different selenoproteins have different evolutionary histories.

4.3 Insects

In our article on the insect selenoproteomes (Chapple and Guigó, 2008), we describe the selenoproteomes of all insect genomes available at the time and identify the first animals to lack selenoprotein genes.

Our findings on insect genomes are without a doubt the most important results of my thesis. During the intensive analysis of the 12 fly genomes (Drosophila 12 genomes Consortium, 2007), we discovered that *Drosophila willistoni* lacks the machinery necessary for

[3]Named J for jellyfish

selenocysteine insertion and, indeed, has only Cys homologs of the *Drosophila melanogaster* selenoproteins SelK, SelH and SPS2.

After publication of the 12 *Drosophila* paper (Drosophila 12 genomes Consortium, 2007), where we first identified *D. willistoni* as the first animal to lack selenoproteins, we extended our analysis to all sequenced insects. It had previously been believed that all insects have the same selenoprotein complement as *D. melanogaster*. Our analysis showed that, in fact, *D. melanogaster* and all the Diptera are the exception rather than the rule. We found a general trend of selenoprotein loss in the Insecta class. Of all the insects investigated, selenoproteins were found only in the Diptera (see Figure 4.2).

The first important point of this paper is that there are animals that lack selenoprotein genes. This came as a complete surprise since it was believed that selenoproteins were essential for animal life. Indeed, mouse *tRNA*^{Sec} knock-outs have been shown to be lethal in-utero (Bösl et al., 1997). Similarly, mutant flies for *sps1* do not contain selenoproteins and are lethal at third instar larvae (Alsina et al., 1999). In contrast, Hirosawa-Takamori et al (Hirosawa-Takamori et al., 2004) have reported that mutant flies for *efsec* also fail to decode TGA as Sec but are viable and fertile.

Our finding that a *Drosophila* species naturally lacks selenoproteins indicates that these proteins are clearly not essential for animal life, and in fact, are probably not so for *D. melanogaster* either. Since SPS1 is known to have functions unrelated to selenoprotein synthesis (Alsina et al., 1999) and given that it was found in all insect genomes irrespective of whether they code for selenoproteins, it seems reasonable to assume that in light of our results, the lethality observed in SPS1 mutant flies is not related to the absence of selenoprotein expression. That EFsec mutant flies are viable also lends support to this hypothesis.

	SPS1	SPS2	SBP2	EFsec	tRNA^{Sec}	SelH	SelK	Secp43	SecS	PSTK
Drosophilas	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
D.willistoni	✓	x	✓	x	x	✓(Cys)	✓(Cys)	✓	x	x
A.gambiae	✓	✓	✓	✓	✓	✓	✓	x	✓	✓
A.aegypti	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A.mellifera	✓	✓	x	x	x	x	✓	✓	x	x
N.vitripennis	✓	x	✓	x	x	x	x	✓	x	x
B.mori	✓	x	x	x	x	x	x	✓	x	x
T.castaneum	✓	x	x	x	x	x	✓(Cys)	✓	x	x

Table 4.1: **A summary of the results for each selenoprotein and selenoprotein factor in all completely sequenced insect genomes.** Species that encode selenoproteins are shown in green, those that don't are shown in red. "✓" means the gene was present and conserved, "✓(Cys)" means the gene was found as a cysteine homolog and "x" means the gene was absent. Adapted from (Chapple and Guigó, 2008).

By checking for the presence/absence of all known selenoprotein specific factors (see Table 4.1) we were able to establish that *efsec* and *pstk* can be used as markers for the presence of selenoproteins in a given genome. Although, ideally, the tRNA^{Sec} would be the perfect marker, unfortunately the tRNA prediction programs are not that good at identifying it because their prediction methods are based on sequence similarity and not enough selenocysteine tRNAs have been annotated.

In an article published just before Chapple and Guigó (2008), Lobanov et al (Lobanov

et al., 2008) also analyzed insect genomes. Their analysis, however, is less comprehensive than ours, they missed *D. willistoni* and *N. vitripennis* and claim that *Apis mellifera* codes for a sec-containing SPS2. In fact, our analysis shows that although *A. mellifera* does have a conserved TGA codon at the "right" place in its *sps2* homolog, it lacks *efsec*, $tRNA^{Sec}$ and *pstk*. In addition, there was only a weak SECIS element (well below normal stability thresholds) in the UTR of the predicted gene. These results led us to posit that *A. mellifera* has lost the ability to code for selenoroteins but is in fact using another readthrough mechanism in this gene. Moreover, we have checked three other hymenoptera, the wasps *Nasonia vitripennis* (Chapple and Guigó, 2008), *Nasonia longicornis* and *Nasonia giraulti* (C.E. Chapple and M. Mariotti unpublished results) and they all show the same pattern. Although all three wasps lack the necessary machinery for selenoprotein synthesis, they have a conserved TGA in-frame. We are therefore convinced that this is, in fact, a novel readthrough mechanism, perhaps connected to the GAPsec readthrough identified by Hirosawa-Takamori et al (Hirosawa-Takamori et al., 2009).

Finally, in this article we identified a clear trend of selenoprotein loss in the Insecta class. I believe that this loss is both recent and still ongoing as evidenced by the situation in the Hymenoptera and the fact that we can still find fossil SECIS elements in some genes. On the other hand, the situation in the Diptera seems different. The selenoprotein genes seem to be very active in this order with loss and gain events still occurring. For example, SelK has been lost as a selenoprotein in *D. persimilis* while it has been duplicated as a Cys paralog in the melanogaster subgroup branch. SelH has been duplicated in *D. grimshawi*.

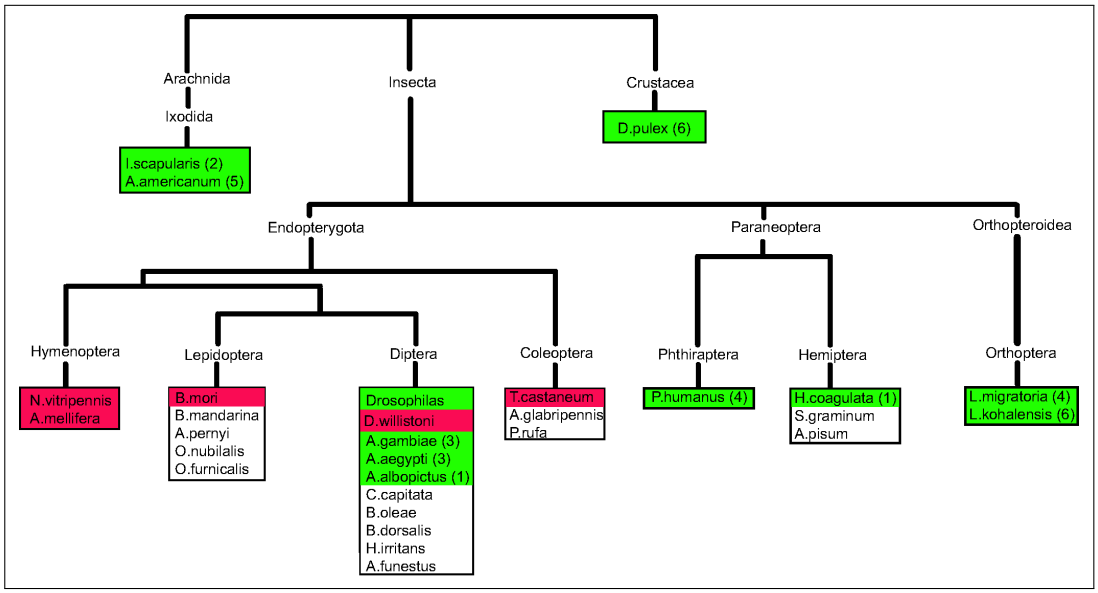

Figure 4.2: **Selenoprotein distribution in the arthropoda**. Species whose genomes do not code for selenoprotein genes are shown in red. Sec encoding species are shown in green with the number of selenoproteins found in each genome in parentheses next to its name. Figure taken from (Chapple and Guigó, 2008).

4.4 SECISaln

SECISaln is the result of a meeting between Roderic Guigó, Alain Krol and myself during the Selenium conference of 2006. Alain suggested that we collaborate to build a structural alignment of eukaryotic SECIS elements. To that end, I went to his lab in Strasbourg in November 2006 and began collecting SECIS sequences. Originally, the idea was to manually collect as many sequences as possible and then (again, manually) align them in terms of their structural units. This structural alignment does not rely on sequence identity, only structure. So, all helix Is would be aligned at the core quartet, all apical loops at the unpaired As etc. As I started building the collection and alignment I realized that it was going to be extremely tedious to do by hand and I started writing a script built around SECISearch which would automate the proces. This script kept growing until it turned into a web-based tool, SECISaln (`http://genome.crg.es/software/secisaln`).

SECISaln is first and foremost a program for building structural alignments of eukaryotic SECIS elements. It is capable of predicting a SECIS in the input sequence, determine its type (I or II), and correctly split it into its constituent structural units, each of which is then aligned against the other units of its type in the database.

SECISaln is also the most extensive (in fact, to my knowledge the only) database of manually curated eukaryotic SECIS elements. Analysis of the alignments produced by SECISaln allowed us to extend the consensus of the eukaryotic SECIS sequence and showed us that there is more variability than we expected (see Figure 1.8, page 21 and Chapple et al. (2009)).

4.5 Selenoprotein gene prediction, past and present

The traditional method for predicting novel selenoproteins has always been to purify a protein and then identify and sequence its cognate cDNA. Because this approach, although clearly effective, is very time consuming, computational techniques for the *in silico* prediction of selenoprotein genes were developed. The first such methods were developed independently by two groups in the late nineties. Kryukov *et al.* (1999) developed SECISearch (see Section 3.2.3, page 107) and used it to identify SelT and SelR in human EST data.

At the same time, Lescure *et al.* (1999) developed a similarly SECIS based approach using `rnamot` instead of `patscan` to find RNA structures and with no thermodynamic assessment. They used their method on ESTs and identified SelX (SelR), SelN and SelZ(SelT).

Probably the most important advance in selenoprotein gene prediction came in 2001 when the genefinder `GeneID` was modified to allow in-frame TGAs and coupled to SECISearch Castellano et al. (2001). `GeneID` is a hierarchical gene predictor developed by our group. It was modified by S. Castellano to allow for the prediction of in-frame TGA codons and incorporate such TGA-containing exons into a gene model in the presence of a suitable SECIS element downstream. However, while very efficient and still the best we have, this method has two major drawbacks. First, like all methods before it, it is SECIS dependent. If `GeneID` is not constrained by the locations of predicted SECIS elements, it will return far

too many false positives. This means that in order to make the results manageable we must rely on SECIS predictions. Therefore, this method can only cope with the identification of selenoprotein genes whose SECIS elements conform to the standard structure. Unfortunately, a variety of recent publications (e.g. Taskov et al. (2005); Lobanov et al. (2007); Novoselov et al. (2007); Chapple et al. (2009)) has shown that the SECIS structure is more permissive than originally thought. The second major drawback of the `GeneID` approach is that the gene finder must be trained on each species investigated, making it hard to test newly sequenced genomes in a high-throughput way.

The bottom line is that we still lack a truly effective method for identifying novel selenoprotein families. Although comparative studies are very promising, they will by definition miss taxa- or species-specific selenoproteins. On the other hand, the *ab initio* approach suffers from a high false positive rate. Coupling the *ab initio* approach to SECIS prediction increases specificity but decreases sensitivity. If we deal with SECIS variability by allowing atypical SECISes, the rate of false positive predictions will skyrocket. Finally,we still lack a reliable method for distinguishing between TGA as STOP and TGA as Sec. Any new *ab initio* technique will have to solve these problems. It must be able to correctly identify both canonical and non-canonical SECIS elements and differentiate between the two functions of TGA. Perhaps as we collect ever more selenoprotein genes we will be able to discern a pattern in the nucleotidic neighborhood of TGA that will enable us to recognize Sec codons.

4.6 Selenoprotein evolution

In my opinion, by far the most important question in the field of selenoproteins is how and why did these proteins evolve. Are they of monophyletic origin? Are they examples of convergent evolution? Why do some animals go to great evolutionary lengths to retain the selenocysteine encoding trait (e.g. *C. elegans*) while others seem to be fine without it (e.g. *D. willistoni*) ? These are still open questions. In the following sections I will give some of my thoughts and conclusions on this subject.

4.6.1 Selenoprotein origin

Where did selenoproteins come from? Are they modified Cysteine homologs, or vice versa? Was UGA originally a Sec codon? A Cys codon? A STOP codon? *Paramecium tetraurelia* has only a single STOP codon (see Section 1.1, page 6), UGA, and yet this one codon is still recoded to insert Sec (C. E. Chapple unpublished data).

Leinfelder and coworkers (1988) suggest that "UGA was originally a codon for Sec in the anaerobic world, perhaps two to three billion years ago, and after introduction of oxygen into biosphere this highly oxidizable amino acid could be maintained only in anaerobic organisms or in aerobic systems which evolved special protective mechanisms". In the aerobic world, the high reactivity of selenocysteine would lead nearly all Sec residues in proteins to be switched to Cys. Consequently, the UGA codon could have "acquired other functions such as its more familiar role in termination" while being retained in the rare case of selenoprotein genes as a sense codon.

Conversely, others (Jukes, 1990; Osawa et al., 1992) hold that the STOP meaning of the UGA codon predates its Sec-inserting function. Osawa and Jukes posited (Osawa et al., 1992) that Sec "is a recent evolutionary improvement and that the more primitive form of the enzymes used Cys with codons UGU and UGC and anticodon GCA. Perhaps, at this stage, UGA was a codon for Trp". They support this theory with the observation that certain selenoproteins are found to function but with a diminished catalytic activity when replacing Sec by Cys (e.g. Lee et al. (2000)). They also posit that the complex nature of the selenocysteine insertion machinery has "the appearance of sophisticated evolutionary innovation rather than the survival of primitive systems".

So, did the TGA codon follow a path from STOP to Sec? Or from Sec to Cys? Although this is clearly still an open question, personally, I find myself leaning more towards the view of Osawa and Jukes. If the original meaning of the UGA codon were Sec, I do not see how we could have arrived at so complex a decoding system as we have today. Its sophistication indicates a modification made to alter, or bypass, existing systems not to return to previously existing ones.

Additionally, whatever the first meaning of TGA, it is reasonable to expect that, at least in some genomes, it would be the *only* meaning of the codon. In all of the genetic codes discovered so far, we have come across none where TGA only codes for Sec. In many species, including most animals, it codes for both Sec and STOP. In one (so far), it codes for both Sec and Cys. In most it codes for only Cys or only Trp or only STOP (see Table 1.1, page 6), making it likelier that TGA was originally a signal for one of these three.

4.6.2 Cys/Sec exchangeablitiy

Most selenoproteins exist in both Sec and Cys forms in different (or even in the same) genomes. This has led to a wide held assumption that the two residues are highly exchangeable. Our finding that *D. willistoni* has only Cys homologs of the *Drosophila* selenoproteins (Chapple and Guigó, 2008) would appear to support that assumption. If the two residues are not functionally equivalent, the change from Sec to Cys would likely have been deleterious to *D. willistoni* and, therefore, would have been selected against. On the other hand, a recent study on Cys/Sec exchangeablitiy (see Castellano (2009) and references therein) found little evidence of functional exchangeability between the two residues in vertebrates. This contradicts our findings in insects. However, in the same article, Castellano posits that, even in insects, the two may be functionally equivalent and that the *D. willistoni* selenoproteome therefore reflects positive selection acting on Cys sites. Although the reasons for such selection are unclear, it is true that our results can be explained by assuming that either Sec homologs became disadvantageous for *D. willistoni* (for example, because of low environmental selenium) and were selected against or, conversely, that Cys homologs confer a selective advantage to this species and were selected for.

Cys/Sec functional exchangeability is still very much an open question. However, as more and more selenoproteomes are described and we get a clearer picture of both inter- and intra-species variation, it is one which today is ready to be examined in more detail.

4.6.3 Mosaic evolution

Selenoprotein families show a mosaic pattern of evolution. In other words, even within the same domain (the eukarya in this case) different families have taken different evolutionary paths in different organisms. This can be seen in the pattern of presence/absence of each protein family in different species as well as the pattern of Cys/Sec usage of each family in each species. So, certain families are present in Sec form in specific taxa, or even species. For example, SelJ is only found as a selenoprotein in actinopterygian fishes and sea urchin (Castellano et al., 2005). SelU is a selenoprotein in various vertebrate and invertebrate species but is only found as a Cys homolog in mammals (Castellano et al., 2004). The Insecta class gives a particularly clear picture of this mosaic evolution (see Figure 4.3).

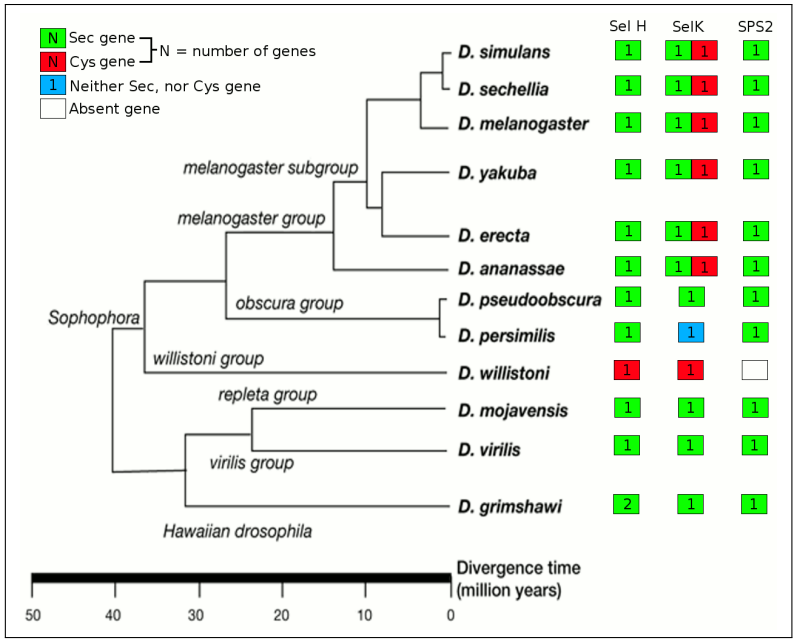

Figure 4.3: **Selenoprotein distribution in the *Drosophilas*.** Green cells are Sec homologs, red ones Cys homologs. SelK in *D. persimilis* is neither Cys, nor Sec and is shown in blue.

As has already been discussed in section 3.3, page 117, although the Arthropod phylum in general has a healthy selenoprotein complement, the Insecta class shows a marked reduction. However, even within insects this reduction is not uniform. Our data show that the different families have evolved independently even within the insects. For example, *D. willistoni* has only Cys homologs of the insect selenoproteins, while the species of the melanogaster subgroup all have two copies of SelK (one Cys and one Sec). In a previously unseen selenoprotein conversion, *D. persimilis*'s SelK is neither a Cys nor a Sec version and

instead has methionine at the same position[4].

This kind of evolutionary mosaicism implies that these families are evolutionarily very active. We cannot trace the distribution of selenoprotein genes across the eukaryotic domain to a few simple loss/gain events. Rather the picture we have is one of a complex pattern of gain/loss occurring independently in each family and each species.

4.7 Ruminations...

In this section I will give some of the questions and thoughts that I find most interesting in the field of selenoproteins today.

- The similarity between the bacterial and eukaryal selenocysteine insertion systems (see page 14 and Böck et al. (2006)) implies a common origin which predates the evolution of introns. However, the essentially non-existent overlap between the bacterial and eukaryal selenoproteomes and the pattern of mosaic evolution discussed in the previous section indicate separate evolution of the two domains.

- Since it has now been demonstrated that selenoproteins are not essential for animal life (Chapple and Guigó, 2008; Lobanov et al., 2008), why is it that such a complex system has been retained in the nematodes? The higher reactivity hypothesis won't cut it. Since Gromer *et al* (Gromer et al., 2003) demonstrated that cysteine thioredoxin reductases (TRs) can have similar activity to their selenocysteine counterparts, why have the nematodes retained the entire selenocysteine insertion machinery for just one protein, and a TR at that?

- What is so special about the UGA codon? Why do both Pyr and Sec use a UGA based recoding system? When I started working on *Paramecium tetraurelia* I expected that it either would have lost the ability to code for selenoproteins or would have used another codon to insert Sec. It seems that the dual meaning of the single STOP codon available in this genome introduces a dangerous ambiguity. It means that every single STOP in the genome could have a dual role. Surely that must lead to an increase in transcriptional errors.

- I expect we will find other cases similar to the insects. I believe it is clear that selenoproteins are very evolutionarily active, and we can expect to find different stories unfolding in different orders.

- Given the previous point, I believe that we will find some Orders or Classes of land plants with selenoproteins as well.

[4]Because of a T nucleotide which has caused a frameshift, eliminating the in-frame TGA, see Chapple and Guigó (2008)

CHAPTER 5

Conclusions

The research carried out during my PhD has resulted in or contributed to:

1. The discovery of a novel selenoprotein, SelJ.
2. The characterization of the selenoproteomes of *C. elegans*, *C. briggsae*, *Tetraodon nigroviridis*, 12 species of Drosophila [1], *Apis mellifera*, *Nasonia vitripennis*, *Nasonia giraulti*, *Tribolium castaneum*, *Anopheles gambiae*, *Aedes aegypti*, *Bombyx mori* and *Bos Taurus*.
3. The identification of the first animals to lack selenoproteins: the fly *D. willistoni*, the honey bee *Apis mellifera*, the wasps *Nasonia vitripennis*and *Nasonia giraulti*, the beetle *Tribolium castaneum* and the moth *Bombyx mori*.
4. The development of a novel, SECIS independent method for selenoprotein gene prediction, the ORF approach.
5. The development of a specialized program for the creation of structural alignments of SECIS elements, `SECISaln`.
6. The building of the largest collection of manually curated and structurally aligned eukaryotic SECIS elements (`http://genome.crg.es/software/secisaln`).
7. The refinement of the SECIS definition.
8. Furthering our understanding of selenoprotein evolution, especially what is termed the "mosaic" pattern of selenoprotein presence/absence.

[1] *D. melanogaster*, *D. willistoni*, *D. simulans*, *D. sechelia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. mojavensis*, *D. virilis* and *D. grimshawi*

CHAPTER 6

Epilogue

Automatization

While working on my PhD I have applied the `GeneID` and `SECISearch` based approach described by Kryukov et al. (2003) to the *Tetraodon* genome, the ORF approach to the *C. elegans* genome, and have performed dozens of comparative analyses of selenoprotein genes. The most effective way to distinguish *bona fide* selenoproteins from false positives I have been able to find is my own eye. Because of the vast amount of data available today, computational biologists have been veering increasingly toward a complete automatization of the prediction process. However, in the special case of selenoprotein genes such automatization is both difficult and, possibly, counterproductive.

I realize this opinion is anathema in a field dominated by computers and computer geeks but hear me out. For example, the past few years I have been involved in overseeing undergraduate student projects in the Pompeu Fabra University. The work of the students was centered around identifying known selenoproteins in novel genomes. The first step of this process was usually BLASTing the known selenoproteins against the genome and then analyzing the significant results. I was one of many supervisors on this project and the majority of the others came from different fields of bioinformatics. They told their students to collect their BLAST results in table format. This is an output mode of BLAST that summarizes the results, showing all relevant statistics (start, end, e-value, score etc) but no alignment. For most analyses, this format is sufficient. It is very concise and extremely parseable. However, selenoprotein homologs are often found with relatively low scores and conservation. Actually looking at the alignment is the only reliable way of recognizing a distant homolog. BLAST alignment scores refer to the entire HSP. Often, however, in the case of selenoproteins, we may find that the overall conservation is very low but the region around the Sec residue is particularly conserved, or that a redox box (CXXU) conserved. This information is lost when looking only at the descriptors of an HSP.

For my work, I have developed a program (`alignthingie`, see Section 3.2.1, page 105) which will filter BLAST HSPs according to various conservation criteria and select only those HSPs where a particular residue (or sequence) is found aligned to another. `alignthingie` also offers the ability to choose conservation levels around the aligned residue. The output of the program is the entire HSP (although gff format is also supported). In this way the researcher can look at the likeliest candidates without having to go through the entire BLAST file. I believe this is a good compromise between the manual and the automatic.

When all is said and done, selenoprotein gene prediction (as so many other things in science) is also a bit of an art. Experience counts. I remember when I first started my PhD and Sergi Castellano, who was finishing his, was showing me the ropes. I kept getting excited about a new result and Sergi would just look at the alignment and discard it. When I asked him to explain, to quantify, how he did this he could not give me any formal rules. In the same way I have seen Alain Krol glance at an SBP2 candidate I had identified and tell me it is not real. As time went by and I got familiar with the field, I found myself doing the same thing. Sometimes there is nothing quite as effective as the accumulation of experience. I seem to have developed a filter in my head capable of distinguishing the real (or at least probable) hits from the false positives. This kind of acquired discriminatory ability cannot be taught and cannot be formalized. As scientists we would do well to remember this and include, whenever possible, a step in our pipelines that brings the researcher into direct contact with the raw data.

REFERENCES

Allmang, C. and Krol, A. (2006). Selenoprotein synthesis: Uga does not end the story. *Biochimie*, 88(11):1561–1571.

Allmang, C., Wurth, L., and Krol, A. (2009). The selenium to selenoprotein pathway in eukaryotes: More molecular partners than anticipated. *Biochim Biophys Acta*.

Alsina, B., Corominas, M., Berry, M. J., na, J. B., and Serras, F. (1999). Disruption of selenoprotein biosynthesis affects cell proliferation in the imaginal discs and brain of drosophila melanogaster. *J Cell Sci*, 112 (Pt 17):2875–2884.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

Altschul, S. F., Madden, T. L., Schffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402.

Andersson, S. G. and Kurland, C. G. (1995). Genomic evolution drives the evolution of the translation system. *Biochem Cell Biol*, 73(11-12):775–787.

Andreesen, J. R. and Ljungdahl, L. G. (1973). Formate dehydrogenase of clostridium thermoaceticum: incorporation of selenium-75, and the effects of selenite, molybdate, and tungstate on the enzyme. *J Bacteriol*, 116(2):867–873.

Barrell, B. G., Bankier, A. T., and Drouin, J. (1979). A different genetic code in human mitochondria. *Nature*, 282(5735):189–194.

Beck, M. A., Levander, O. A., and Handy, J. (2003). Selenium deficiency and viral infection. *J Nutr*, 133(5 Suppl 1):1463S–1467S.

Beier, H. and Grimm, M. (2001). Misreading of termination codons in eukaryotes by natural nonsense suppressor trnas. *Nucleic Acids Res*, 29(23):4767–4782.

Berry, M. J., Banu, L., Chen, Y. Y., Mandel, S. J., Kieffer, J. D., Harney, J. W., and Larsen, P. R. (1991). Recognition of uga as a selenocysteine codon in type i deiodinase requires sequences in the 3′ untranslated region. *Nature*, 353(6341):273–276.

Birney, E., Clamp, M., and Durbin, R. (2004). Genewise and genomewise. *Genome Res*, 14(5):988–995.

Birringer, M., Pilawa, S., and Flohé, L. (2002). Trends in selenium biochemistry. *Nat Prod Rep*, 19(6):693–718.

Böck, A., Rother, M., Leibundgut, M., and Ban, N. (2006). *Selenium: Its molecular biology and role in human health, 2nd edition*, chapter Selenium metabolism in prokaryotes, pages 9–28. Springer Science+Business Media, New York, New York.

Bösl, M. R., Takaku, K., Oshima, M., Nishimura, S., and Taketo, M. M. (1997). Early embryonic lethality caused by targeted disruption of the mouse selenocysteine trna gene (trsp). *Proc Natl Acad Sci U S A*, 94(11):5531–5534.

Bovine Genome Sequencing and Analysis Consortium (2009). The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science*, 324(5926):522–528.

Brenner, S., Barnett, L., Katz, E. R., and Crick, F. H. (1967). Uga: a third nonsense triplet in the genetic code. *Nature*, 213(5075):449–450.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic dna. *J Mol Biol*, 268(1):78–94.

Carlson, B. A., Xu, X.-M., Kryukov, G. V., Rao, M., Berry, M. J., Gladyshev, V. N., and Hatfield, D. L. (2004). Identification and characterization of phosphoseryl-trna[ser]sec kinase. *Proc Natl Acad Sci U S A*, 101(35):12848–12853.

Carlson, B. A., Xu, X.-M., Shrimali, R., Sengupta, A., Yoo, M.-H., Irons, R., Zhong, N., and Hatfield, D. L. (2006). *Selenium: Its molecular biology and role in human health, 2nd edition*, chapter Mammalian and other eukaryotic selenocysteine tRNAs, pages 29–37. Springer Science+Business Media, New York, New York.

Castellano, S. (2009). On the unique function of selenocysteine - insights from the evolution of selenoproteins. *Biochim Biophys Acta*.

Castellano, S., Lobanov, A. V., Chapple, C., Novoselov, S. V., Albrecht, M., Hua, D., Lescure, A., Lengauer, T., Krol, A., Gladyshev, V. N., and Guigó, R. (2005). Diversity and functional plasticity of eukaryotic selenoproteins: identification and characterization of the selj family. *Proc Natl Acad Sci U S A*, 102(45):16188–16193.

Castellano, S., Morozova, N., Morey, M., Berry, M. J., Serras, F., Corominas, M., and Guigó, R. (2001). In silico identification of novel selenoproteins in the drosophila melanogaster genome. *EMBO Rep*, 2(8):697–702.

Castellano, S., Novoselov, S. V., Kryukov, G. V., Lescure, A., Blanco, E., Krol, A., Gladyshev, V. N., and Guigó, R. (2004). Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution. *EMBO Rep*, 5(1):71–77.

Chambers, I., Frampton, J., Goldfarb, P., Affara, N., McBain, W., and Harrison, P. R. (1986). The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded by the 'termination' codon, tga. *EMBO J*, 5(6):1221–1227.

Chapple, C. E. and Guigó, R. (2008). Relaxation of selective constraints causes independent selenoprotein extinction in insect genomes. *PLoS ONE*, 3(8):e2968.

Chapple, C. E., Guigó, R., and Krol, A. (2009). Secisaln, a web-based tool for the creation of structure-based alignments of eukaryotic secis elements. *Bioinformatics*, 25(5):674–675.

Chavatte, L., Brown, B. A., and Driscoll, D. M. (2005). Ribosomal protein l30 is a component of the uga-selenocysteine recoding machinery in eukaryotes. *Nat Struct Mol Biol*, 12(5):408–416.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003). Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res*, 31(13):3497–3500.

Commans, S. and Böck, A. (1999). Selenocysteine inserting trnas: an overview. *FEMS Microbiol Rev*, 23(3):335–351.

Cone, J. E., Río, R. M. D., Davis, J. N., and Stadtman, T. C. (1976). Chemical characterization of the selenoprotein component of clostridial glycine reductase: identification of selenocysteine as the organoselenium moiety. *Proc Natl Acad Sci U S A*, 73(8):2659–2663.

Copeland, P. R. and Driscoll, D. M. (1999). Purification, redox sensitivity, and rna binding properties of secis-binding protein 2, a protein involved in selenoprotein biosynthesis. *J Biol Chem*, 274(36):25447–25454.

Copeland, P. R., Stepanik, V. A., and Driscoll, D. M. (2001). Insight into mammalian selenocysteine insertion: domain structure and ribosome binding properties of sec insertion sequence binding protein 2. *Mol Cell Biol*, 21(5):1491–1498.

Crick, F. H. (1968). The origin of the genetic code. *J Mol Biol*, 38(3):367–379.

Crick, F. H., Barnett, L., Brenner, S., and Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, 192:1227–1232.

Crick, F. H., Griffith, J. S., and Orgel, L. E. (1957). Codes without commas. *Proc Natl Acad Sci U S A*, 43(5):416–421.

Cruveiller, S., Jabbari, K., Clay, O., and Bernardi, G. (2003). Compositional features of eukaryotic genomes for checking predicted genes. *Brief Bioinform*, 4(1):43–52.

Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). *A model of evolutionary change in proteins in Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Washington, DC.

Dillon, D. M. B. and Stapleton, J. T. (2007). Selenium effects on hiv rna and cd4 cell counts. *Arch Intern Med*, 167(14):1556–7; author reply 1557.

Ding, Y., Chan, C. Y., and Lawrence, C. E. (2004). Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res*, 32(Web Server issue):W135–W141.

Donovan, J., Caban, K., Ranaweera, R., Gonzalez-Flores, J. N., and Copeland, P. R. (2008). A novel protein domain induces high affinity selenocysteine insertion sequence binding and elongation factor recruitment. *J Biol Chem*, 283(50):35129–35139.

Driscoll, D. M. (2006). *Selenium: Its molecular biology and role in human health, 2nd edition*, chapter SECIS binding proteins and eukaryotic selenoprotein synthesis, pages 63–72. Springer Science+Business Media, New York, New York.

Drosophila 12 genomes Consortium (2007). Evolution of genes and genomes on the drosophila phylogeny. *Nature*, 450(7167):203–218.

Dsouza, M., Larsen, N., and Overbeek, R. (1997). Searching for patterns in genomic data. *Trends Genet*, 13(12):497–498.

Epstein, C. J. (1966). Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. *Nature*, 210(5031):25–28.

Fagegaltier, D., Lescure, A., Walczak, R., Carbon, P., and Krol, A. (2000). Structural analysis of new local features in secis rna hairpins. *Nucleic Acids Res*, 28(14):2679–2689.

Flohe, L., Gnzler, W. A., and Schock, H. H. (1973). Glutathione peroxidase: a selenoenzyme. *FEBS Lett*, 32(1):132–134.

Forstrom, J. W., Zakowski, J. J., and Tappel, A. L. (1978). Identification of the catalytic site of rat liver glutathione peroxidase as selenocysteine. *Biochemistry*, 17(13):2639–2644.

Ganichkin, O. M., Xu, X.-M., Carlson, B. A., Mix, H., Hatfield, D. L., Gladyshev, V. N., and Wahl, M. C. (2008). Structure and catalytic mechanism of eukaryotic selenocysteine synthase. *J Biol Chem*, 283(9):5849–5865.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., and Snyder, M. (2007). What is a gene, post-encode? history and updated definition. *Genome Res*, 17(6):669–681.

Giulio, M. D. (2005). The origin of the genetic code: theories and their relationships, a review. *Biosystems*, 80(2):175–184.

Gladyshev, V. N., Liu, A., Novoselov, S. V., Krysan, K., Sun, Q. A., Kryukov, V. M., Kryukov, G. V., and Lou, M. F. (2001). Identification and characterization of a new mammalian glutaredoxin (thioltransferase), grx2. *J Biol Chem*, 276(32):30374–30380.

Gromer, S., Johansson, L., Bauer, H., Arscott, L. D., Rauch, S., Ballou, D. P., Williams, C. H., Schirmer, R. H., and Arnr, E. S. J. (2003). Active sites of thioredoxin reductases: why selenoproteins? *Proc Natl Acad Sci U S A*, 100(22):12618–12623.

Grundner-Culemann, E., Martin, G. W., Harney, J. W., and Berry, M. J. (1999). Two distinct secis structures capable of directing selenocysteine incorporation in eukaryotes. *RNA*, 5(5):625–635.

Gu, Q. P., Beilstein, M. A., Vendeland, S. C., Lugade, A., Ream, W., and Whanger, P. D. (1997). Conserved features of selenocysteine insertion sequence (secis) elements in selenoprotein w cdnas from five species. *Gene*, 193(2):187–196.

Guigó, R., Flicek, P., Abril, J. F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V. B., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras, T. R., Harrow, J., Hubbard, T., Lewis, S. E., and Reese, M. G. (2006). Egasp: the human encode genome annotation assessment project. *Genome Biol*, 7 Suppl 1:S2.1–S231.

Guigó, R., Knudsen, S., Drake, N., and Smith, T. (1992). Prediction of gene structure. *J Mol Biol*, 226(1):141–157.

Guimaraes, M. J., Peterson, D., Vicari, A., Cocks, B. G., Copeland, N. G., Gilbert, D. J., Jenkins, N. A., Ferrick, D. A., Kastelein, R. A., Bazan, J. F., and Zlotnik, A. (1996). Identification of a novel seld homolog from eukaryotes, bacteria, and archaea: is there an autoregulatory mechanism in selenocysteine metabolism? *Proc Natl Acad Sci U S A*, 93(26):15086–15091.

Harrow, J., Nagy, A., Reymond, A., Alioto, T., Patthy, L., Antonarakis, S., and Guig, R. (2009). Identifying protein-coding genes in genomic sequences. *Genome Biol*, 10(1):201.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919.

Hirosawa-Takamori, M., Chung, H.-R., and Jckle, H. (2004). Conserved selenoprotein synthesis is not critical for oxidative stress defence and the lifespan of drosophila. *EMBO Rep*, 5(3):317–322.

Hirosawa-Takamori, M., Ossipov, D., Novoselov, S. V., Turanov, A. A., Zhang, Y., Gladyshev, V. N., Krol, A., Vorbrggen, G., and Jckle, H. (2009). A novel stem loop control element-dependent uga read-through system without translational selenocysteine incorporation in drosophila. *FASEB J*, 23(1):107–113.

Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee apis mellifera. *Nature*, 443(7114):931–949.

Hurwitz, B. E., Klaus, J. R., Llabre, M. M., Gonzalez, A., Lawrence, P. J., Maher, K. J., Greeson, J. M., Baum, M. K., Shor-Posner, G., Skyler, J. S., and Schneiderman, N. (2007). Suppression of human immunodeficiency virus type 1 viral load with selenium supplementation: a randomized controlled trial. *Arch Intern Med*, 167(2):148–154.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Bimont, C., Skalli, Z., Cattolico, L., Poulain, J., Berardinis, V. D., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.-P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K. J., McEwan, P., Bosak, S., Kellis, M., Volff, J.-N., Guigó, R., Zody, M. C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Qutier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E. S., Weissenbach, J., and Crollius, H. R. (2004). Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957.

Jukes, T. H. (1990). Genetic code 1990. outlook. *Experientia*, 46(11-12):1149–1157.

Katoh, K., Misawa, K., ichi Kuma, K., and Miyata, T. (2002). Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res*, 30(14):3059–3066.

Kinzy, S. A., Caban, K., and Copeland, P. R. (2005). Characterization of the secis binding protein 2 complex required for the co-translational insertion of selenocysteine in mammals. *Nucleic Acids Res*, 33(16):5172–5180.

Koch, N. C. (1967). Disappearance of the dinosaurs. *Journal of paleontology*, 41:970–972.

Koonin, E. V. and Novozhilov, A. S. (2009). Origin and evolution of the genetic code: the universal enigma. *IUBMB Life*, 61(2):99–111.

Korf, I., Flicek, P., Duan, D., and Brent, M. R. (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics*, 17 Suppl 1:S140–S148.

Kryukov, G. V., Castellano, S., Novoselov, S. V., Lobanov, A. V., Zehtab, O., Guigó, R., and Gladyshev, V. N. (2003). Characterization of mammalian selenoproteomes. *Science*, 300(5624):1439–1443.

Kryukov, G. V., Kryukov, V. M., and Gladyshev, V. N. (1999). New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *J Biol Chem*, 274(48):33888–33897.

Krzycki, J. A. (2005). The direct genetic encoding of pyrrolysine. *Curr Opin Microbiol*, 8(6):706–712.

Lassmann, T. and Sonnhammer, E. L. L. (2006). Kalign, kalignvu and mumsa: web servers for multiple sequence alignment. *Nucleic Acids Res*, 34(Web Server issue):W596–W599.

Lee, S. R., Bar-Noy, S., Kwon, J., Levine, R. L., Stadtman, T. C., and Rhee, S. G. (2000). Mammalian thioredoxin reductase: oxidation of the c-terminal cysteine/selenocysteine active site forms a thioselenide, and replacement of selenium with sulfur markedly reduces catalytic activity. *Proc Natl Acad Sci U S A*, 97(6):2521–2526.

Leinfelder, W., Zehelein, E., Mandrand-Berthelot, M. A., and Bck, A. (1988). Gene for a novel trna species that accepts l-serine and cotranslationally inserts selenocysteine. *Nature*, 331(6158):723–725.

Lescure, A., Allmang, C., Yamada, K., Carbon, P., and Krol, A. (2002). cdna cloning, expression pattern and rna binding analysis of human selenocysteine insertion sequence (secis) binding protein 2. *Gene*, 291(1-2):279–285.

Lescure, A., Gautheret, D., Carbon, P., and Krol, A. (1999). Novel selenoproteins identified in silico and in vivo by using a conserved rna structural motif. *J Biol Chem*, 274(53):38147–38154.

Lesoon, A., Mehta, A., Singh, R., Chisolm, G. M., and Driscoll, D. M. (1997). An rna-binding protein recognizes a mammalian selenocysteine insertion sequence element required for cotranslational incorporation of selenocysteine. *Mol Cell Biol*, 17(4):1977–1985.

Lobanov, A. V., Delgado, C., Rahlfs, S., Novoselov, S. V., Kryukov, G. V., Gromer, S., Hatfield, D. L., Becker, K., and Gladyshev, V. N. (2006). The plasmodium selenoproteome. *Nucleic Acids Res*, 34(2):496–505.

Lobanov, A. V., Fomenko, D. E., Zhang, Y., Sengupta, A., Hatfield, D. L., and Gladyshev, V. N. (2007). Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life. *Genome Biol*, 8(9):R198.

Lobanov, A. V., Hatfield, D. L., and Gladyshev, V. N. (2008). Selenoproteinless animals: Selenophosphate synthetase sps1 functions in a pathway unrelated to selenocysteine biosynthesis. *Protein Sci*, 17(1):176–182.

Low, S. C., Harney, J. W., and Berry, M. J. (1995). Cloning and functional characterization of human selenophosphate synthetase, an essential component of selenoprotein synthesis. *J Biol Chem*, 270(37):21659–21664.

Manichanh, C., Chapple, C. E., Frangeul, L., Gloux, K., Guigó, R., and Dore, J. (2008). A comparison of random sequence reads versus 16s rdna sequences for estimating the biodiversity of a metagenomic library. *Nucleic Acids Res*, 36(16):5180–5188.

Martin-Romero, F. J., Kryukov, G. V., Lobanov, A. V., Carlson, B. A., Lee, B. J., Gladyshev, V. N., and Hatfield, D. L. (2001). Selenium metabolism in drosophila: selenoproteins, selenoprotein mrna expression, fertility, and mortality. *J Biol Chem*, 276(32):29798–29804.

Mathews, D. H. (2006). Revolutions in rna secondary structure prediction. *J Mol Biol*, 359(3):526–532.

Mattick, J. S. and Makunin, I. V. (2006). Non-coding rna. *Hum Mol Genet*, 15 Spec No 1:R17–R29.

Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., Kanamori, H., Namiki, N., Kitagawa, M., Yamashita, H., Yasukochi, Y., Kadono-Okuda, K., Yamamoto, K., Ajimura, M., Ravikumar, G., Shimomura, M., Nagamura, Y., Shin-I, T., Abe, H., Shimada, T., Morishita, S., and Sasaki, T. (2004). The genome sequence of silkworm, bombyx mori. *DNA Res*, 11(1):27–35.

Mix, H., Lobanov, A. V., and Gladyshev, V. N. (2007). Secis elements in the coding regions of selenoprotein transcripts are functional in higher eukaryotes. *Nucleic Acids Res*, 35(2):414–423.

Navarro-Alarcon, M. and Cabrera-Vique, C. (2008). Selenium in food and the human body: a review. *Sci Total Environ*, 400(1-3):115–141.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217.

Novoselov, S. V., Hua, D., Lobanov, A. V., and Gladyshev, V. N. (2006). Identification and characterization of fep15, a new selenocysteine-containing member of the sep15 protein family. *Biochem J*, 394(Pt 3):575–579.

Novoselov, S. V., Lobanov, A. V., Hua, D., Kasaikina, M. V., Hatfield, D. L., and Gladyshev, V. N. (2007). A highly efficient form of the selenocysteine insertion sequence element in protozoan parasites and its use in mammalian cells. *Proc Natl Acad Sci U S A*, 104(19):7857–7862.

Novoselov, S. V., Rao, M., Onoshko, N. V., Zhi, H., Kryukov, G. V., Xiang, Y., Weeks, D. P., Hatfield, D. L., and Gladyshev, V. N. (2002). Selenoproteins and selenocysteine insertion system in the model plant cell system, chlamydomonas reinhardtii. *EMBO J*, 21(14):3681–3693.

Obata, T. and Shiraiwa, Y. (2005). A novel eukaryotic selenoprotein in the haptophyte alga emiliania huxleyi. *J Biol Chem*, 280(18):18462–18468.

Osawa, S., Jukes, T. H., Watanabe, K., and Muto, A. (1992). Recent evidence for evolution of the genetic code. *Microbiol Rev*, 56(1):229–264.

Parra, G., Agarwal, P., Abril, J. F., Wiehe, T., Fickett, J. W., and Guigó, R. (2003). Comparative gene prediction in human and mouse. *Genome Res*, 13(1):108–117.

Parra, G., Blanco, E., and Guigó, R. (2000). Geneid in drosophila. *Genome Res*, 10(4):511–515.

Passaretti, C. and Gupta, A. (2007). Selenium and hiv-1: hope or hype? *Arch Intern Med*, 167(22):2530–2531.

Rayman, M. P. (2000). The importance of selenium to human health. *Lancet*, 356(9225):233–241.

Rotruck, J. T., Pope, A. L., Ganther, H. E., Swanson, A. B., Hafeman, D. G., and Hoekstra, W. G. (1973). Selenium: biochemical role as a component of glutathione peroxidase. *Science*, 179(73):588–590.

Schuster, P., Fontana, W., Stadler, P. F., and Hofacker, I. L. (1994). From sequences to shapes and back: a case study in rna secondary structures. *Proc Biol Sci*, 255(1344):279–284.

Shchedrina, V. A., Novoselov, S. V., Malinouski, M. Y., and Gladyshev, V. N. (2007). Identification and characterization of a selenoprotein family containing a diselenide bond in a redox motif. *Proc Natl Acad Sci U S A*, 104(35):13919–13924.

Shen, Q., Fan, L., and Newburger, P. E. (2006). Nuclease sensitive element binding protein 1 associates with the selenocysteine insertion sequence and functions in mammalian selenoprotein translation. *J Cell Physiol*, 207(3):775–783.

Shrimali, R. K., Lobanov, A. V., Xu, X.-M., Rao, M., Carlson, B. A., Mahadeo, D. C., Parent, C. A., Gladyshev, V. N., and Hatfield, D. L. (2005). Selenocysteine trna identification in the model organisms dictyostelium discoideum and tetrahymena thermophila. *Biochem Biophys Res Commun*, 329(1):147–151.

Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31.

Small-Howard, A., Morozova, N., Stoytcheva, Z., Forry, E. P., Mansell, J. B., Harney, J. W., Carlson, B. A., Xu, X.-M., Hatfield, D. L., and Berry, M. J. (2006). Supramolecular complexes mediate selenocysteine incorporation in vivo. *Mol Cell Biol*, 26(6):2337–2346.

Squires, J. E., Stoytchev, I., Forry, E. P., and Berry, M. J. (2007). Sbp2 binding affinity is a major determinant in differential selenoprotein mrna translation and sensitivity to nonsense-mediated decay. *Mol Cell Biol*, 27(22):7848–7855.

Takeuchi, A., Schmitt, D., Chapple, C., Babaylova, E., Karpova, G., Guigó, R., Krol, A., and Allmang, C. (2009). A short motif in drosophila secis binding protein 2 provides differential binding affinity to secis rna hairpins. *Nucleic Acids Res*, 37:2126–2141.

Taskov, K., Chapple, C., Kryukov, G. V., Castellano, S., Lobanov, A. V., Korotkov, K. V., Guigó, R., and Gladyshev, V. N. (2005). Nematode selenoproteome: the use of the selenocysteine insertion system to decode one codon in an animal genome? *Nucleic Acids Res*, 33(7):2227–2238.

The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, 447(7146):799–816.

Tian, B., Bevilacqua, P. C., Diegelman-Parente, A., and Mathews, M. B. (2004). The double-stranded-rna-binding motif: interference and much more. *Nat Rev Mol Cell Biol*, 5(12):1013–1023.

Tormay, P., Wilting, R., Heider, J., and Bck, A. (1994). Genes coding for the selenocysteine-inserting trna species from desulfomicrobium baculatum and clostridium thermoaceticum: structural and evolutionary implications. *J Bacteriol*, 176(5):1268–1274.

Tsugita, A. and Fraenkel-Conrat, H. (1962). The composition of proteins of chemically evoked mutants of tmv rna. *J Mol Biol*, 4:73–82.

Turanov, A. A., Lobanov, A. V., Fomenko, D. E., Morrison, H. G., Sogin, M. L., Klobutcher, L. A., Hatfield, D. L., and Gladyshev, V. N. (2009). Genetic code supports targeted insertion of two amino acids by one codon. *Science*, 323(5911):259–261.

Turner, D. C. and Stadtman, T. C. (1973). Purification of protein components of the clostridial glycine reductase system and characterization of protein a as a selenoprotein. *Arch Biochem Biophys*, 154(1):366–381.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin,

D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Walczak, R., Westhof, E., Carbon, P., and Krol, A. (1996). A novel rna structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mrnas. *RNA*, 2(4):367–379.

Woese, C. R. (1965). On the evolution of the genetic code. *Proc Natl Acad Sci U S A*, 54(6):1546–1552.

Wong, J. T. (1975). A co-evolution theory of the genetic code. *Proc Natl Acad Sci U S A*, 72(5):1909–1912.

Wu, R., Shen, Q., and Newburger, P. E. (2000). Recognition and binding of the human selenocysteine insertion sequence by nucleolin. *J Cell Biochem*, 77(3):507–516.

Xu, X.-M., Carlson, B. A., Mix, H., Zhang, Y., Saira, K., Glass, R. S., Berry, M. J., Gladyshev, V. N., and Hatfield, D. L. (2007). Biosynthesis of selenocysteine on its trna in eukaryotes. *PLoS Biol*, 5(1):e4.

Xu, X.-M., Mix, H., Carlson, B. A., Grabowski, P. J., Gladyshev, V. N., Berry, M. J., and Hatfield, D. L. (2005). Evidence for direct roles of two additional factors, secp43 and soluble liver antigen, in the selenoprotein synthesis machinery. *J Biol Chem*, 280(50):41568–41575.

Zhang, W., Ramanathan, C. S., Nadimpalli, R. G., Bhat, A. A., Cox, A. G., and Taylor, E. W. (1999). Selenium-dependent glutathione peroxidase modules encoded by rna viruses. *Biol Trace Elem Res*, 70(2):97–116.

Zinoni, F., Birkmann, A., Stadtman, T. C., and Bck, A. (1986). Nucleotide sequence and expression of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from escherichia coli. *Proc Natl Acad Sci U S A*, 83(13):4650–4654.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–3415.

Zuker, M. and Sankoff, D. (1984). Rna secondary structures and their prediction. *Bull. Math. Biol.*, 46.

Acknowledgements

OK, given that in this section I am free of the normal constraints of academic writing, I will let my hair down and take a page out of Sol's book. So, the following is written in many languages, if a given phrase is in one you don't understand then, well, its not about you anyway!

Family

So, family first. **Μουτ** μου, τι να σου πρωτοθυμηθώ, τα παπούτσια τα μικρά... Δεν είναι το σωστό μέρος να σου δώσω τις ευχαριστίες που σου αρμόζουν. Ξέρεις όμως. **Dad**, thanks for the commas, and the lawnmowers, and that patience whose end I have yet to come up against. **Aliki** heh, thanks for teaching me to be an uncle, something far more formidable than a mere sibling. **Σπύρο**, αν και για χρόνια νόμιζα ότι εσύ ασχολείσαι με βιολιά (βιολόγος δεν είσαι;), μετά τους γονείς μου εσύ είσαι ο άνθρωπος που περισσότερο με έσπρωξε να φτάσω εκεί που βρίσκομαι σήμερα. **Πόπια**, να 'σαι καλά που μας αντέχεις όλους τους τσαπελοαρταβανέους. **Κατερινιό**, τώρα που έβγαλα και εγώ doctorat, να δούμε πότε θα σκάσω επιτέλους...**Γιαγιά**, εσένα τι να σου γράψω που από εσένα πηγάζουν όλα τα παραπάνω;

Friends

Αρί, να ήξερα πως με αντέχεις τόσα χρόνια... Να ήξερες πόσο με έχει βοηθήσει και με βοηθά η παρουσία σου... Ευχαριστώ. **Φοιβάκι**, να δω πότε θα ξαναβρεθούμε στην ίδια χώρα, δεν ξέρεις πόσο μου έχεις λείψει αυτά τα χρόνια. **Solenn**, mi Sol, nunca lo hubiera acabado todo eso sin tu presencia, y todavia me extraña tu ausencia. **Φίλιππε**, χε,

τρέχα γύρευε τώρα να σου διαλέξω γλώσσα... Thanks for all the multilingual conversaciones que hemos hecho toutes ces années. Still owe you a whisky or twelve. **Hog**, φέτος αγόρι μου επιτέλους πάλι Σκύρο! **Luisa**, nunca he dado un mechero con tan buenos resultados. **Marco**, selenobrothers in concert! **Manu, Elena, Ale, Sara, Dalila**, parezco a ABBA pero thank you for the music! **Solène**, thanks for your dreams and nightmares. **Bet**, a ti te debo abrazos. Hay que hacer cuentas... , **Esteve**, bombs happen, like s**t! But we don't mind! **Ramon**, espero que un dia hables el castellano como yo. **Roser**, moltes gràcies pel resum! Y por esos chats de esquizofrenia nocturna pretesica... **Axele, Ale**, gracias por todo que sin vosotros nunca hubiera venido a barcelona. **Sofia**, gracias petita por haber me hecho sentir mas de lo que soy. **Sergio, Enric**, fuisteis los primeros amigos que me hice en este ciudad. Barcelona no es la misma sin vosotros. **Jairo** ai este toque latino! **Camille**, je, gracias por este año tan lleno. **The geek parade**, otherwise known as **Roderic's tame bionformaticians**, thanks for making me take pride in my own geekiness! **Nico**, y moyollon! **Chays**, thanks for teaching me what "metagenomics" means. Not to mention irritable bowel syndrome! **Scott** I should have thanked you 7 years ago, I do so now. **Magali, Mariana** todabia os hechamos de menos aqui en Ventallo! **David, Thien** thanks for proving that post docs are human too! **Judit**, "hey Jud", "Sleeping Philippos blues", tantes cançons que has generat! **Davina**, gracias por volver. **Isa** ya que he acabado la tesis, supershifteamos? **Kreis, Paul, Dan, Zaxo, Zaf, Da, Pano, Bego** thanks! To all those people, too many to list, who have variously helped, entertained, intrigued or infuriated me over the years at the **PRBB** (and the **CMIMA**) thanks! **To those I missed**, sorry sorry sorry but ones mind is not at its best when finishing a thesis...

miju, a ti no te escribo nada pero a ti te lo debo todo.

I would also like to thank **Alain** for having me in his lab and so freely sharing his expertise over the years, and the people in his group, especially **Artemy**, for making me feel welcome.

I am also grateful to **my thesis tribunal** for taking the time to read my thesis. Especially those who made it all the way to the acknowledgements!

Finally, special thanks must needs go out to **Roderic** for not only guiding me through this thesis but actually putting up with me, above and beyond the call of duty, all these years and to **Sergi** who showed me that selenoproteins, contrary to popular belief, are COOL!

Appendices

APPENDIX A

Articles

A.1 Jaillon *et al*, 2004

In this publication (Jaillon et al., 2004), the first analysis of the *Tetraodon* genome is described. *Tetraodon* is a puffer fish with a compact genome and its preliminary gene annotation is presented here.

Our contribution to this multitudenary paper is the description and annotation of the *Tetraodon* selenoproteome. In short, it consists of 19 selenoprotein families, two more than in human. Apart the SelU family originally found in *Takifugu*, the *Tetraodon* genome also contains the novel SelJ family, which is widely distributed in, but restricted to, actinopterygians among vertebrates. Since the original article is 17 pages long, only the first page and the section relevant to selenoproteins are included here.

- Article abstract : http://www.nature.com/nature/journal/v431/n7011/abs/nature03025.html
- Full text : http://www.nature.com/nature/journal/v431/n7011/full/nature0302html
- PDF : http://www.nature.com/nature/journal/v431/n7011/pdf/nature03025.pdf
- Supplementary material : http://www.nature.com/nature/journal/v431/n7011/suppinfo/nature03025.html
- Database : http://www.genoscope.cns.fr/cgi-bin/ggb/gbrowse2?source=tetraodon
- Datasets : http://genome.crg.es/datasets/sptetra2004/

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E *et al*.
Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype.
Nature. 2004 Oct 21;431(7011):946-57.

Comment in:
Nature. 2004 Oct 21;431(7011):916-7.

A.2 Manichanh C. et al 2008

Manichanh C, Chapple CE, Frangeul L, Gloux K, Guigo R, Dore J. *A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library.*
Nucleic Acids Res. 2008 Sep;36(16):5180-8. Epub 2008 Aug 5.

This is an article (Manichanh et al., 2008) I did in collaboration with Chaysavannh Manichanh in which we present a novel method for the rapid phylogenetic classification of a metagenomic library.

_ Article abstract :
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2532719&rendertype=abstract
_ Full text :
http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18682527
_ PDF : http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=2532719&blobtype=pdf

A.3 Drosophila 12 Genomes Consortium, 2008

The following is the entire published article mentioned in section 2.3 (Drosophila 12 genomes Consortium, 2007).

APPENDIX B

IUPAC-IUB/GCG Ambiguity Codes

IUPAC-IUB/GCG Code	Meaning	Complement
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
X/N	G or A or T or C	X
.	not G or A or T or C	.

APPENDIX C

List of publications

Articles

The Bovine Genome Sequencing and Analysis Consortium (including C. E. Chapple and R. Guigó), Christine G. Elsik, Ross L. Tellam, Kim C. Worley.
The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science*, **324**(5926):522-528 (2009)

Takeuchi, A., Schmitt, D., Chapple, C., Babaylova, E., Karpova, G., Guigó, R., Krol, A. & Allmang, C.
A short motif in Drosophila SECIS Binding Protein 2 provides differential binding affinity to SECIS RNA hairpins. *Nucleic Acids Res*, Advance Access published on February 17, 2009

Chapple C.E., Guigó R. and Krol A.
SECISaln, a web-based tool for the creation of structure-based alignments of eukaryotic SECIS elements. *Bioinformatics* , **25**:674-675 (2009)

Chapple C.E., Guigó R.
Relaxation of selective constraints causes independent selenoprotein extinction in insect genomes. *PLoS ONE*, **3**(8):e2968 (2008).

Manichanh C, Chapple CE, Frangeul L, Gloux K, Guigó R, Dore J.
A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. *Nucleic Acids Res*, **36**(16):5180-8 (2008)

Drosophila 12 Genomes Consortium (including C. E. Chapple and R. Guigó)
Evolution of genes and genomes on the Drosophila phylogeny., *Nature* **450**(7167):203-18 (2007).

S. Castellano, A.V. Lobanov, C. Chapple, S.V. Novoselov, M. Albrecht, D. Hua, A. Lescure, T. Lengauer, A. Krol, V.N. Gladyshev and R. Guigó
Diversity and functional plasticity of eukaryotic selenoproteins: identification and characterization of the SelJ family. *PNAS*, **102**(45):16188-16193 (2005)

Taskov K., Chapple C., Kryukov G.V., Castellano S., Lobanov A.V.,Korotkov K.V., Guigó R. and Gladyshev V.N.
Nematode selenoproteome: the use of selenocysteine insertion system to decode one codon in an animal genome?, *Nucleic Acids Res.*, **33**:2227-2238 (2005)

Jaillon *et al.* (including C. E. Chapple and R. Guigó)
Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, **431**(7011):946-57 (2004)

Posters

Castellano S., Chapple C., and Guigó R.
Annotation of Eukaryotic Selenoproteins:
Finding the Needle in the Haystack
The Biology of Genomes, CSHL, New York (USA) (2004)

INDEX

GLOSSARY

CDS	A gene's **C**o**d**ing **S**equence is that part of a gene that is transcribed into protein (as opposed to, for example, introns and UTRs)
EFsec	Selenocysteine-specific elongation factor.
EST	**E**xpressed **S**equence **T**ags are short subsequences of a transcribed cDNA, produced by one-shot sequencing of a cloned mRNA.
PSTK	Phosphoseryl tRNA kinase, phosphorylates the Ser-tRNA^{Sec}.
Ribosomal protein L30	A component of the ribosome which has also been shown to bind the SECIS element.
SBP2	The **S**ecis **B**inding **P**rotein 2 is a eukaryotic protein, necessary for selenoprotein gene transcription. It interacts with the SECIS element, EFsec and the ribosome.
SECIS	The **Sec** **I**nsertion **S**equence, SECIS, is a three dimensional stem-loop structure on the 3' UTR of eukaryotic mRNAs which is necessary for the correct recodification of sec-encoding TGA codons.
secp43	Forms part of the SBP2/tRNA^{Sec}/EFsec complex but its exact role is unclear.
SecS	Eukaryotic selenocysteine Synthetase (previously SLA/LP), converts Ser-tRNA^{Sec} to Sec-tRNA^{Sec}.

Selenocysteine	Selenocysteine (Sec), the 21^{st} amino acid is an analog of cysteine with Selenium replacing Sulfur.
Selenoproteins	Selenoproteins are proteins that contain at least one selenocysteine residue
SPS1 and SPS2	Selenophosphate Synthetases
$tRNA^{Sec}$	Selenocysteine-specific transfer RNA.
UTR	The **U**n**t**ranslated **R**egions are stretches of DNA on one or both (5′ and 3′) sides of the coding sequence of a gene that survive into the mature mRNA but are not translated into protein.